Modeling Intention in Email:

Speech Acts, Information Leaks and User Ranking Methods

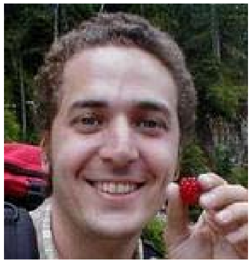
Vitor R. Carvalho

Carnegie Mellon University

William Cohen

Tom Mitchell

Jon Elsas



Ramnath Balasubramanyan

Outline

1. Motivation
2. Email Speech Acts
   - Modeling textual intention in email messages
3. Intelligent Email Addressing
   - Preventing information leaks
   - Ranking potential recipients
   - Cut Once – a Mozilla Thunderbird extension
4. Fine-tuning Ranking Models
   - Ranking in two optimization steps

Why Email

- The most successful e-communication application.
  - Great tool to collaborate, especially in different time zones.
  - Very cheap, fast, convenient and robust. It just works.

- Increasingly popular

[Shipley & Schwalbe, 2007]

  - Clinton adm. left 32 million emails to the National Archives
  - Bush adm.…more than 100 million in 2009 (expected)

- *Visible* impact
  - Office workers in the U.S. spend at least 25% of the day on email – not counting handheld use

Hard to manage

- People get overwhelmed.

  [Dabbish & Kraut, CSCW-2006].

  [Belloti et al. HCI-2005]

  - Costly interruptions
  - Serious impacts on work productivity
  - Increasingly difficult to manage requests, negotiate shared tasks and keep track of different commitments

- People make horrible mistakes.
  - *“I accidentally sent that message to the wrong person”*
  - *“Oops, I forgot to CC you his final offer”*
  - *“Oops, Did I just hit reply-to-all?”*

Outline

1. Motivation
2. Email Speech Acts
   - Modeling textual intention in email messages
3. Intelligent Email Addressing
   - Preventing information leaks
   - Ranking potential recipients
   - Cut Once – a Mozilla Thunderbird extension
4. Fine-tuning Ranking Models
   - Ranking in two optimization steps

Example

From: Benjamin Han
To: Vitor Carvalho
Subject: LTI Student Research Symposium

Hey Vitor

When exactly is the LTI SRS submission deadline?

Also, don't forget to ask Eric about the SRS webpage.

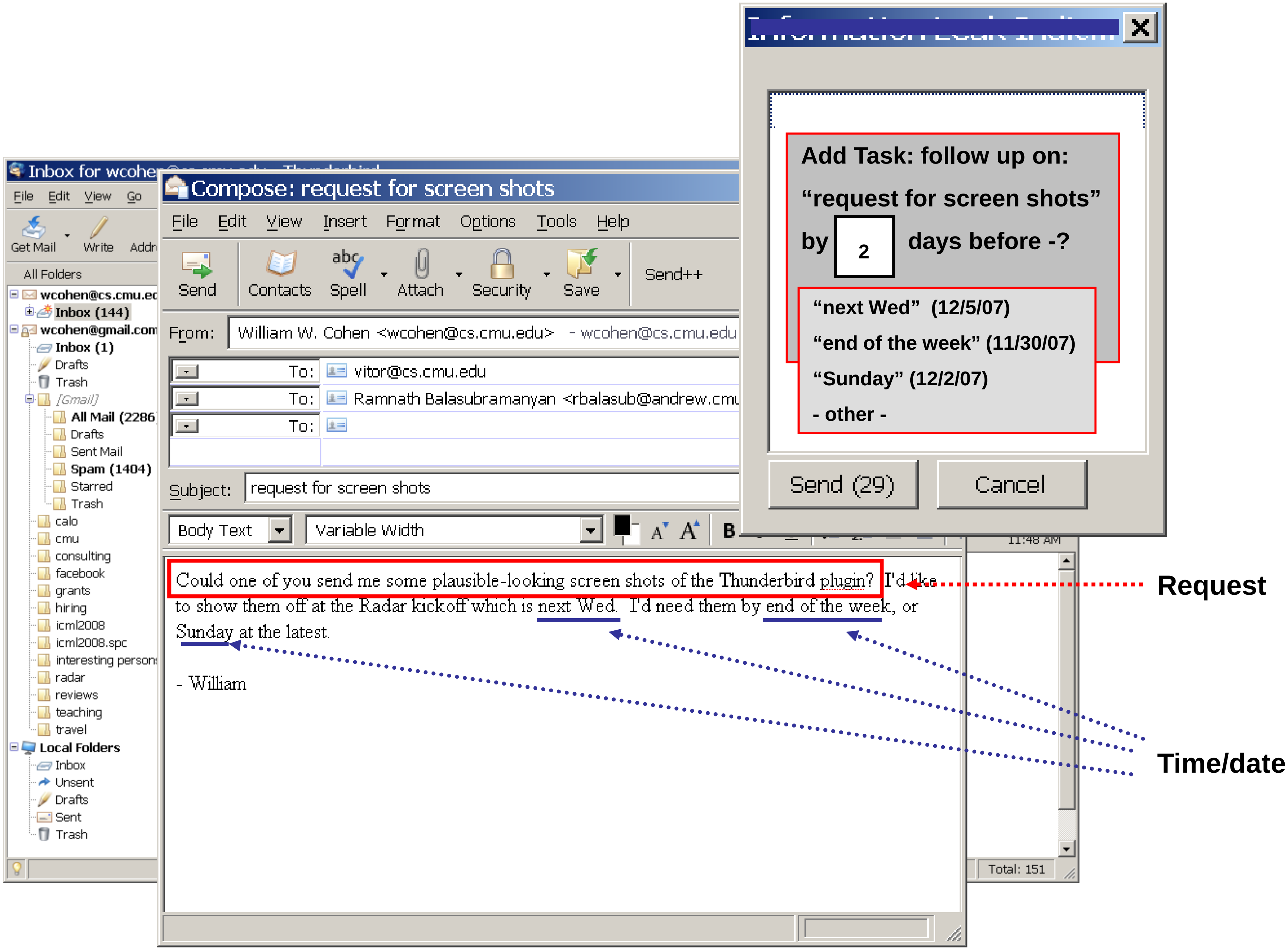
Thanks.

Ben

Request - Information

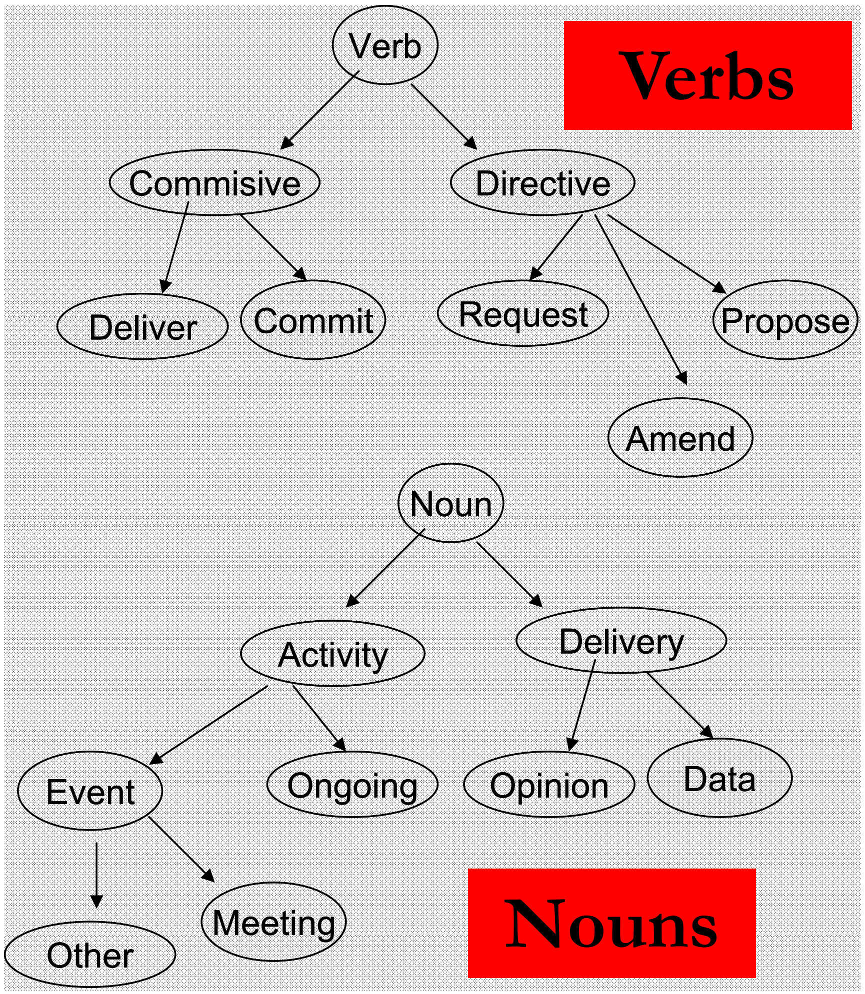
Reminder - Action/Task

- ✓ Prioritize email by "intention"
- ✓ Help keep track of your tasks:
  - ✓ pending requests, commitments, reminders, answers, etc.
- ✓ Better integration with to-do lists


Inbox for wcohen
File Edit View Go
Get Mail
Write
All Folders
wcohen@cs.cmu.ec
Inbox (144)
wcohen@gmail.com
Inbox (1)
Drafts
Trash
[Gmail]
All Mail (2286
Drafts
Sent Mail
Spam (1404)
Starred
Trash
calo
cmu
consulting
facebook
grants
hiring
icml2008
icml2008.spc
interesting person
radar
reviews
teaching
travel
Local Folders
Inbox
Unsent
Drafts
Sent
Trash
Compose: request for screen shots
File Edit View Insert Format Options Tools Help
Send
Contacts
Spell
Attach
Security
Save
Send++
From: William W. Cohen <wcohen@cs.cmu.edu> - wcohen@cs.cmu.edu
To: vitor@cs.cmu.edu
To: Ramnath Balasubramanyan <rbalasub@andrew.cmu
To:
Subject: request for screen shots
Body Text
Variable Width
Could one of you send me some plausible-looking screen shots of the Thunderbird plugin? I'd like to show them off at the Radar kickoff which is next Wed. I'd need them by end of the week, or Sunday at the latest.
- William
Total: 151
Add Task: follow up on:
"request for screen shots"
by 2 days before -?
"next Wed" (12/5/07)
"end of the week" (11/30/07)
"Sunday" (12/2/07)
- other -
Send (29)
Cancel
Request
Time/date

Classifying Email into Acts

[Cohen, Carvalho & Mitchell, EMNLP-04]



- **An Act is described as a <u>verb-noun</u> pair (e.g., propose meeting, request information) - Not all pairs make sense**
- **One single email message may contain multiple acts**
- **Try to describe commonly observed behaviors, rather than all possible speech acts in English**
- **Also include non-linguistic usage of email (e.g. delivery of files)**

Data & Features

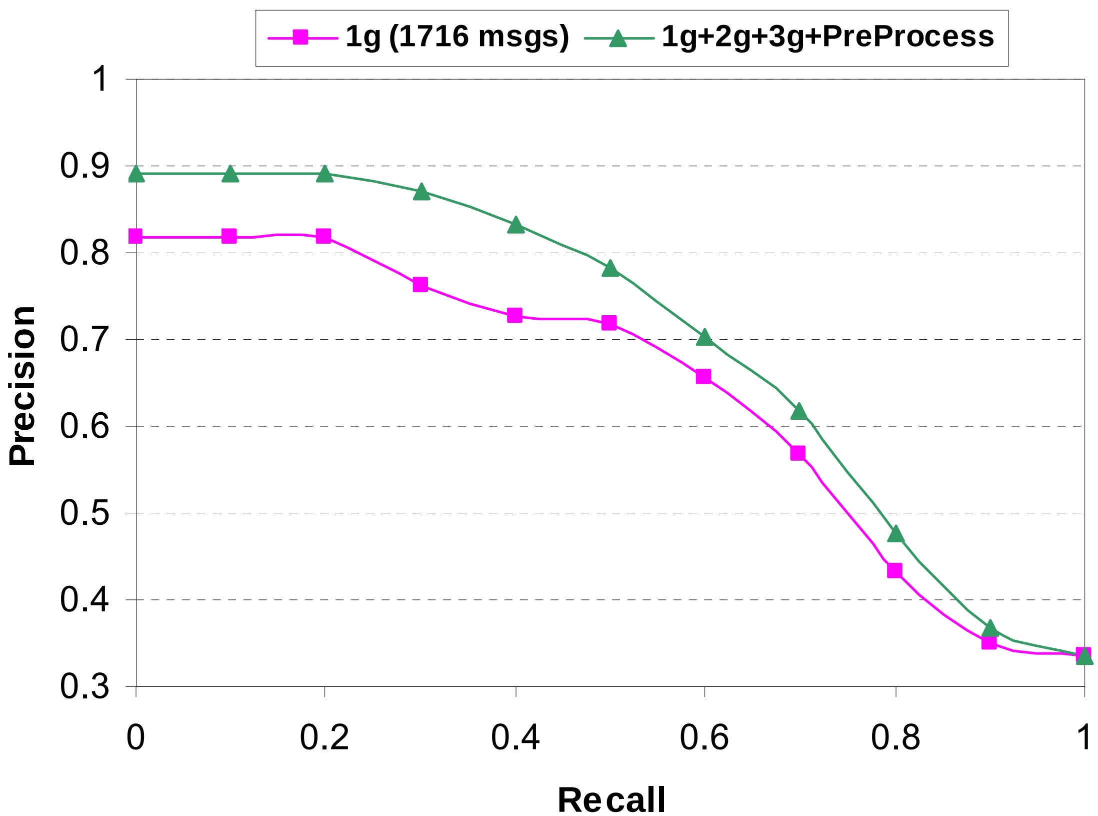
- **Data: Carnegie Mellon MBA students competition**
  - Semester-long project for CMU MBA students. Total of 277 students, divided in 50 teams (4 to 6 students/team). Rich in task negotiation.
  - 1700+ messages (from 5 teams) were manually labeled. One of the teams was double labeled, and the inter-annotator agreement ranges from 0.72 to 0.83 (Kappa) for the most frequent acts.

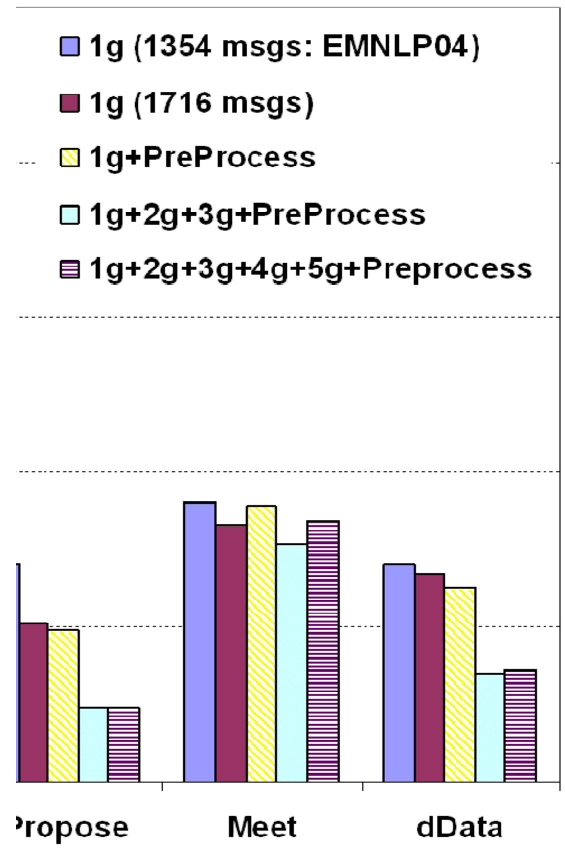
- **Features:**
  - N-grams: 1-gram, 2-gram, 3-gram,4-gram and 5-gram
  - Pre-Processing
    - Remove Signature files, quoted lines (in-reply-to) [Jangada package]
    - Entity normalization and substitution patterns:
      - “Sunday”…”Monday” →[day], [number]:[number] → [hour],
      - “me, her, him ,us or them” → [me],
      - “after, before, or during” → [time], etc

Error Rate for Various Acts

[Carvalho & Cohen, HLT-ACTS-06]

[Cohen, Carvalho & Mitchell, EMNLP-04]





5-fold cross-validation over 1716 emails, SVM with linear kernel

Best features

(selected by Information Gain)

Request	Commit	Meeting
[wwhh] do [person] think	is good for [me]	[day] at [hour] [pm]
[me] know [wwhh]	i will bring copies	let's plan to meet
that would be great	i will do the	meet at [hour] [pm]

1-gram	2-gram	3-gram	4-gram	5-gram
?	do [person]	[person] need to	[wwhh] do [person] think	[wwhh] do [person] think ?
please	? [person]	[wwhh] do [person]	do [person] need to	let [me] know [wwhh] [person]
[wwhh]	could [person]	let [me] know	and let [me] know	a call [number]-[number]
could	[person] please	would [person]	call [number]-[number]	give [me] a call [number]
do	? thanks	do [person] think	would be able to	please give give [me] a call
can	are [person]	are [person] meeting	[person] think [person] need	[person] would be able to
of	can [person]	could [person] please	let [me] know [wwhh]	take a look at it
[me]	need to	do [person] need	do [person] think ?	[person] think [person] need to

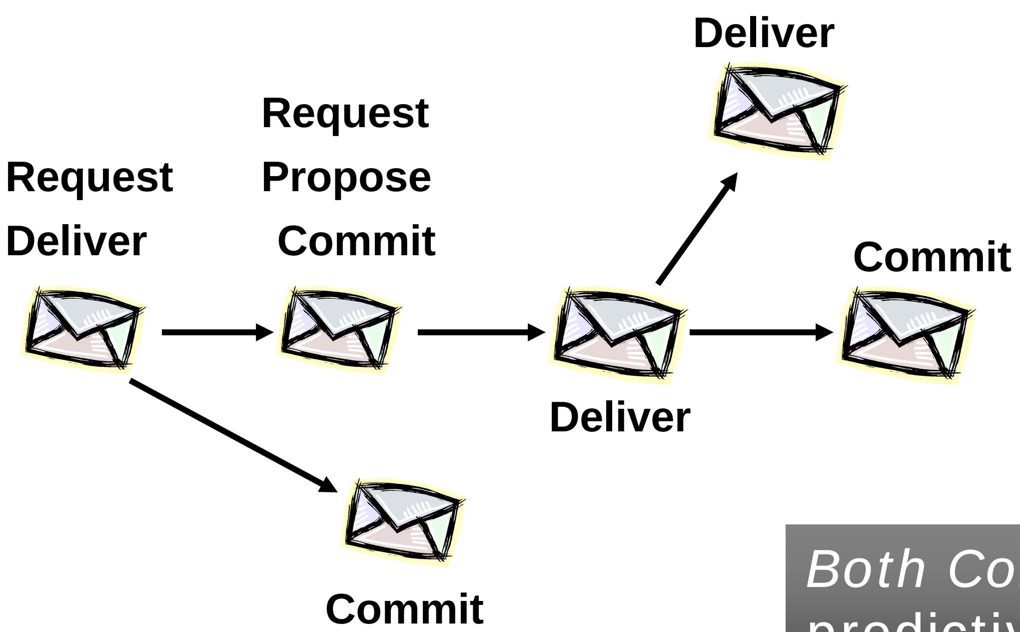
Table 2: Request Act:Top eight N-grams Selected by Information Gain.

Ciranda:
Java package for Email Speech Act Classification

Idea: Predicting Acts from Surrounding Acts

[Carvalho & Cohen, SIGIR-05]

Example of Email Thread Sequence



Strong correlation between previous and next message's acts

Act has little or no correlation with other acts of *same* message

Both Context and Content have predictive value for email act classification

Context: Collective classification problem

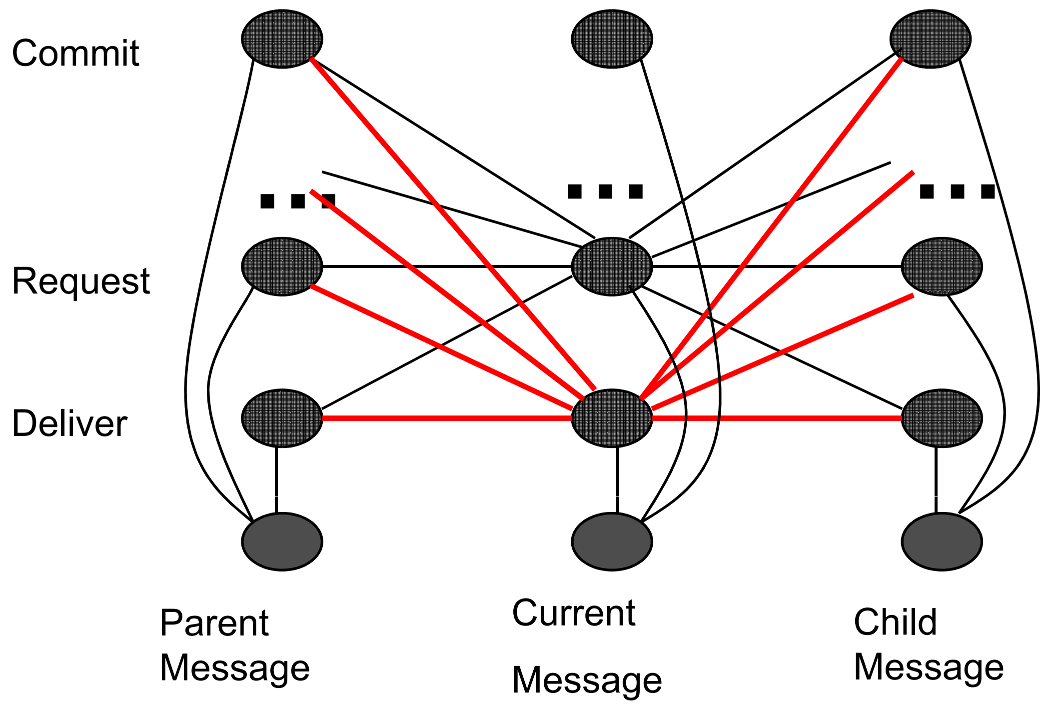
Collective Classification with Dependency Networks (DN) [Carvalho & Cohen, SIGIR-05]

- In DNs, the full joint probability distribution is approximated with a set of conditional distributions that can be learned independently. The conditional probabilities are calculated for each node given its *Markov blanket*.

$$\Pr(\vec{X}) = \prod_i \Pr(X_i \mid Blanket(X_i))$$

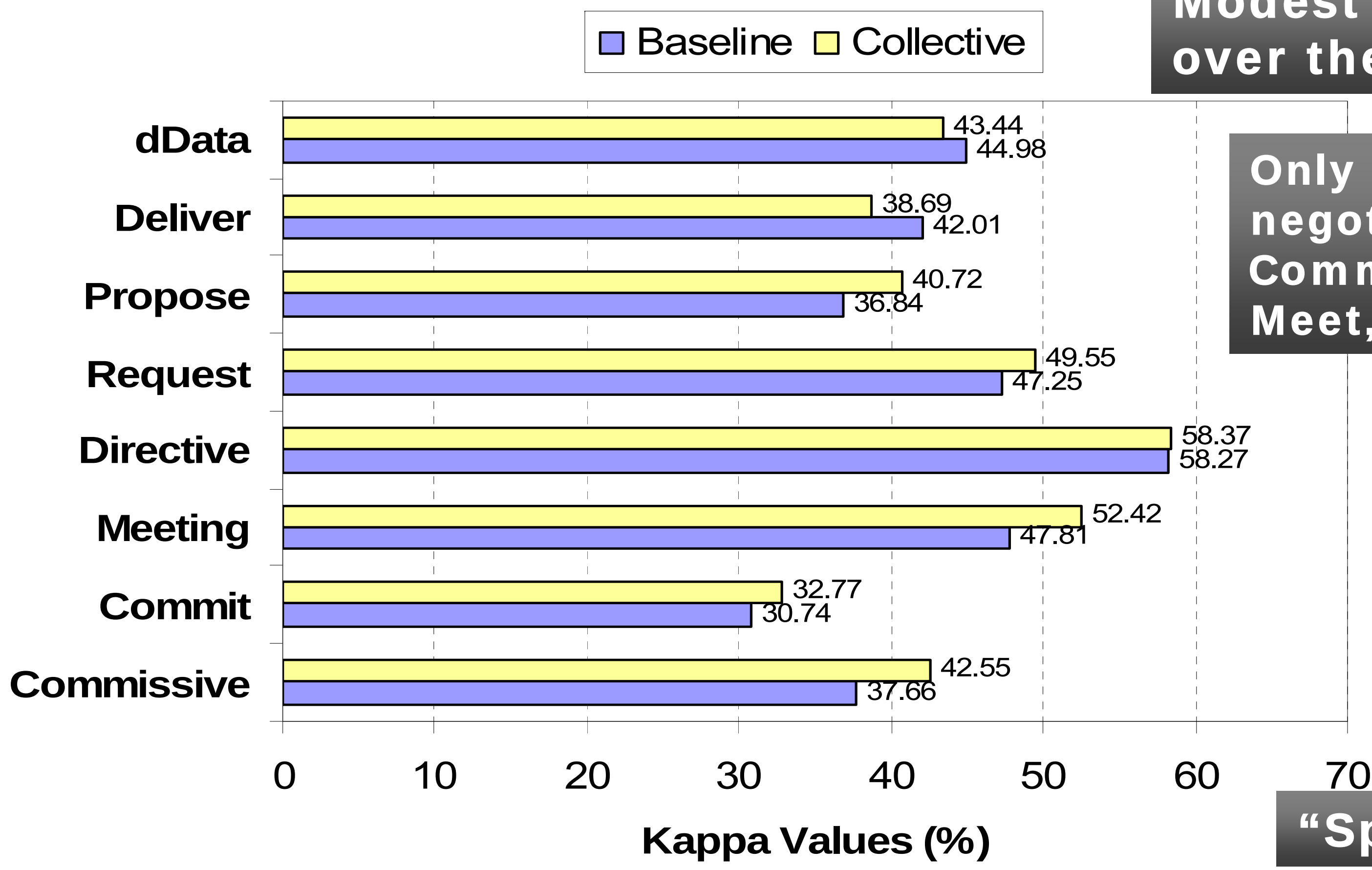
[Heckerman et al., JMLR-00]

[Neville & Jensen, JMLR-07]



Inference: Temperature-driven Gibbs sampling

Act by Act Comparative Results



Modest improvements over the baseline

Only on acts related to negotiation: Request, Commit, Propose, Meet, Commissive, etc.

"Sparse" links

Kappa values with and without collective classification, averaged over four team test sets in the leave-one-team out experiment.

Applications of Email Acts

- Iterative Learning of Email Tasks and Email Acts

  [Kushmerick & Khousainov, IJCAI-05]

- Predicting Social Roles and Group Leadership

  [Leusky,SIGIR-04][Carvalho,Wu & Cohen, CEAS-07]

- Detecting Focus on Threaded Discussions

  [Feng et al., HLT/NAACL-06]

- Automatically Classifying Emails into Activities

  [Dredze, Lau & Kushmerick, IUI-06]

Outline

1. Motivation √
2. Email Speech Acts √
   - Modeling textual intention in email messages
3. Intelligent Email Addressing
   - Preventing information leaks
   - Ranking potential recipients
4. Fine-tuning User Ranking Models
   - Ranking in two optimization steps

FreeRepublic.com *"A Conservative News Forum"*

[Last | Latest Posts | Latest Articles | Self Search | Add Bookmark | Post | Abuse | Help!]

California Power-Buying Data Disclosed in Misdirected E-Mail

News/Current Events **Breaking News News** Keywords: **CALPOWERCRISIS CALIFORNIA POWER CRISIS**
Source: **Bloomberg.com**
Published: **July 6, 2001**
Posted on **07/06/2001 12:30:48 PDT** by **John Jorsett**

Sacramento, California, July 6 (Bloomberg) -- California Governor Gray Davis's office released data on the state's purchases in the spot electricity market -- information Davis has been trying to keep secret -- through a misdirected e-mail.

The e-mail, containing data on California's power purchases yesterday, was intended for members of the governor's staff, said Davis spokesman Steve Maviglio. It was accidentally sent to some reporters on the office's press list, he said.

Davis is fighting disclosure of state power purchases, saying it would compromise negotiations for future contracts. This week, Davis appealed a state judge's order to release spot-market invoices, purchase orders and confirmation sheets for power contracts signed through June 27. The state is buying electricity on behalf of utilities, which are burdened by debt.

``It's an internal document,'' Maviglio said of the e-mail. ``We have a meeting every morning where

Done

Lilly's $1 Billion E-Mailstrom

by Katherine Eban | Feb 5 2008

A secret memo meant for a colleague lands in a *Times* reporter's in-box.

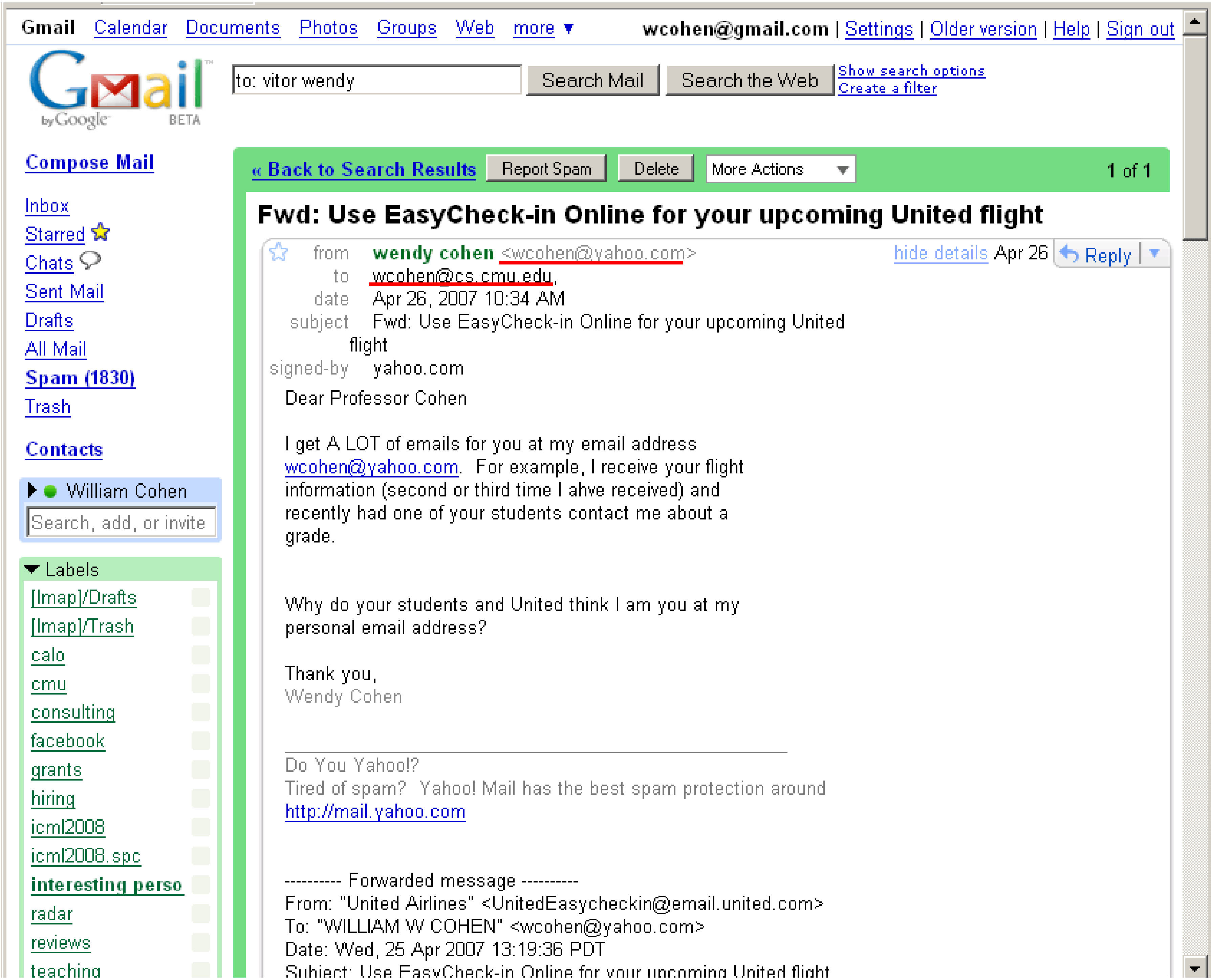
When the *New York Times* broke the story last week

Eli Lilly & Co. was in confidential settlement talks wi

government, angry calls flew behind the scenes as th

As the company's lawyers began turning over rocks closer to home, however, they discovered what could be called *A Nightmare on Email Street,* a pharmaceutical consultant told Portfolio.com. One of its outside lawyers at Philadelphia-based Pepper Hamilton had mistakenly emailed confidential information on the talks to *Times* reporter *Alex* Berenson instead of *Bradford* Berenson, her co-counsel at Sidley Austin.

With the negotiations over alleged marketing improprieties reaching a mind-boggling sum of $1 billion, Eli Lilly had every reason to want to keep the talks under wraps. It was paying the two fancy law firms a small fortune to negotiate deftly and quietly.


Gmail Calendar Documents Photos Groups Web more ▾
wcohen@gmail.com | Settings | Older version | Help | Sign out
Gmail by Google BETA
to: vitor wendy
Search Mail
Search the Web
Show search options
Create a filter
Compose Mail
Inbox
Starred
Chats
Sent Mail
Drafts
All Mail
Spam (1830)
Trash
Contacts
William Cohen
Search, add, or invite
Labels
[Imap]/Drafts
[Imap]/Trash
calo
cmu
consulting
facebook
grants
hiring
icml2008
icml2008.spc
interesting perso
radar
reviews
teaching
« Back to Search Results
Report Spam
Delete
More Actions
1 of 1
Fwd: Use EasyCheck-in Online for your upcoming United flight
from wendy cohen <wcohen@yahoo.com>
hide details Apr 26
Reply
to wcohen@cs.cmu.edu,
date Apr 26, 2007 10:34 AM
subject Fwd: Use EasyCheck-in Online for your upcoming United flight
signed-by yahoo.com
Dear Professor Cohen
I get A LOT of emails for you at my email address wcohen@yahoo.com. For example, I receive your flight information (second or third time I ahve received) and recently had one of your students contact me about a grade.
Why do your students and United think I am you at my personal email address?
Thank you,
Wendy Cohen
Do You Yahoo!?
Tired of spam? Yahoo! Mail has the best spam protection around
http://mail.yahoo.com
---------- Forwarded message ----------
From: "United Airlines" <UnitedEasycheckin@email.united.com>
To: "WILLIAM W COHEN" <wcohen@yahoo.com>
Date: Wed, 25 Apr 2007 13:19:36 PDT
Subject: Use EasyCheck-in Online for your upcoming United flight

http://www.sophos.com/

In this section

Journalists' guide
News archive

Talk to our experts

Press contacts

Resources

Sophos Podcasts
SophosLabs blog
Image gallery
Customer case studies
Free anti-virus
White papers
Awards and reviews
Industry affiliations

Info feeds

Security news
Company news

Security news
Company news

What are info feeds?

16 November 2007

70% of businesses concerned about data leakage via email

With half of employees admitting to sending email to the wrong person, firms are right to be worried

Research conducted by IT security and control firm Sophos has revealed that 70 percent of businesses are concerned about sensitive material falling into the wrong hands as a result of data leakage via email.

50% of computer users have accidentally sent a sensitive email to the wrong person.

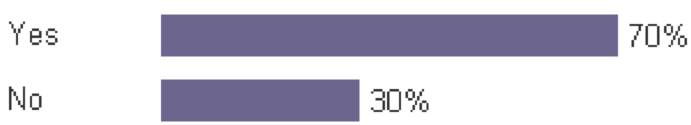
A further 50 percent of employees admit to having accidentally sent an embarrassing or sensitive email to the wrong person from the workplace, demonstrating that email leakage is a very real concern. Sophos experts note that it can potentially cause corporate embarrassment, compliance breaches and the loss of business critical information.

Sophos experts note that there can also be a significant financial impact from data such as customer lists, engineering information, and financial statements falling into the wrong hands. Suffering economic loss is undoubtedly the most serious potential consequence of data leakage.

"As more and more business, and indeed personal interaction, is conducted via work email, the risk of slipping up and clicking send without double-checking the recipient's details is ever-growing," said Graham Cluley, senior technology consultant at Sophos. "The fact that as many as half of employees have experienced that heart-stopping moment when they realize that their message is hurtling towards the wrong person shows that the human error factor is too significant to ignore. Businesses would be wise to check that their email security solutions have the facility to prevent this from happening by identifying when sensitive data or attachments are contained in the message, and if they don't, to consider a more water-tight alternative."

Survey results

Are you worried about sensitive data leaking from your company via email?

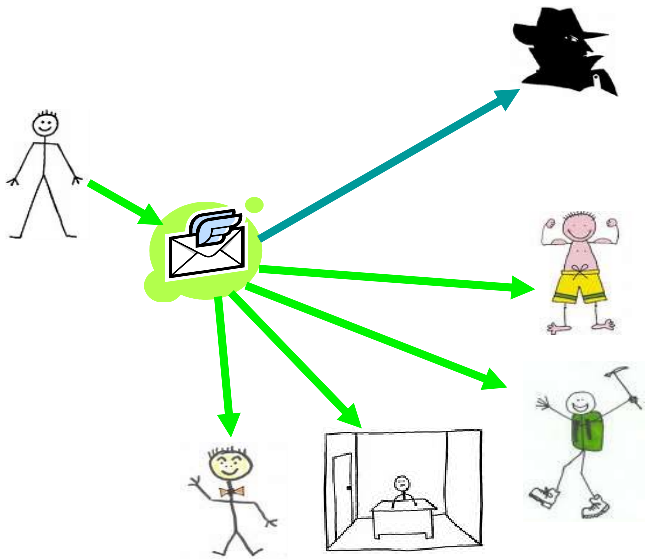


Sophos online survey, 200 respondents, November 2007

Preventing Email Info Leaks

[Carvalho & Cohen, SDM-07]

Email Leak: email accidentally sent to wrong person



Disastrous consequences: expensive law suits, brand reputation damage, negotiation setbacks, etc.

No labeled data

- **Who would give me this kind of data?**

1. **Similar first or last names, aliases, etc**
2. **Aggressive auto-completion of email addresses**
3. **Typos**
4. **Keyboard settings**

Preventing Email Info Leaks

[Carvalho & Cohen, SDM-07]

- **Method**
  1. Create simulated/artificial email recipients

  2. Build model for (msg.recipients): train classifier on real data to detect synthetically created outliers (added to the true recipient list).
     - **Features**: textual(subject, body), network features (frequencies, co-occurrences, etc).

  3. Detect outlier and warn user based on confidence.

1. Similar first or last names, aliases, etc

2. Aggressive auto-completion of email addresses

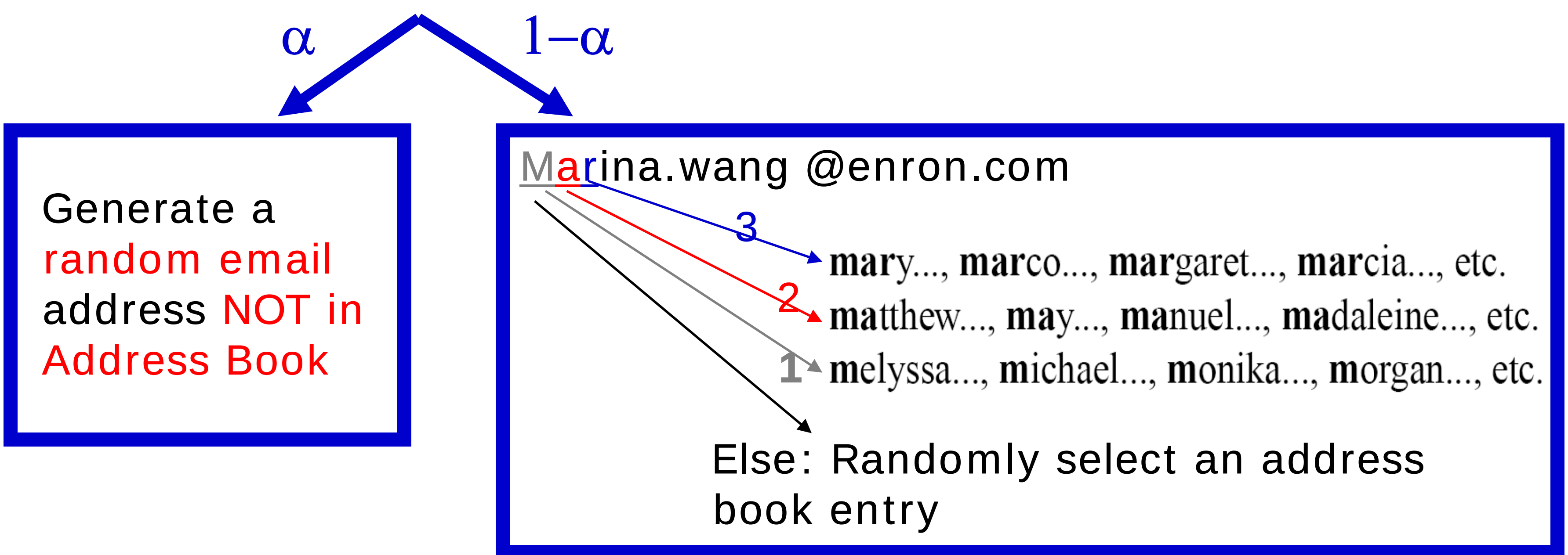
3. Typos

4. Keyboard settings

Simulating Email Leaks

- Several options:
  - Frequent typos, same/similar last names, identical/similar first names, aggressive auto-completion of addresses, etc.

- We adopted the *3g-address* criteria:
  - On each trial, one of the msg recipients is randomly chosen and an outlier is generated according to:



Data and Baselines

- Enron email dataset, with a realistic setting
  - For each user, ~10% most recent sent messages were used as test
  - Some basic preprocessing

- Baseline methods:
  - Textual similarity
  - Common baselines in IR

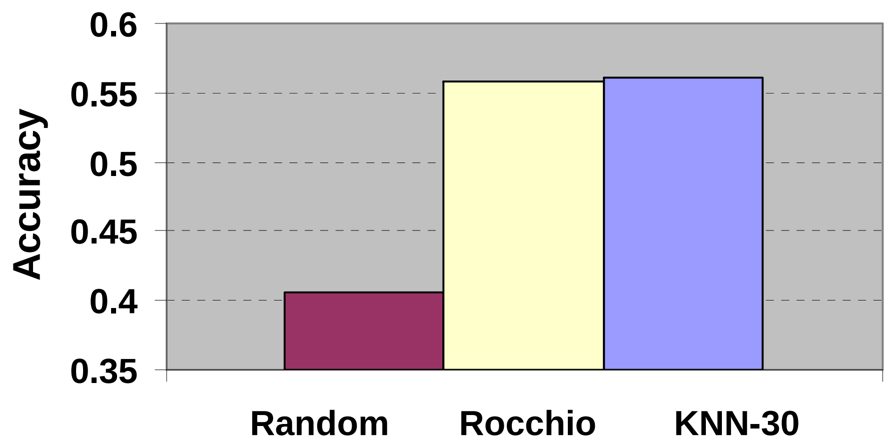
- **Rocchio/TFIDF Centroid [1971]**
  Create a "TfIdf centroid" for each user in Address Book. For testing, rank according to **cosine similarity** between test msg and each centroid.
- **Knn-30 [Yang & Chute, 1994]**
  Given a test msg, get 30 most similar msgs in training set. Rank according to "sum of similarities" of a given user on the 30-msg set.

Enron Data Preprocessing

- **ISI version of Enron**
  - Remove repeated messages and inconsistencies
- **Disambiguate Main Enron addresses**
  - List provided by Corrada-Emmanuel from UMass
- **Bag-of-words**
  - Messages were represented as the union of BOW of *body* and BOW of *subject*
- **Some stop words removed**
- **Self-addressed messages were removed**

Leak Results 1

Enron user	Random	Rocc	Knn-30	
			(sent)	(s+r)
rapp	0.236	0.470	0.547	0.459
hernandez	0.349	0.226	0.247	0.353
pereira	0.459	0.490	0.450	0.465
dickson	0.462	0.627	0.641	0.659
lavorato	0.463	0.697	0.668	0.637
hyatt	0.400	0.488	0.533	0.586
germany	0.352	0.570	0.620	0.588
white	0.389	0.648	0.626	0.616
whitt	0.426	0.478	0.522	0.563
zufferli	0.479	0.628	0.654	0.697
campbell	0.385	0.454	0.422	0.451
geaccone	0.367	0.413	0.423	0.420
hyvl	0.455	0.523	0.467	0.436
giron	0.444	0.551	0.588	0.616
horton	0.460	0.646	0.604	0.615
derrick	0.454	0.784	0.758	0.668
kaminski	0.471	0.711	0.753	0.739
hayslett	0.304	0.547	0.561	0.551
corman	0.466	0.782	0.728	0.695
Kitchen	0.300	0.424	0.379	0.415
Average	0.406	0.558	0.560	0.561

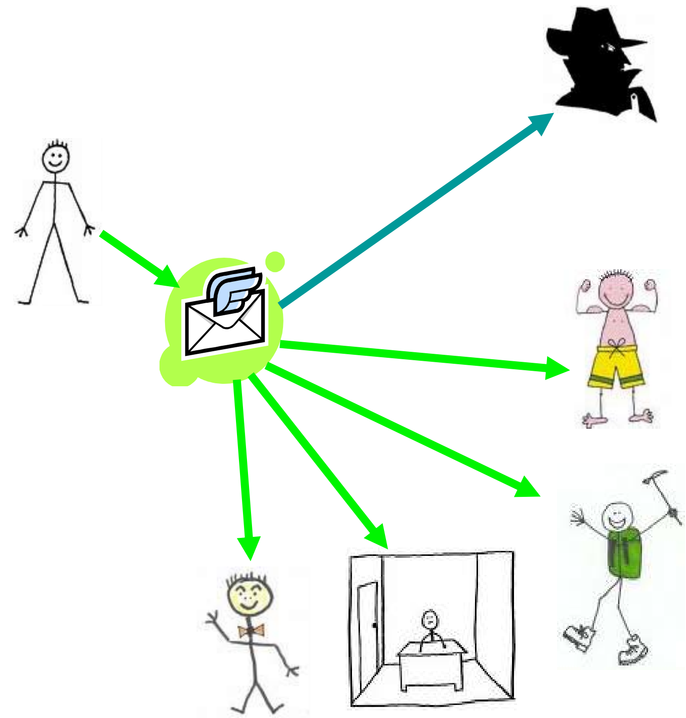


Average Accuracy in 10 trials:

On each trial, a different set of outliers is generated

Using Network Features

1. **Frequency features**
   – Number of received, sent and sent+received messages (from this user)
2. **Co-Occurrence Features**
   – Number of times a user co-occurred with all other recipients.
3. **Auto features**
   – For each recipient R, find Rm (=address with max score from 3g-address list of R), then use score(R)-score(Rm) as feature.



Using Network Features

1. **Frequency features**
   - Number of received, sent and sent+received messages (from this user)
2. **Co-Occurrence Features**
   - Number of times a user co-occurred with all other recipients.
3. **Auto features**
   - For each recipient R, find Rm (=address with max score from 3g-address list of R), then use score(R)-score(Rm) as feature.

Combine with text-only scores using perceptron-based reranking, trained on simulated leaks

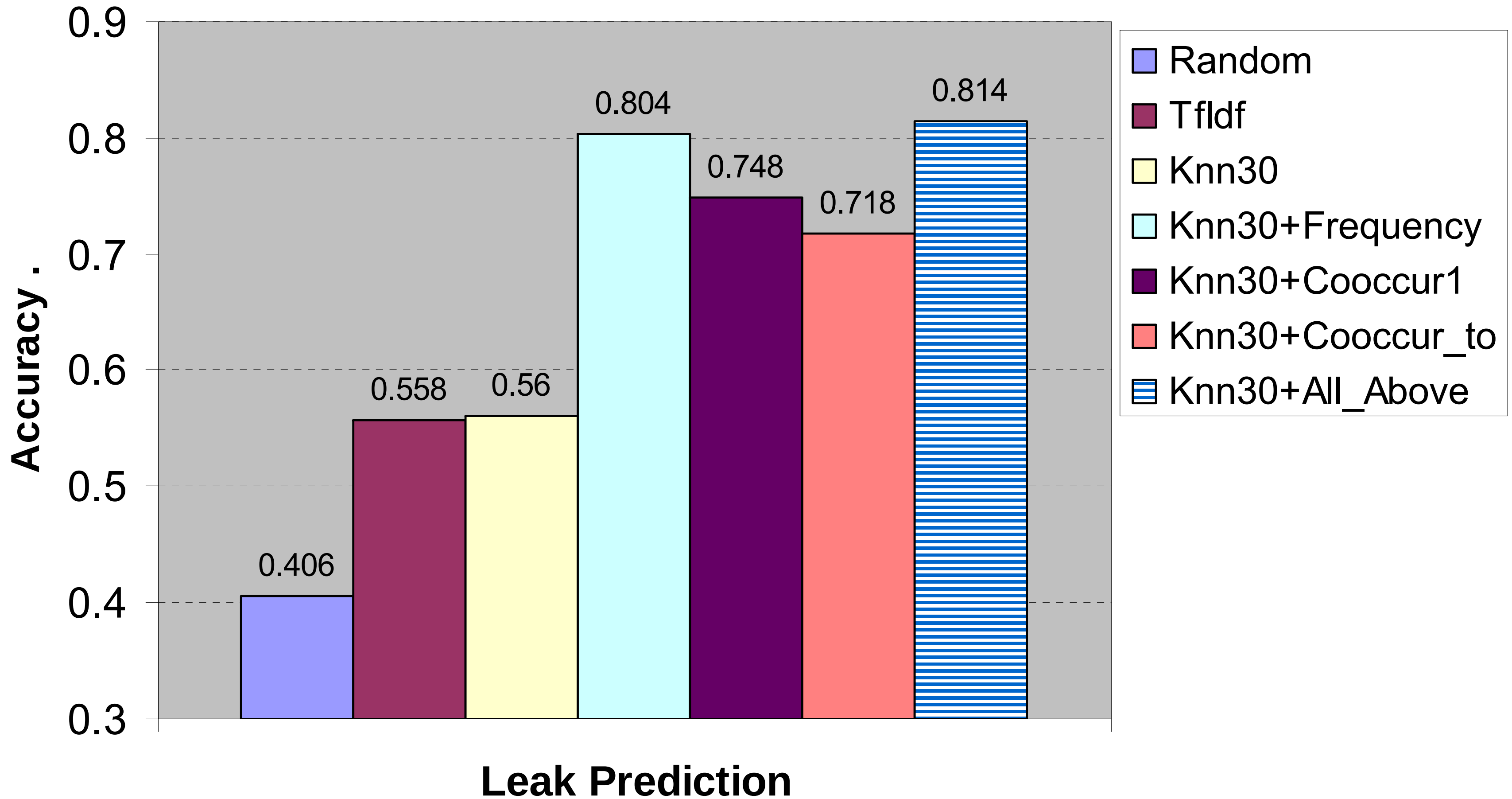
$$Score = \alpha_0 F_0 + \sum_i \alpha_i F_i$$

Text-based Feature

Network Features

Email Leak Results

[Carvalho & Cohen, SDM-07]



Finding Real Leaks in Enron

- How can we find it?
  - Grep for “mistake”, “sorry” or “accident”
  - Note: must be *from* one of the Enron users

“Sorry. Sent this to you by mistake.”, “I accidentally sent you this reminder”

- Found 2 good cases:
  1. Message germany-c/sent/930, message has 20 recipients, leak is alex.perkins@
  2. kitchen-l/sent items/497, it has 44 recipients, leak is rita.wynne@

- Prediction results:
  - The proposed algorithm was able to find these two leaks

Not the
only problem
when addressing emails…

Sometimes people just… forget an intended recipient

- Particularly in large organizations, it is not uncommon to forget to CC an important collaborator: a manager, a colleague, a contractor, an intern, etc.

[Carvalho & Cohen, ECIR-2008]

- **More frequent than expected** (from Enron Collection)
  - at least 9.27% of the users have forgotten to add a desired email recipient.
  - At least 20.52% of the users were not included as recipients (even though they were intended recipients) in at least one received message.

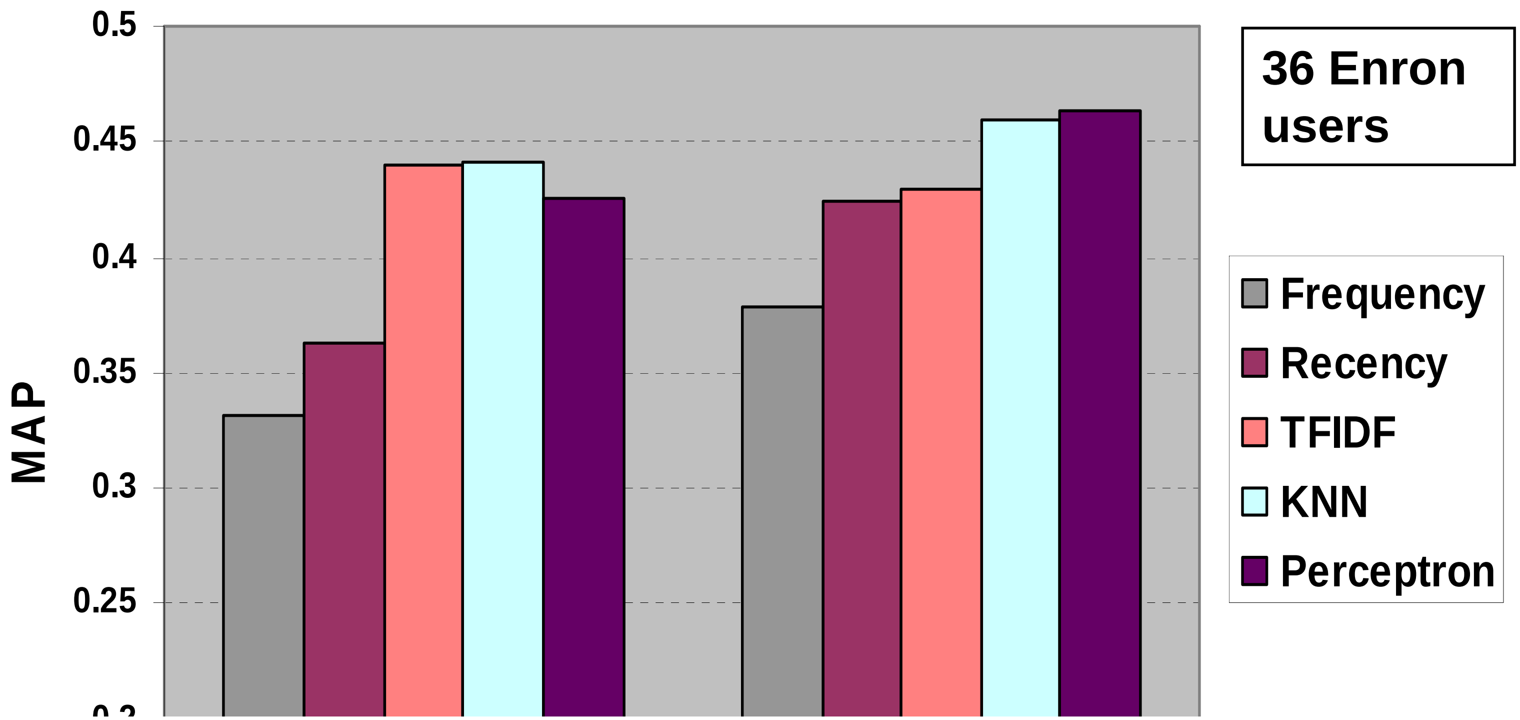
- Cost of errors in task management can be high:
  - Communication delays, Deadlines can be missed
  - Opportunities wasted, Costly misunderstandings, Task delays

Data and Features

- Two Ranking problems:
  - Predicting TO+CC+BCC
  - Predicting CC+BCC
- Easy to obtain labeled data

- Features
  - Textual: Rocchio (TfIdf) and KNN
  - Network (from Email Headers)
    - Frequency
      - # messages received and/or sent (from/to this user)
    - Recency
      - How often was a particular user addressed in the last 100 msgs
    - Co-Occurrence
      - Number of times a user co-occurred with all other recipients. Co-occurr means “two recipients were addressed in the same message in the training set”

Email Recipient Recommendation



	MAP	MRR	R-Prec	P@5	P@10
TOCCBCC	0.441	0.516	0.398	0.225	0.157
CCBCC	0.459	0.540	0.425	0.239	0.156

44000+ queries

Avg: ~1267 q/user

[Carvalho & Cohen, ECIR-08]

Rank Aggregation (Data Fusion)

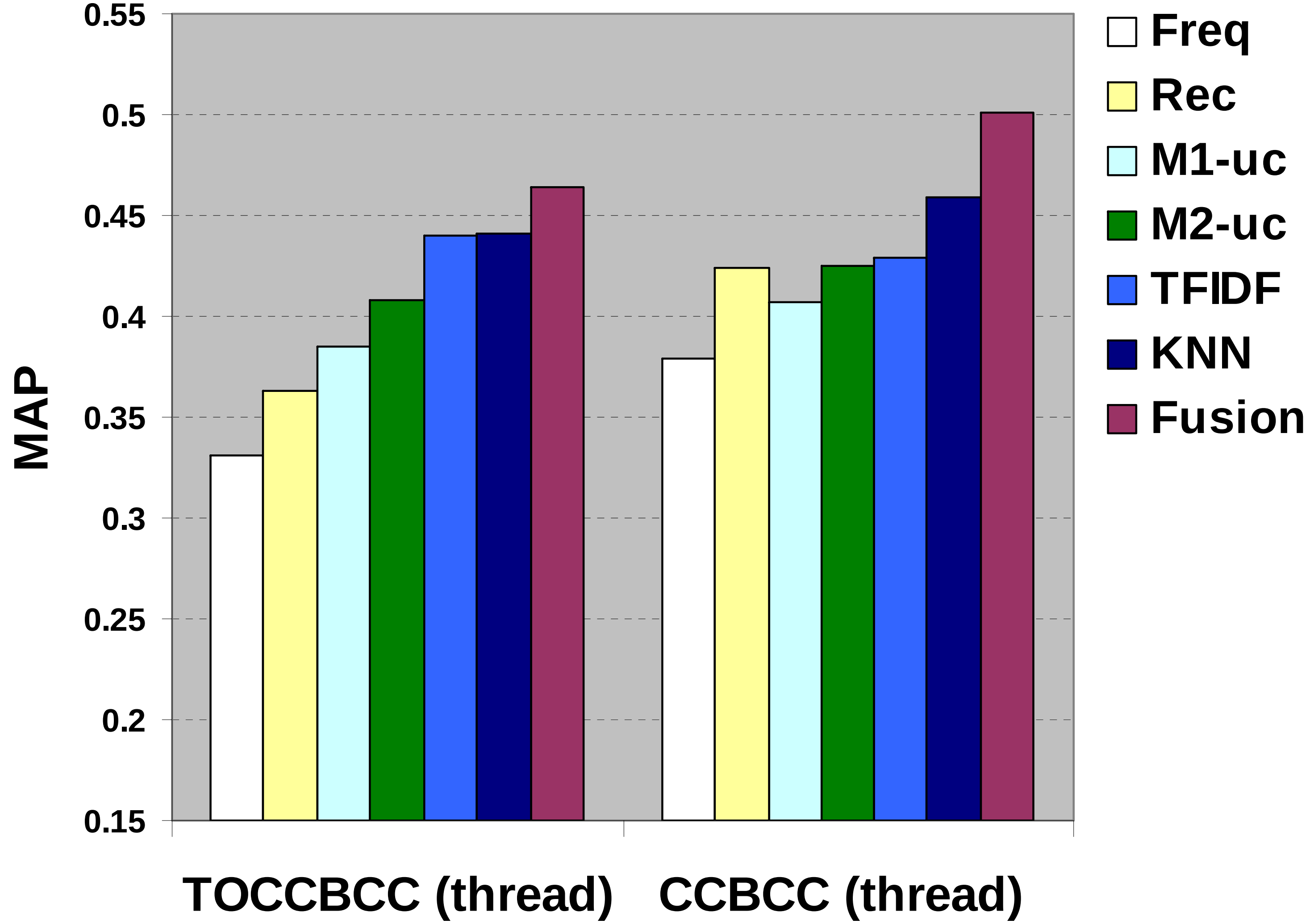
Ranking combined by Reciprocal Rank:

$$RR(d_i) = \sum_{q \in Rankings} \frac{1}{rank_q(d_i)}$$

Table 4. MAP values for model aggregations with Reciprocal Rank. The $*$ and $**$ symbols indicate statistically significant results over the Knn baseline.

Task		Freq	Recency	TFIDF	M2-uc
TOCCBCC	Knn $\odot$	0.417**	0.432	**0.457****	0.444
	Knn $\odot$ TFIDF $\odot$	0.455**	**0.464****	—	0.461**
Baseline: Knn	Knn $\odot$ TFIDF $\odot$ Rec $\odot$	0.451**	—	—	**0.470****
MAP = 0.441	Knn $\odot$ TFIDF $\odot$ Rec $\odot$ M2-uc $\odot$	**0.464****	—	—	—
CCBCC	Knn $\odot$	0.455	0.470	0.462	**0.474***
	Knn $\odot$ M2-uc $\odot$	0.476**	**0.491****	0.482**	—
Baseline: Knn	Knn $\odot$ M2-uc $\odot$ Rec $\odot$	0.491**	—	**0.494****	—
MAP = 0.458	Knn $\odot$ M2-uc $\odot$ Rec$\odot$ TFIDF $\odot$	**0.501****	—	—	—

Rank Aggregation Results



Intelligent Email Auto-completion

[Carvalho & Cohen, ECIR-08]

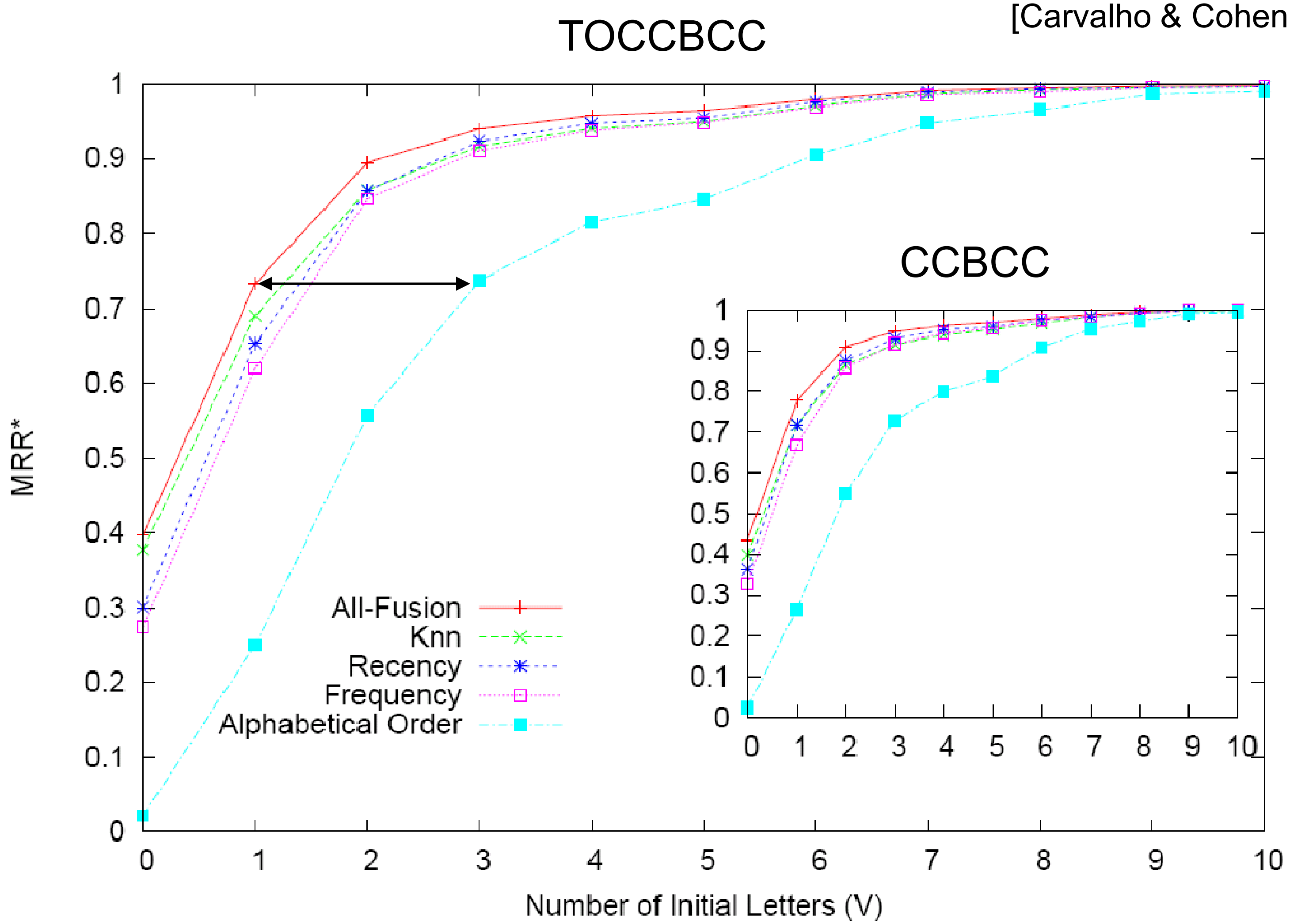


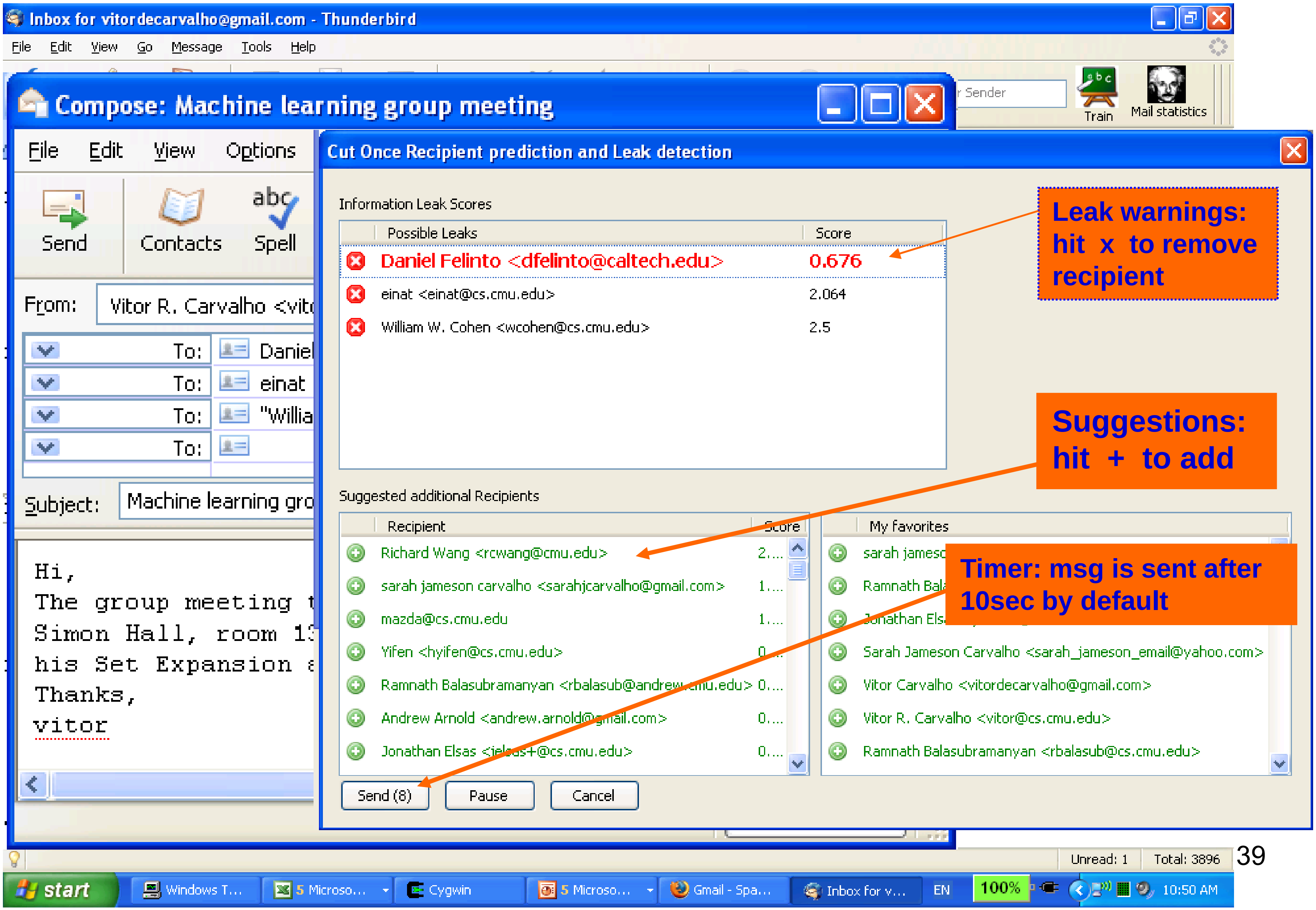
Fig. 1. Auto-completion performance for different number of initial letters V.

Intelligent Email Auto-completion

Table 5. Auto-completion Experiments. Performance values for different models and V values. Statistical significance relative to the previous column value is indicated with the symbols ** ($p < 0.01$) and * ($p < 0.05$).

V	Alpha	Freq	Rec	Knn	Fus	Δ(Knn-Rec)	Δ(Fus-Rec)	Δ(Fus-Knn)
	Primary Prediction							
0	0.022	0.274**	0.300**	0.377**	0.394**	25.542%	31.124%	4.447%
1	0.250	0.620**	0.653**	0.690**	0.731**	5.753%	11.893%	5.806%
2	0.557	0.846**	0.857	0.858	0.895**	0.078%	4.412%	4.331%
3	0.737	0.911**	0.923*	0.917	0.942**	-0.683%	2.001%	2.702%
	Secondary Prediction							
0	0.025	0.329**	0.364**	0.398*	0.436**	9.526%	19.927%	9.496%
1	0.265	0.668**	0.718**	0.717	0.777**	-0.125%	8.289%	8.424%
2	0.549	0.858**	0.875	0.865	0.910**	-1.189%	3.928%	5.178%
3	0.729	0.915**	0.929	0.915	0.946**	-1.558%	1.811%	3.423%

Mozilla Thunderbird plug-in (Cut Once)



Mozilla Thunderbird extension (Cut Once)

- Interested? **Just google:**
  “mozilla extension carnegie mellon”
  “email leak carnegie mellon”

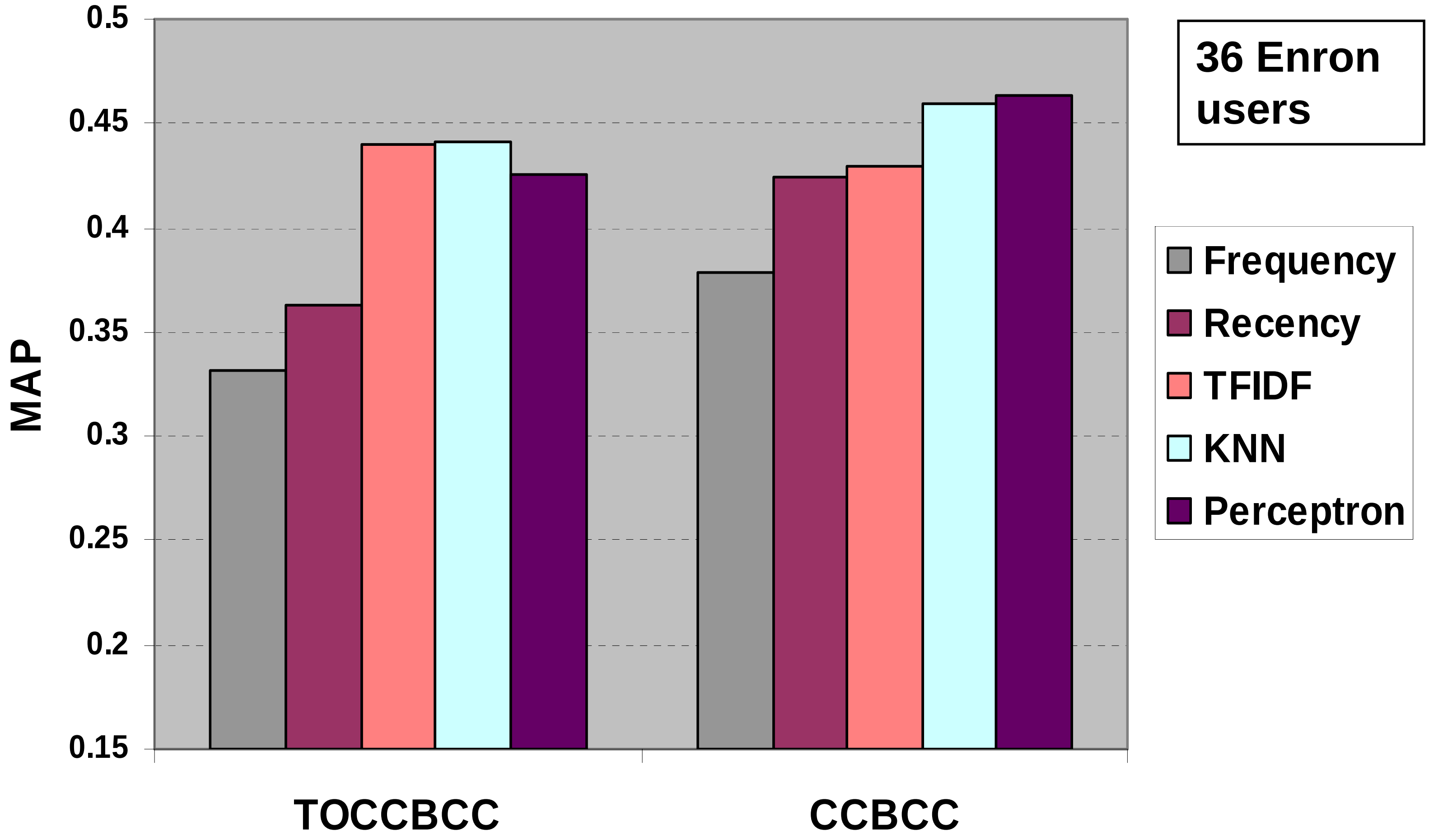
- From Mozilla website: 64 active daily users.

- User Study using *Cut Once*
  - Strong TFIDF preference
  - *write-then-address* behavior (instead of address-then-write)

Outline

1. Motivation √
2. Email Speech Acts √
   - Modeling textual intention in email messages
3. Intelligent Email Addressing √
   - Preventing information leaks
   - Ranking potential recipients
4. Fine-tuning User Ranking Models
   - Ranking in two optimization steps

Email Recipient Recommendation



Learning to Rank

- Can we do better ranking?
  - Learning to Rank: machine learning to improve ranking
  - Recently proposed feature-based ranking methods:
    - RankSVM [Joachims, KDD-02]
    - ListNet [Cao et al. , ICML-07]
    - RankBoost [Freund et al, 2003]
    - Perceptron Variations [Elsas, Carvalho & Carbonell, WSDM-08]
      - Online, scalable.
  - Learning to rank in 2 optimization steps
    - Pairwise-based ranking framework (like many of the above)

Pairwise-based Ranking

Rank q

d_1

d_2

d_3

d_4

d_5

$d_6 = (x_{16}, x_{26}, \ldots, x_{m6})$

...

d_T

Goal: induce a ranking function f(d) s.t.

$$d_i \succ d_j \Leftrightarrow f(d_i) > f(d_j)$$

We assume a linear function f

$$f(d_i) = \langle \mathrm{w}, d_i \rangle = w_1 x_{1i} + w_2 x_{2i} + \ldots + w_m x_{mi}$$

Constraints:

$$d_i \succ d_j \Leftrightarrow \langle w, d_i - d_j \rangle > 0$$

Pairwise-based Ranking

- Advantages
  1. Most classification methods can be easily adapted to the ranking problem
  2. This framework can be generalized to any graded relevance levels (e.g. definitely relevant, somewhat relevant, non-relevant).
  3. In many practical scenarios, it is easier to obtain large amounts of pairwise preference data [Joachims:2002]
  4. Also, there is evidence that pairwise preference judgment is easier for accessors [Carterette, 2008].

Pairwise-based Ranking

- Disadvantages
  - One single human labeling error creates many outliers - since pairs of documents of different labels are used as instances in the learning scheme.
  - Discrimination of multi-level labeling scheme (1-2, 2-3, versus 1-5)
  - In real labeled ranking datasets, many of the documents are unjudged and typically considered non-relevant for pairwise learning algorithms.

Method 1: Ranking with Perceptrons

- Nice convergence and mistake bounds
  - bound on the number of misranks
- Online, fast and scalable

[Collins, 2002; Gao et al, 2005]

[Elsas, Carvalho & Carbonell, 2008]

- Many variants
  - Voting, averaging, committee, pocket, etc.
  - General update rule:

$$W^{t+1} = W^{t} + [d_R - d_{NR}]$$

  - Here: Averaged version of perceptron

Method 2: Rank SVM

[Joachims, KDD-02],

[Herbrich et al, 2000]

$$\min_{w} L_{ranksvm} = \frac{1}{2}\|w\|^2 + C\sum_{i \in RP} \varepsilon_i,$$

subject to $\varepsilon_{\mathrm{i}} \geq 0, \langle w, d_R - d_{NR} \rangle \geq 1 - \varepsilon_i, RP = \{(d_R, d_{NR})\}$

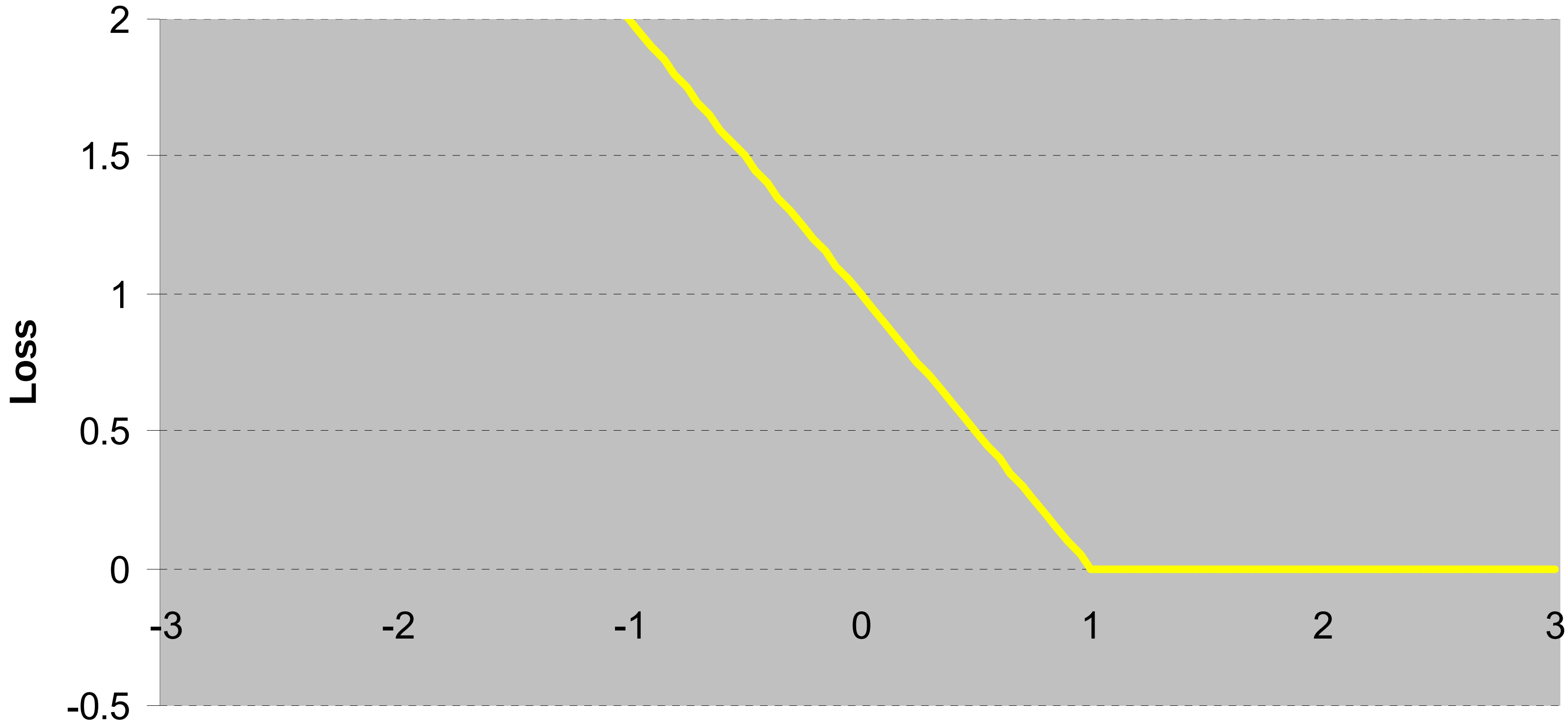
Equivalent to:

$$\boxed{\min_{w} L_{ranksvm} = \lambda\|w\|^2 + \sum_{RP} [1 - \underline{\langle w, d_R - d_{NR} \rangle}]_+ \quad \text{, where } \lambda = \frac{1}{2\mathrm{C}}.}$$

- Minimizing number of misranks (hinge loss approx.)
- Equivalent to maximizing AUC
- Lowerbound on MAP, precision@K, MRR, etc.

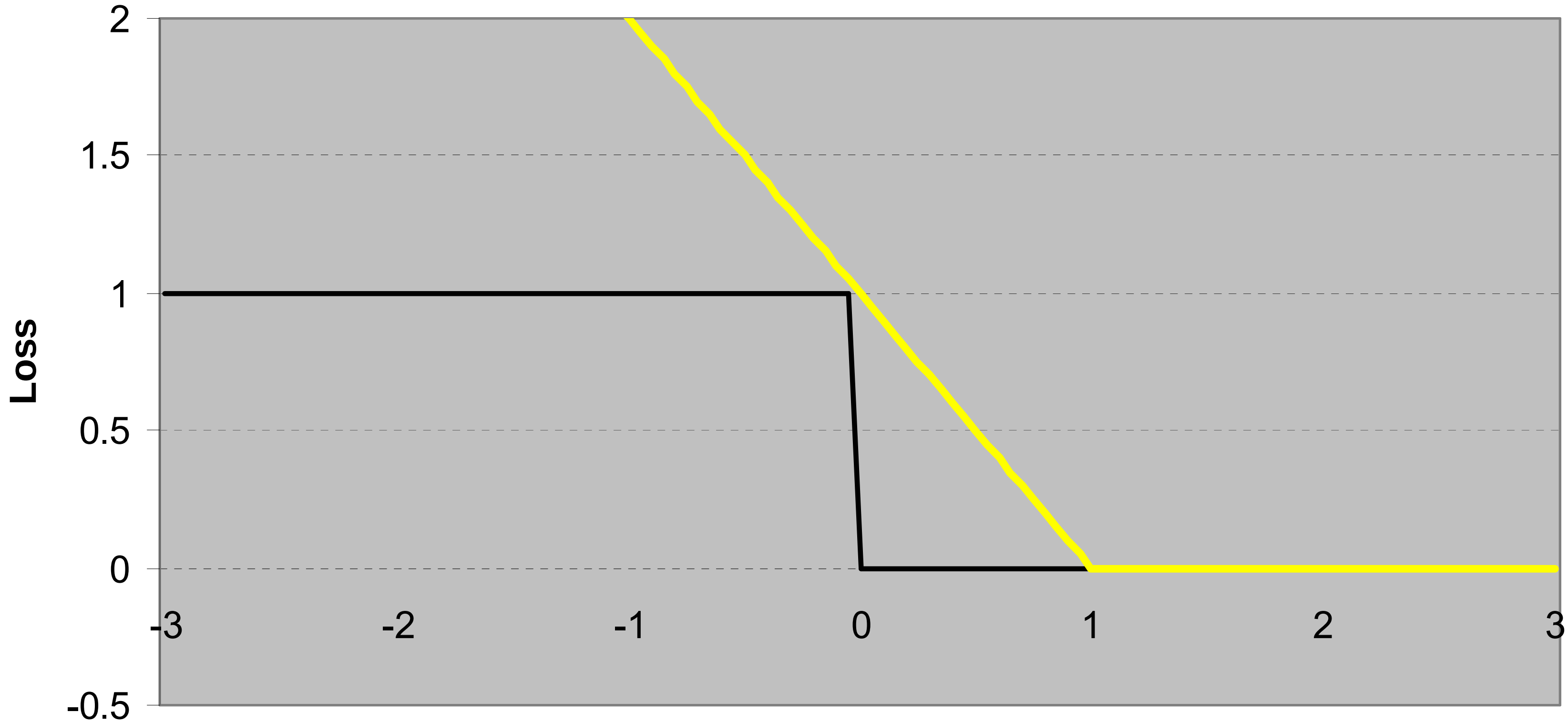
[Elsas, Carvalho & Carbonell, WSDM-08]

Loss Function



$$x = \langle w, d_R - d_{NR} \rangle$$

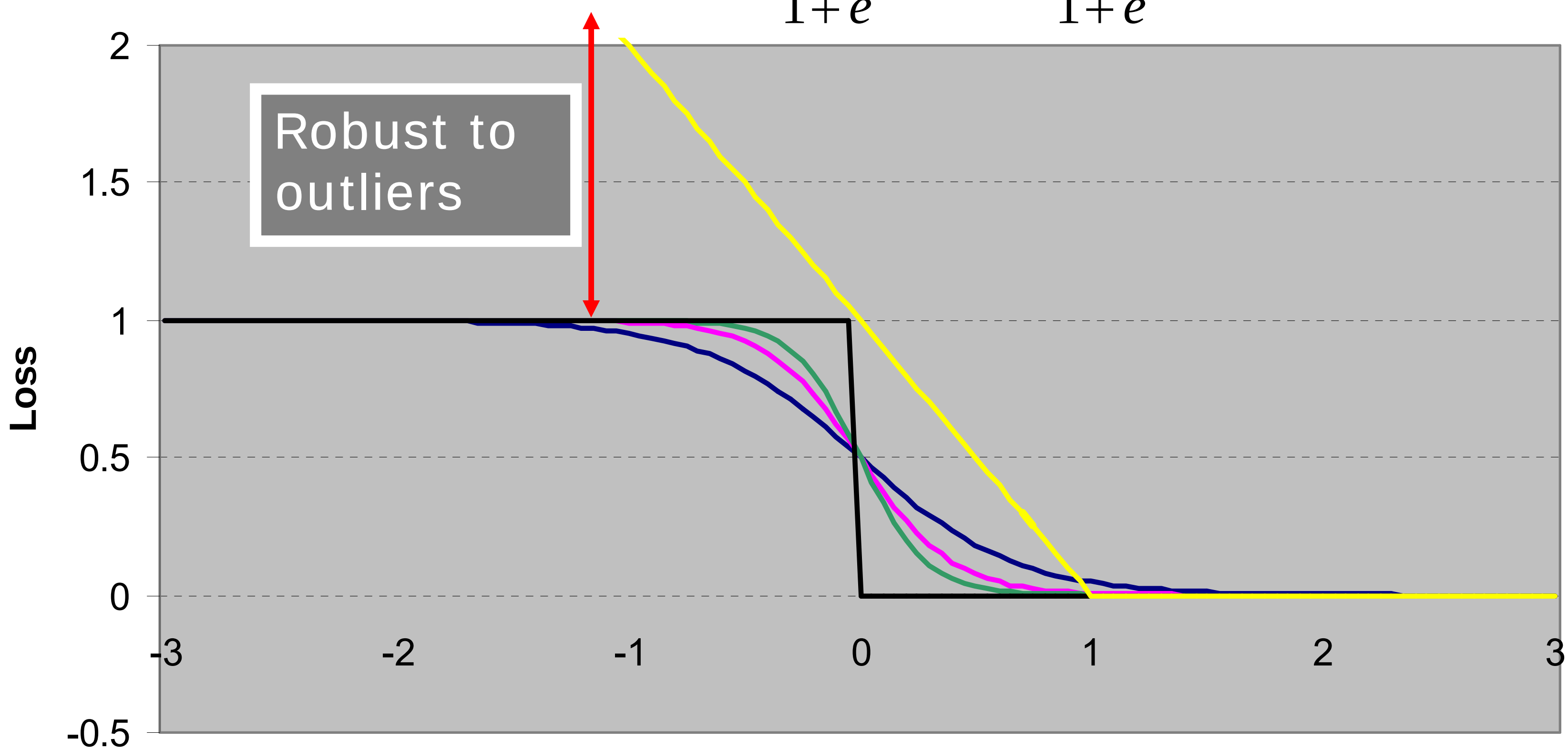
Loss Function



$$x = \langle w, d_R - d_{NR} \rangle$$

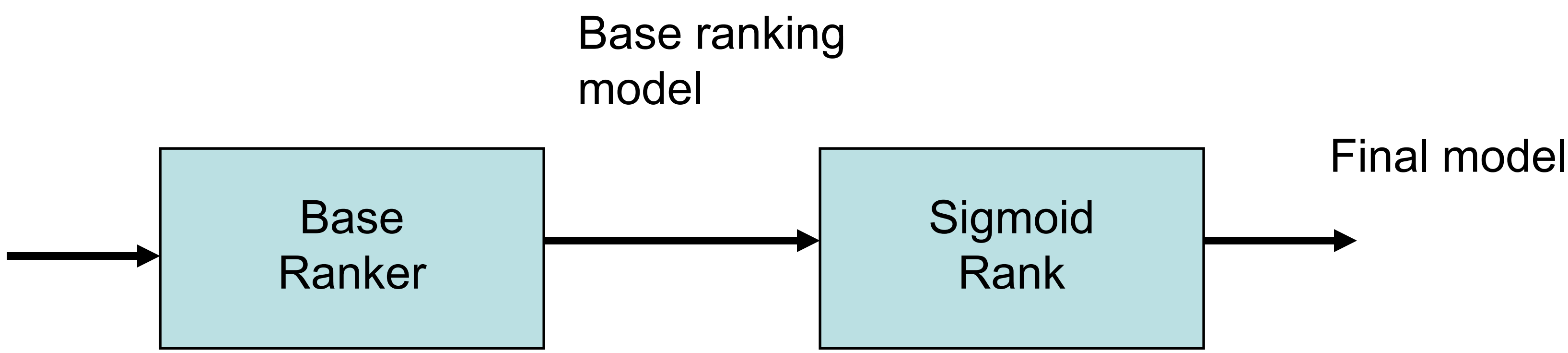
Loss Function

$$\frac{e^{-x\sigma}}{1+e^{-x\sigma}} = 1 - \frac{1}{1+e^{-x\sigma}} = 1 - sigmoid(x\sigma)$$



$$x = \langle w, d_R - d_{NR} \rangle$$

Fine-tuning Ranking Models



e.g.,
RankSVM,
Perceptron,
ListNet,

etc.

Non-convex:

Minimize (a very close approximation for) the empirical error: number of misranks

Robust to outliers (label noise)

Learning

- SigmoidRank Loss

$$sigmoid(x) = \frac{1}{1+e^{-\sigma x}}$$

$$\boxed{\min_{w} L_{SigmoidRank} = \lambda \|w\|^2 + \sum_{RP} [1 - sigmoid(\langle w, d_R - d_{NR} \rangle)]}$$

- Learning with Gradient Descent

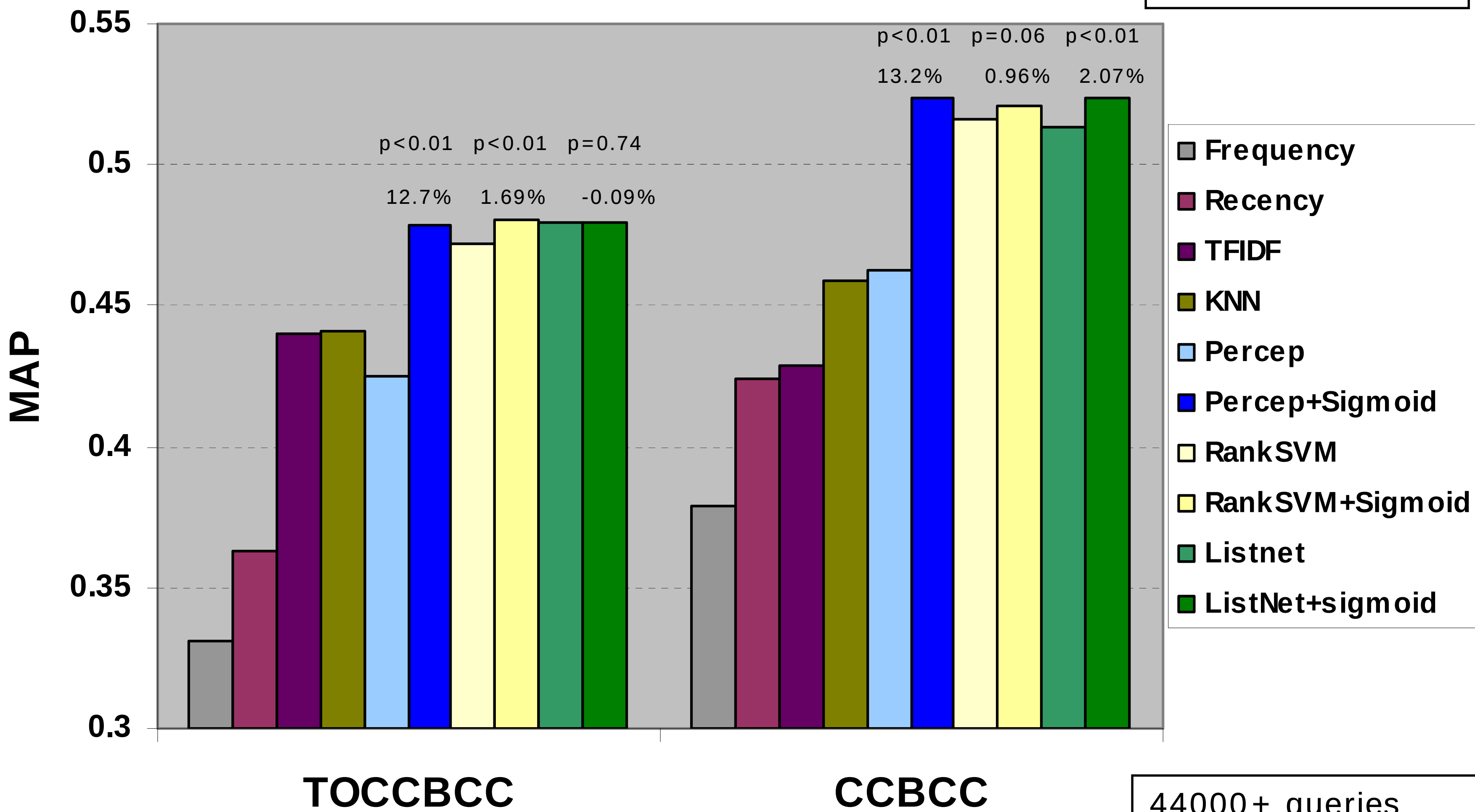
$$w^{(k+1)} = w^{(k)} + \eta_k \Delta w^{(k)}$$

$$\Delta w^{(k)} = -\nabla L_{rankSigmoid}(w^{(k)})$$

$$\nabla L_{rankSigmoid}(w^{(k)}) = 2\lambda w - \sum_{RP} \sigma\, sigmoid(\langle w, d_R - d_{NR} \rangle)[1 - sigmoid(\langle w, d_R - d_{NR} \rangle)]$$

Email Recipient Results

36 Enron users

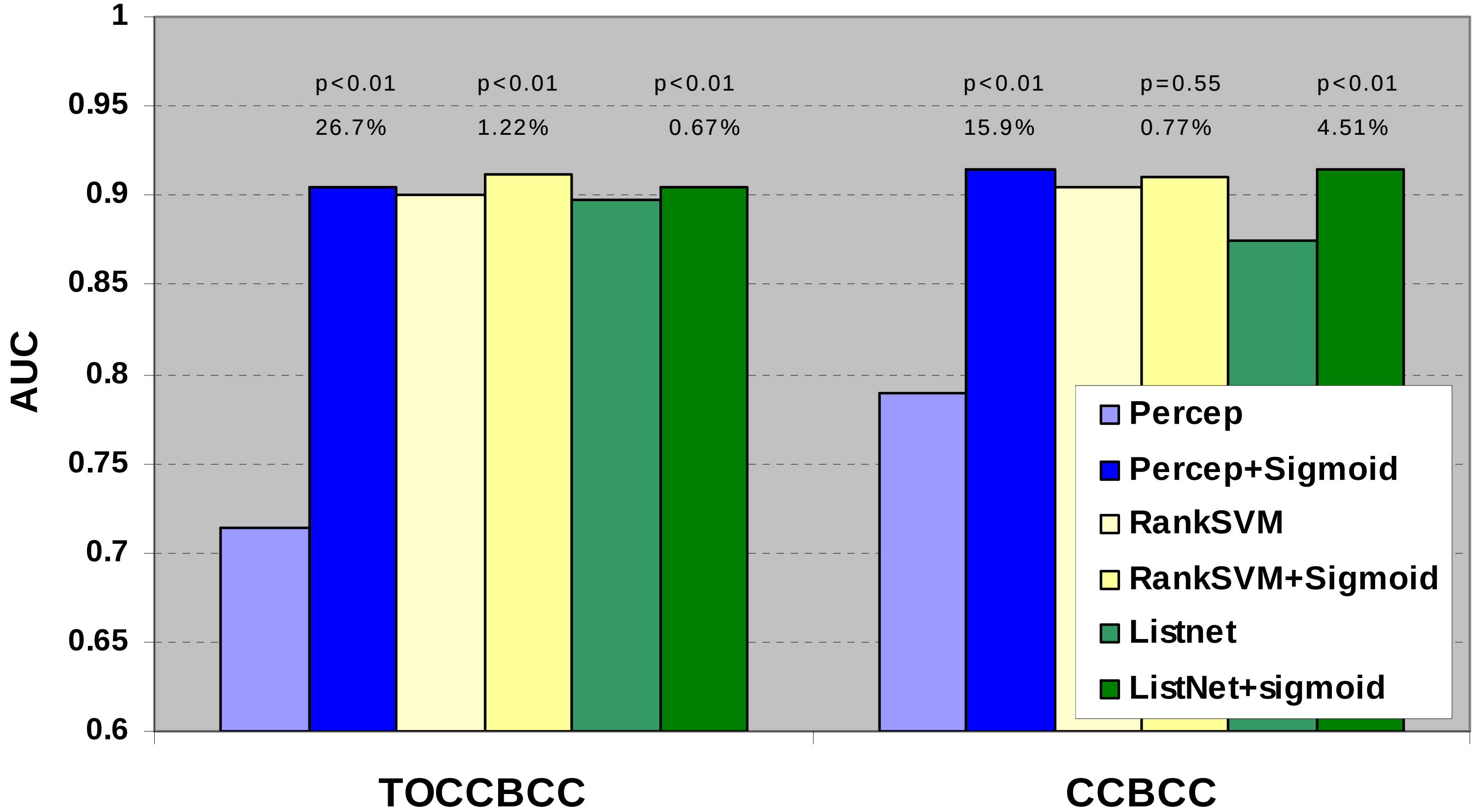


44000+ queries

Avg: ~1267 q/user

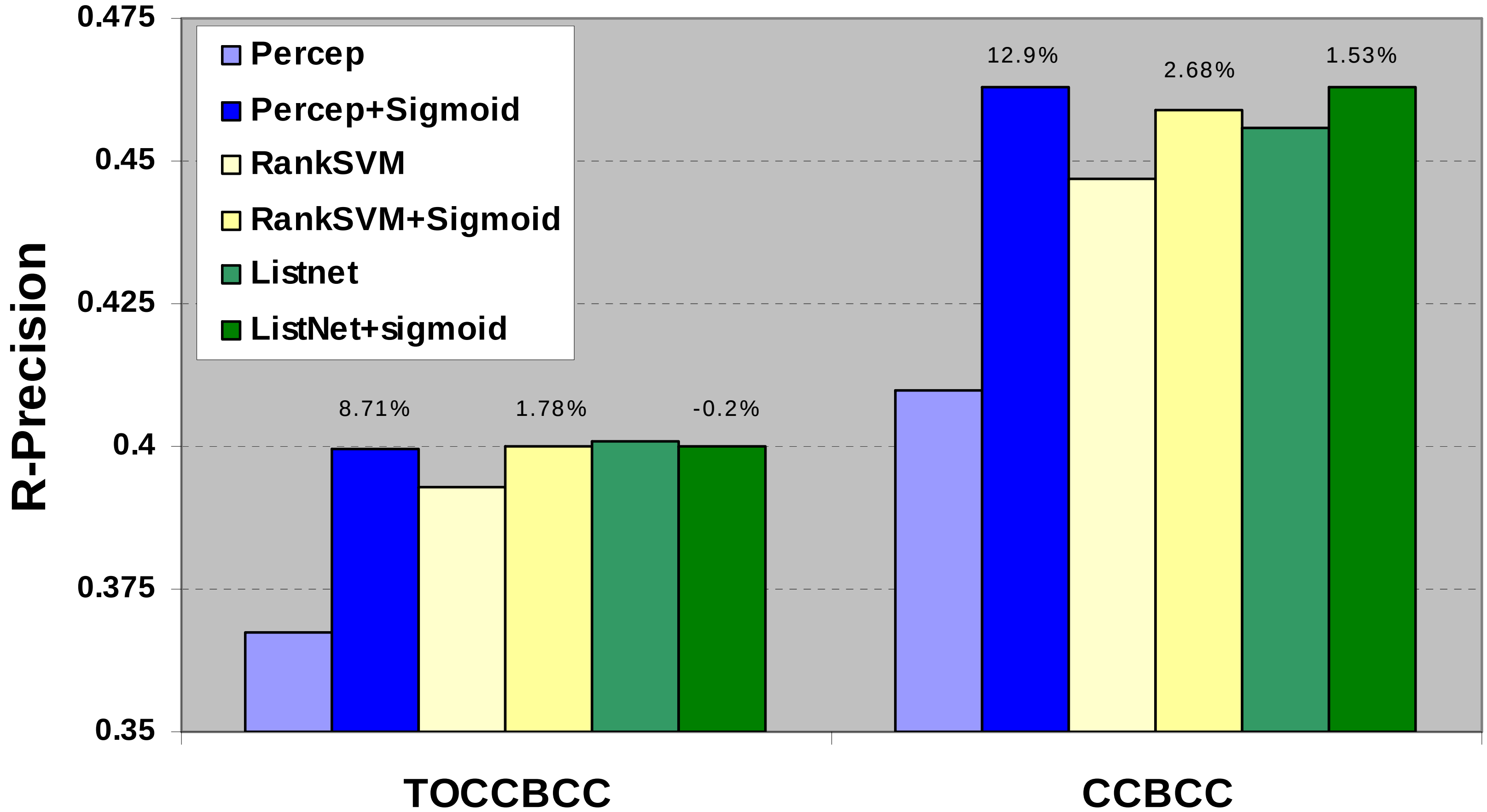
Email Recipient Results

36 Enron users



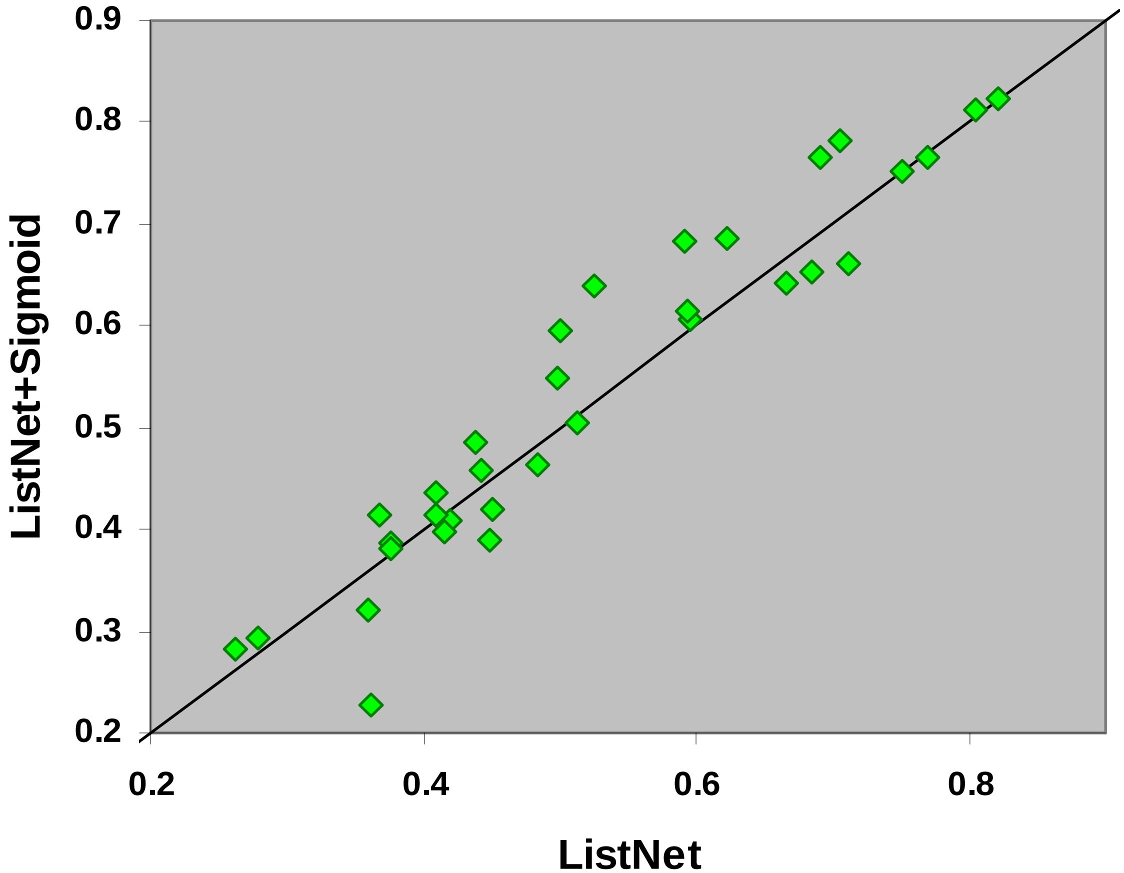
Email Recipient Results

36 Enron users



Email Recipient Results

36 Enron users

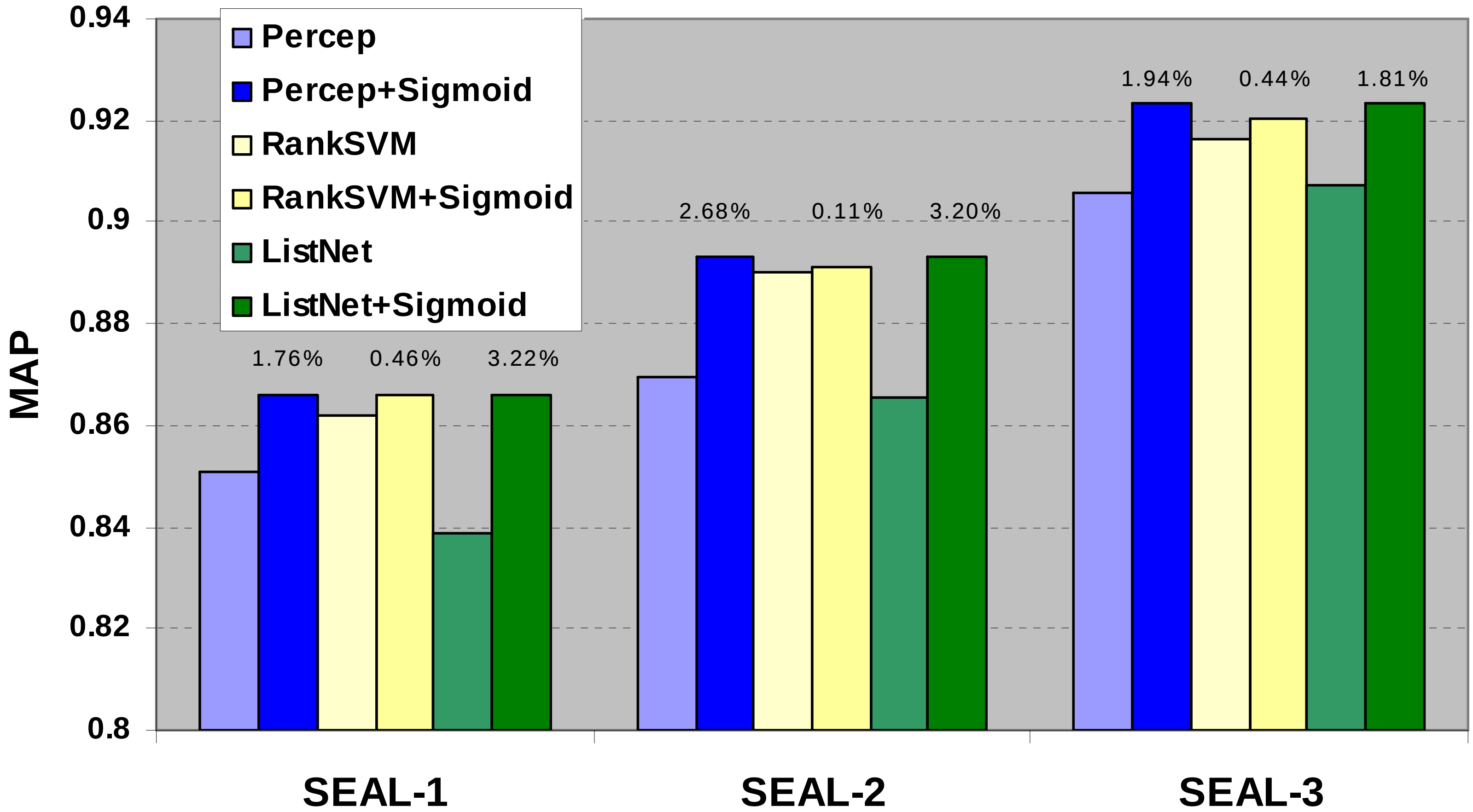


MAP values

CCBCC task

Set Expansion (SEAL) Results

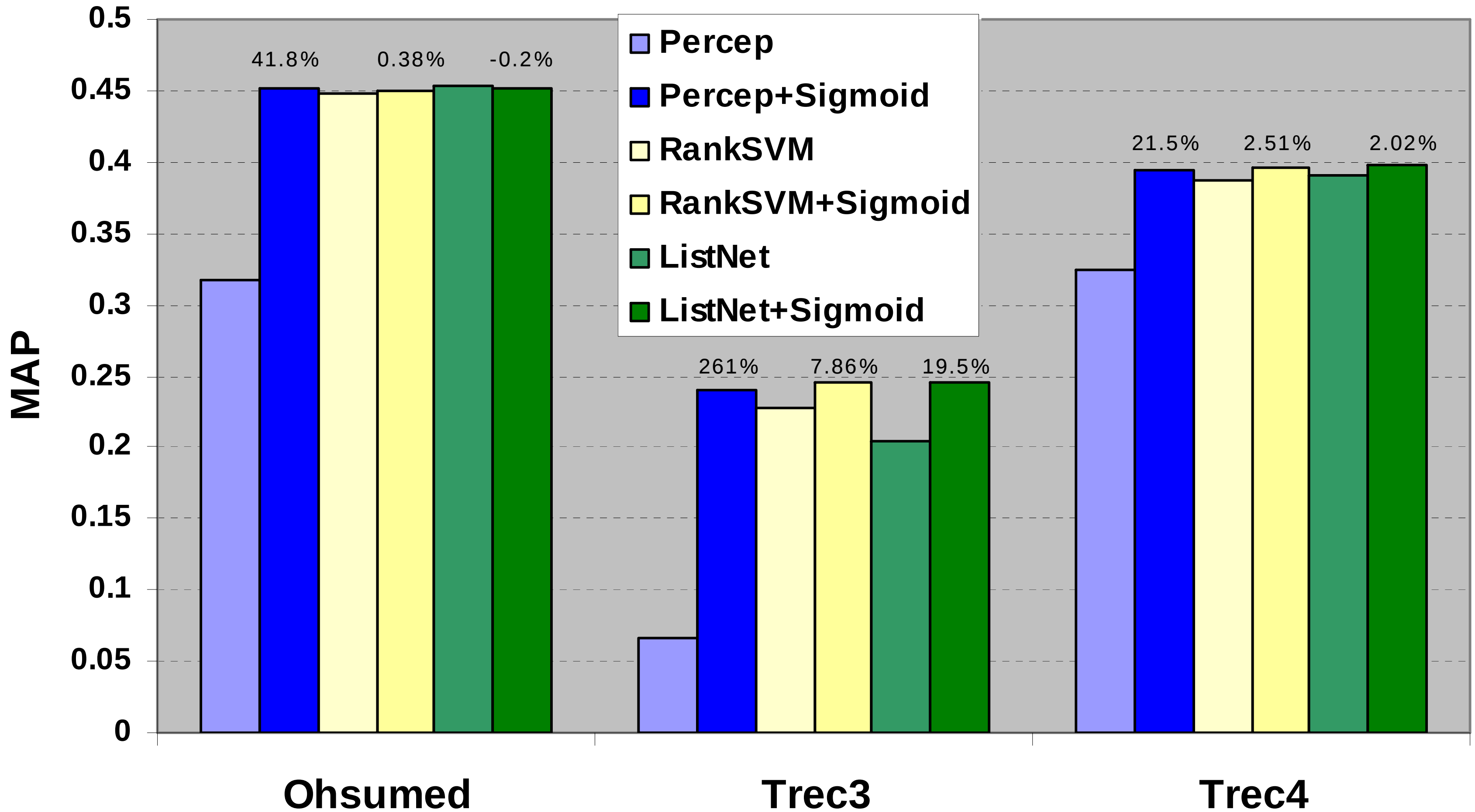
[Wang & Cohen, ICDM-2007]



[18 features, ~120/60 train/test splits, ~half relevant]

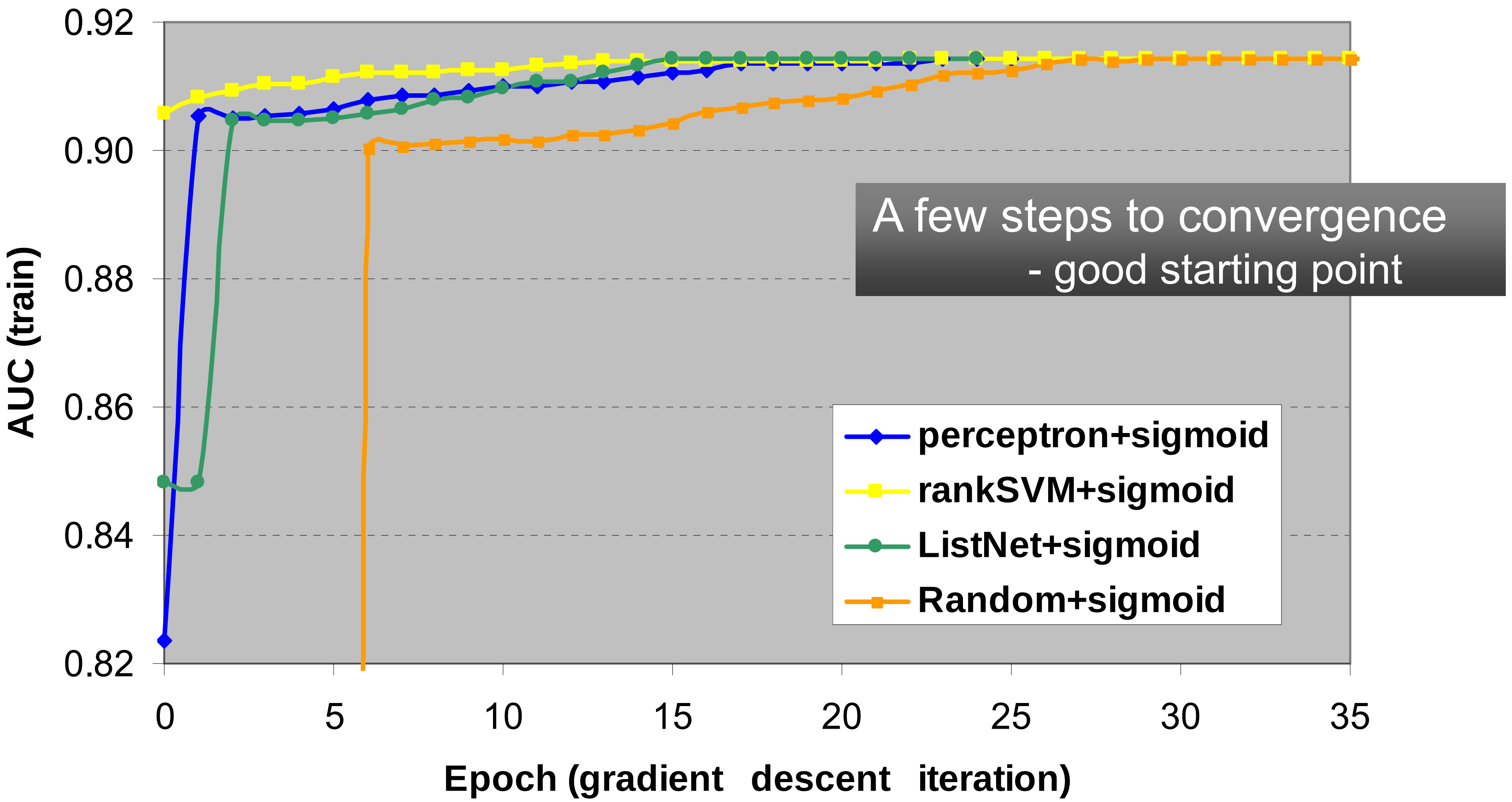
Letor Results

[Liu et al, SIGIR-LR4IR 2007]



[#queries/#features: (106/25) (50/44) (75/44)]

Learning Curve



TOCCBCC Enron: user lokay-m

Conclusions

- Email acts
  - Managing/tracking commits, requests... (semi) automatically
- Preventing User's Mistakes
  - *Email Leaks* (accidentally adding non-intended recipients)
  - Recipient prediction (forgetting intended recipients)
  - Mozilla Thunderbird extension
- Ranking in two-optimization steps
  - Robust to outliers (when compared to convex losses)
  - Closer approximation minimizing number of misranks (empirical risk minimization framework)
  - Fine-tune any base learner in few steps - good starting point

Related Work 1

- Email acts:
  - Speech Act Theory [Austin, 1962;Searle,1969]
  - Email classification: spam, folder, etc.
  - Dialog Acts for Speech Recognition, Machine Translation, and other dialog-based systems. [Stolcke et al., 2000] [Levin et al., 03]
    - Typically, 1 act per utterance (or sentence) and more fine-grained taxonomies, with larger number of acts.
    - Email is new domain
  - Winograd's Coordinator (1987)
    - *users manually annotated email with intent.*
  - Related applications:
    - Focus message in threads/discussions [Feng et al, 2006], Action-items discovery [Bennett & Carbonell, 2005], Activity classification [Dredze et al., 2006], Task-focused email summary [Corsten-Oliver et al, 2004], Predicting Social Roles [Leusky, 2004], etc.

Related Work 2

- Email Leak
  - [Boufaden et al., 2005]
    - *proposed a privacy enforcement system to monitor specific privacy breaches (student names, student grades, IDs).*

- Recipient Recommendation
  - [Pal & McCallum, 2006], [Dredze et al., 2008]
    - *CC Prediction problem, Recipient prediction based on summary keywords*

  - Expert Search in Email
    - *[Dom et al.,2003], [Campbell et al,2003], [Balog & de Rijke, 2006], [Balog et al, 2006],[Soboroff, Craswell, de Vries (TREC-Enterprise 2005-06-07…)]*

Related Work 3

- Ranking in two-optimization steps
  - [Perez-Cruz et al, 2003]
    - *similar idea for the SVM-classification context (Empirical Risk Minimization)*
  - [Xu, Crammer & Schuurman, 2006][Krause & Singer, 2004][Zhan & Shen, 2005], etc.
    - *SVM robust to outliers and label noise*
  - [Collobert et al, 2006], [Liu et al, 2005]
    - *convexity tradeoff*

Thank you.