Statistical Models of Semantics and Unsupervised Language Discovery

Lecture #18

Introduction to Natural Language Processing

CMPSCI 585, Fall 2007

Andrew McCallum

Computer Science Department

University of Massachusetts Amherst

Including slides from Chris Manning, Dan Klein, Rion Snow & Patrick Pantel.

Attachment Ambiguity

- Where to attach a phrase in the parse tree?
- *“I saw the man with the telescope.”*
  - What does “with a telescope” modify?
  - Is the problem AI complete? Yes, but…
  - Proposed simple structural factors
    - Right association [Kimball 1973]
      ‘low’ or ‘near’ attachment = ‘early closure’ of NP
    - Minimal attachment [Frazier 1978]
      (depends on grammar) = ‘high’ or ‘distant’ attachment = ‘late closure’ (of NP)

Attachment Ambiguity

- “The children ate the cake with a spoon.”
- “The children ate the cake with frosting.”

- “Joe included the package for Susan.”
- “Joe carried the package for Susan.”

- *Ford, Bresnan and Kaplan (1982):*
  “It is quite evident, then, that the closure effects in these sentences are induced in some way by the choice of the lexical items.”

Lexical acquisition, semantic similarity

- Previous models give same estimate to all unseen events.
- Unrealistic - could hope to refine that based on semantic classes of words
- Examples
  - "Susan ate the cake with a durian."
  - "Susan had never eaten a fresh durian before."
  - Although never seen "eating pineapple" should be more likely than "eating holograms" because pineapple is similar to apples, and we have seen "eating apples".

An application: selectional preferences

- Most verbs prefer arguments of a particular type. Such regularities are called *selectional preferences* or *selectional restrictions*.
- “Bill drove a…” Mustang, car, truck, jeep

- Selectional preference strength: how strongly does a verb constrain direct objects
- “see” versus “unknotted”

Measuring selectional preference strength

- Assume we are given a clustering of (direct object) nouns. Resnick (1993) uses WordNet.

- Selectional association between a verb and a class

$$S(v) = D(P(C|v)||P(C) = \sum_{c} P(c|v) \log \frac{P(c|v)}{P(c)}$$

Proportion that its summand contributes to preference strength.

$$A(v,c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S(v)}$$

- For nouns in multiple classes, disambiguate as most likely sense: $A(v,n) = \max_{c \in \text{classes}(n)} A(v,c)$

Selection preference strength (made up data)

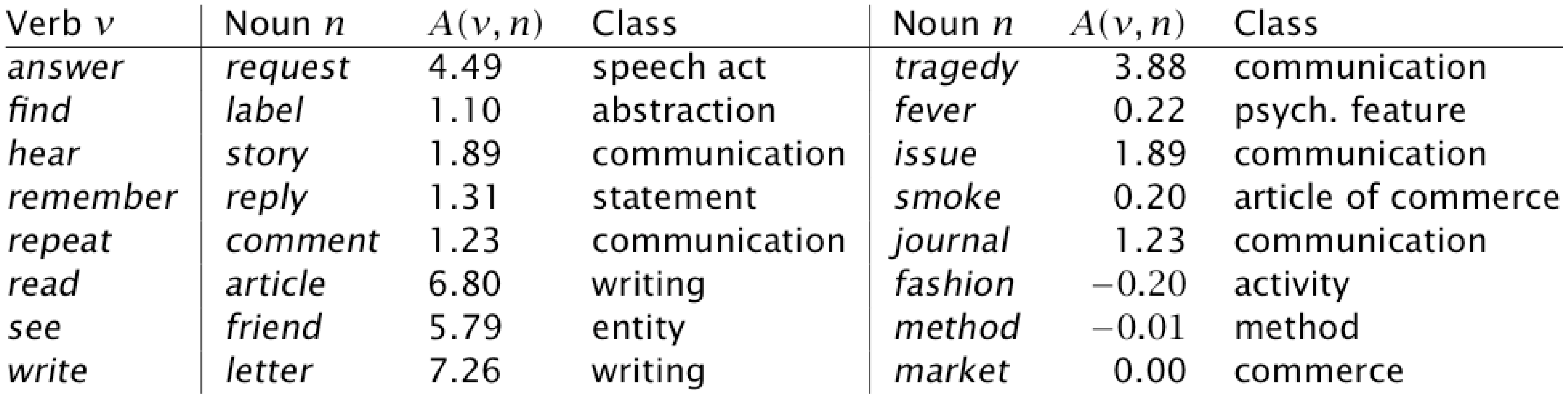
Noun class c	P(c)	P(c|eat)	P(c|see)	P(c|find)
people	0.25	0.01	0.25	0.33
furniture	0.25	0.01	0.25	0.33
food	0.25	0.97	0.25	0.33
action	0.25	0.01	0.25	0.01
SPS S(v)		**1.76**	**0.00**	**0.35**

A(eat, food) = 1.08

A(find, action) = -0.13

Selectional Preference Strength example

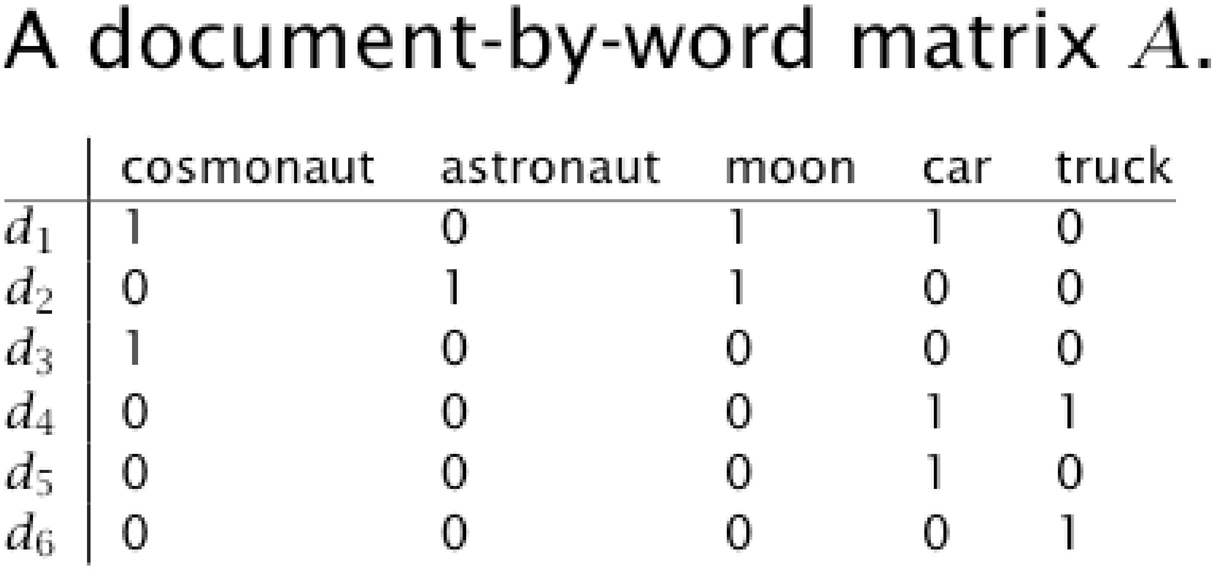
(Resnick, Brown corpus)



Verb v	Noun n	$A(v, n)$	Class	Noun n	$A(v, n)$	Class
answer	*request*	4.49	speech act	*tragedy*	3.88	communication
find	*label*	1.10	abstraction	*fever*	0.22	psych. feature
hear	*story*	1.89	communication	*issue*	1.89	communication
remember	*reply*	1.31	statement	*smoke*	0.20	article of commerce
repeat	*comment*	1.23	communication	*journal*	1.23	communication
read	*article*	6.80	writing	*fashion*	-0.20	activity
see	*friend*	5.79	entity	*method*	-0.01	method
write	*letter*	7.26	writing	*market*	0.00	commerce

But how might we measure word similarity for word classes?

- Vector spaces



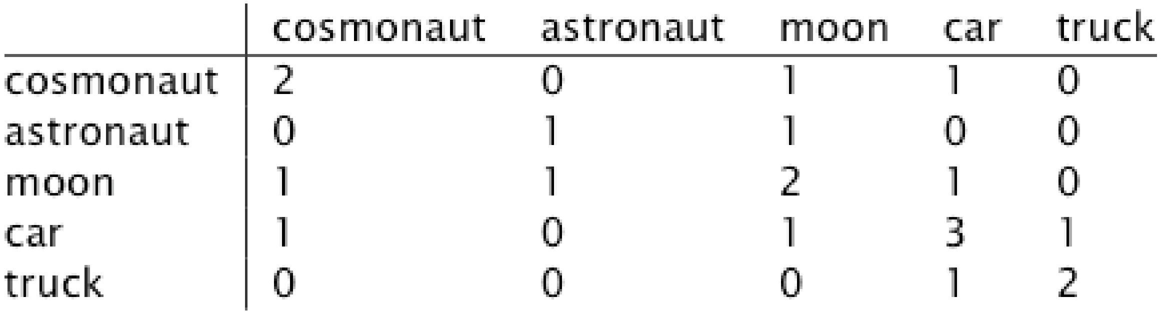
A document-by-word matrix A.

	cosmonaut	astronaut	moon	car	truck
d_1	1	0	1	1	0
d_2	0	1	1	0	0
d_3	1	0	0	0	0
d_4	0	0	0	1	1
d_5	0	0	0	1	0
d_6	0	0	0	0	1

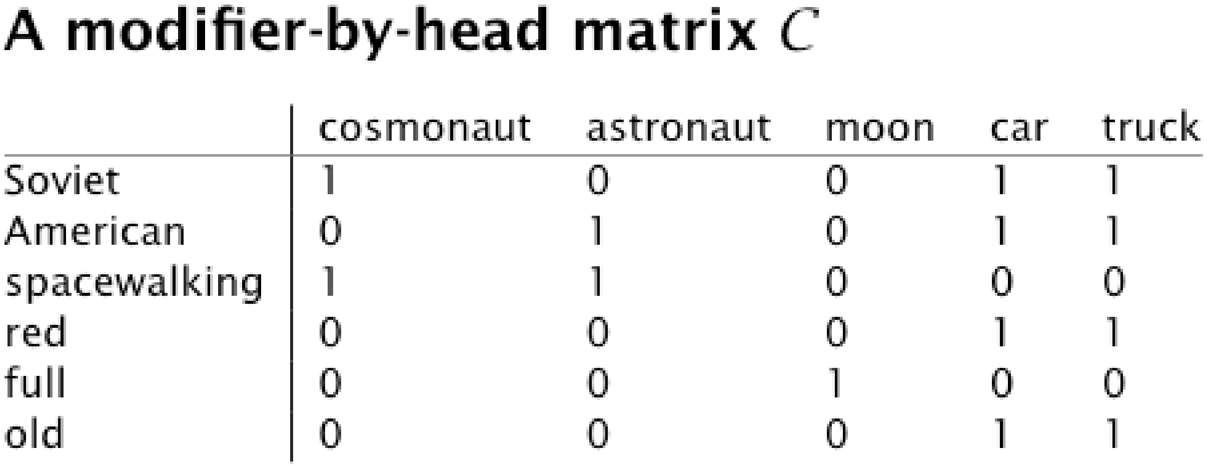
But how might we measure word similarity for word classes?

- Vector spaces

word-by-word matrix B



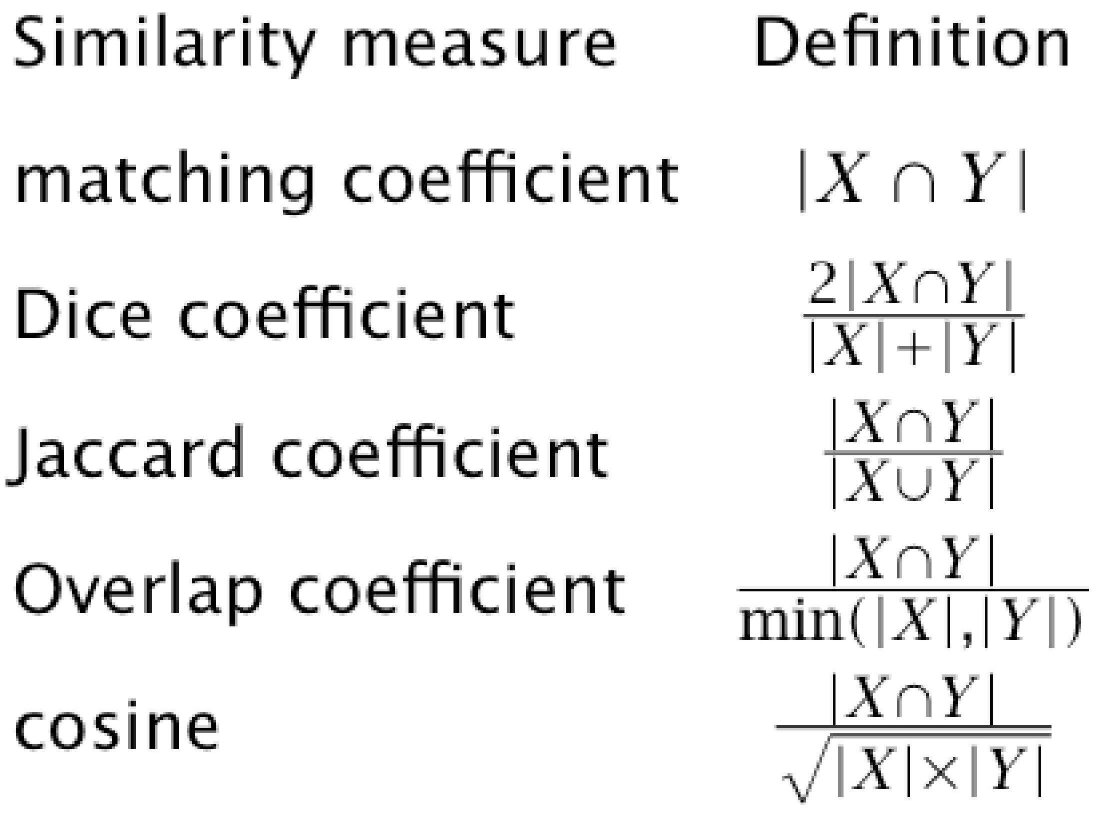
	cosmonaut	astronaut	moon	car	truck
cosmonaut	2	0	1	1	0
astronaut	0	1	1	0	0
moon	1	1	2	1	0
car	1	0	1	3	1
truck	0	0	0	1	2



A modifier-by-head matrix *C*

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

Similarity measures for binary vectors



Similarity measure	Definition
matching coefficient	$\|X \cap Y\|$
Dice coefficient	$\frac{2\|X \cap Y\|}{\|X\|+\|Y\|}$
Jaccard coefficient	$\frac{\|X \cap Y\|}{\|X \cup Y\|}$
Overlap coefficient	$\frac{\|X \cap Y\|}{\min(\|X\|,\|Y\|)}$
cosine	$\frac{\|X \cap Y\|}{\sqrt{\|X\| \times \|Y\|}}$

Cosine measure

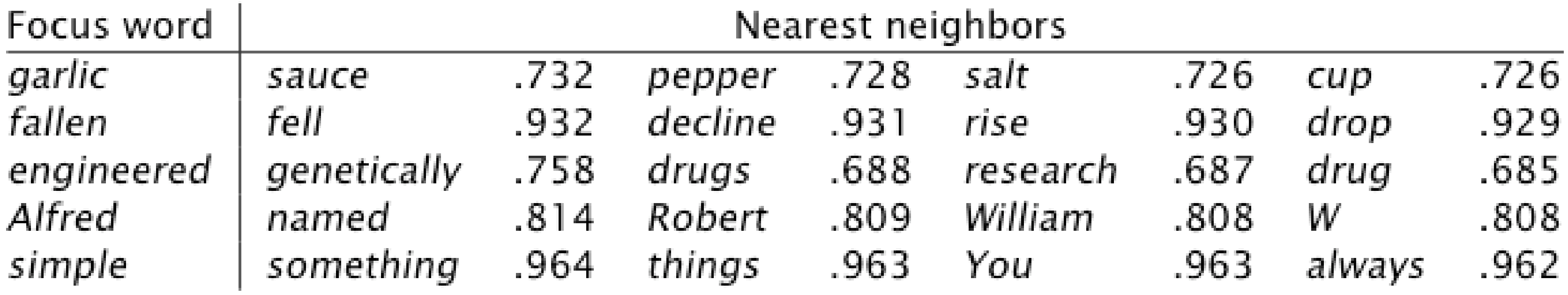
$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

maps vectors onto unit circle by dividing through

by lengths:

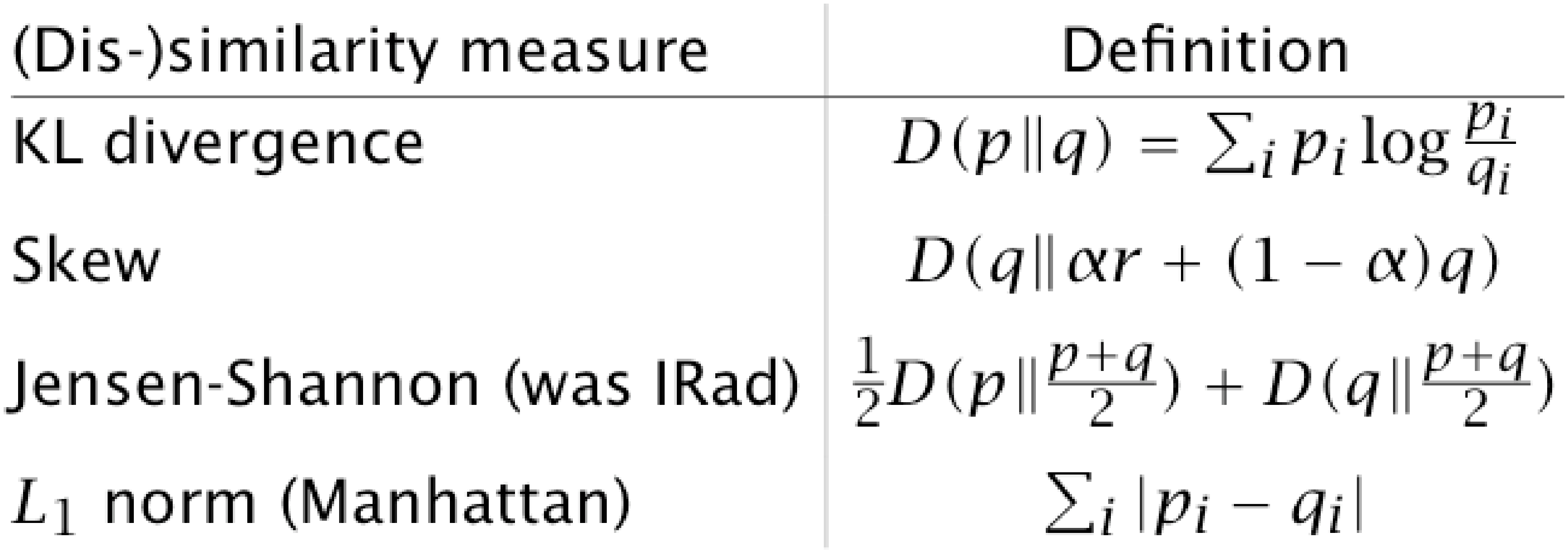
$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Example of cosine measure on word-by-word matrix on NYT



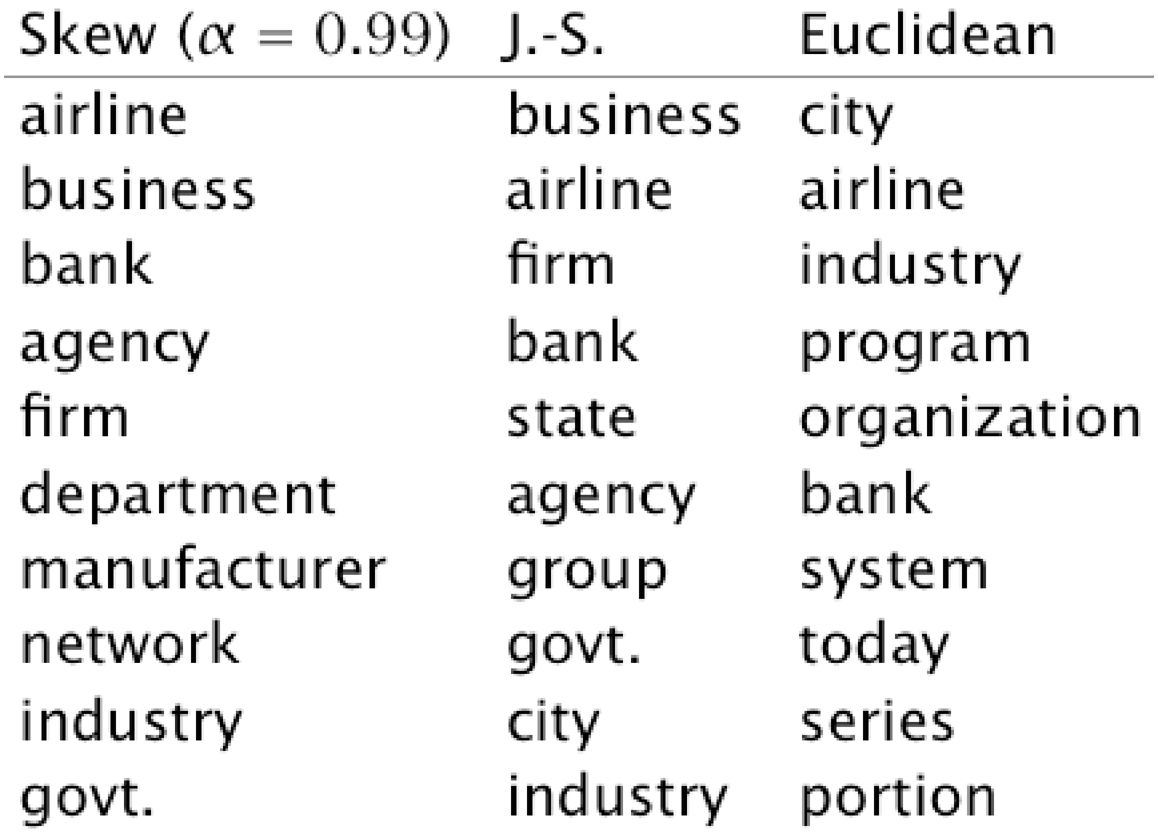
Focus word	Nearest neighbors							
garlic	*sauce*	.732	*pepper*	.728	*salt*	.726	*cup*	.726
fallen	*fell*	.932	*decline*	.931	*rise*	.930	*drop*	.929
engineered	*genetically*	.758	*drugs*	.688	*research*	.687	*drug*	.685
Alfred	*named*	.814	*Robert*	.809	*William*	.808	*W*	.808
simple	*something*	.964	*things*	.963	*You*	.963	*always*	.962

Probabilistic measures



(Dis-)similarity measure	Definition
KL divergence	$D(p \Vert q) = \sum_i p_i \log \frac{p_i}{q_i}$
Skew	$D(q \Vert \alpha r + (1-\alpha)q)$
Jensen-Shannon (was IRad)	$\frac{1}{2}D(p \Vert \frac{p+q}{2}) + D(q \Vert \frac{p+q}{2})$
L_1 norm (Manhattan)	$\sum_i \lvert p_i - q_i \rvert$

Neighbors of word “company” [Lee]

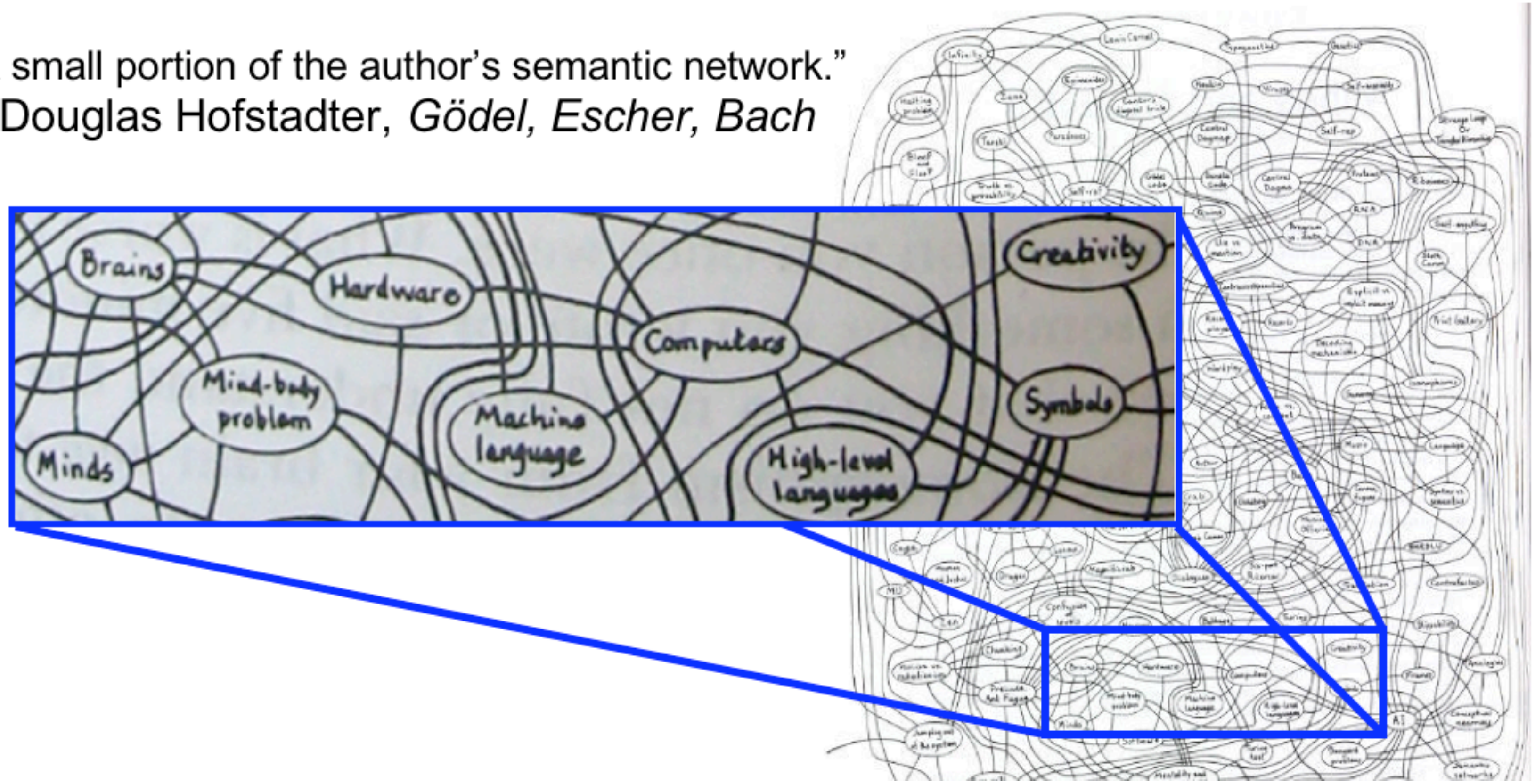


Skew ($\alpha = 0.99$)	J.-S.	Euclidean
airline	business	city
business	airline	airline
bank	firm	industry
agency	bank	program
firm	state	organization
department	agency	bank
manufacturer	group	system
network	govt.	today
industry	city	series
govt.	industry	portion

Learning syntactic patterns for automatic hypernym discovery

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng.

- It has long been a goal of AI to automatically acquire structured knowledge directly from text, e.g, in the form of a semantic network.



“A small portion of the author’s semantic network.”
– Douglas Hofstadter, *Gödel, Escher, Bach*

We aim to classify whether a noun pair (*X*, *Y*) participates in one of the following semantic relationships:

Hypernymy (ancestor)

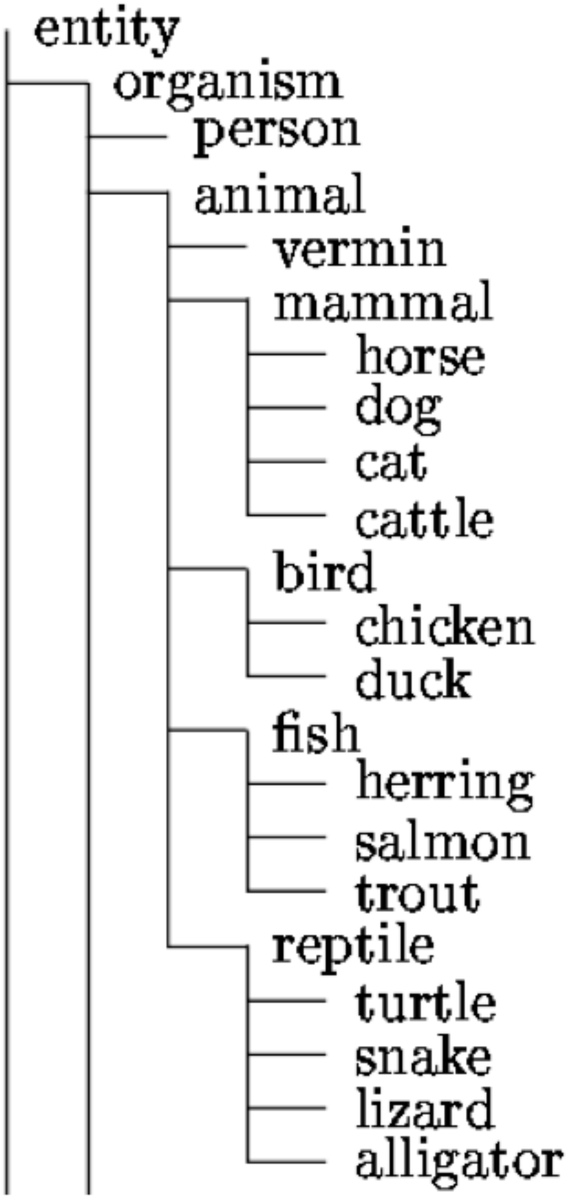
$Y \underset{H}{>} X$ if "*X* is a kind of *Y*".

$$entity \underset{H}{>} organism \underset{H}{>} person$$

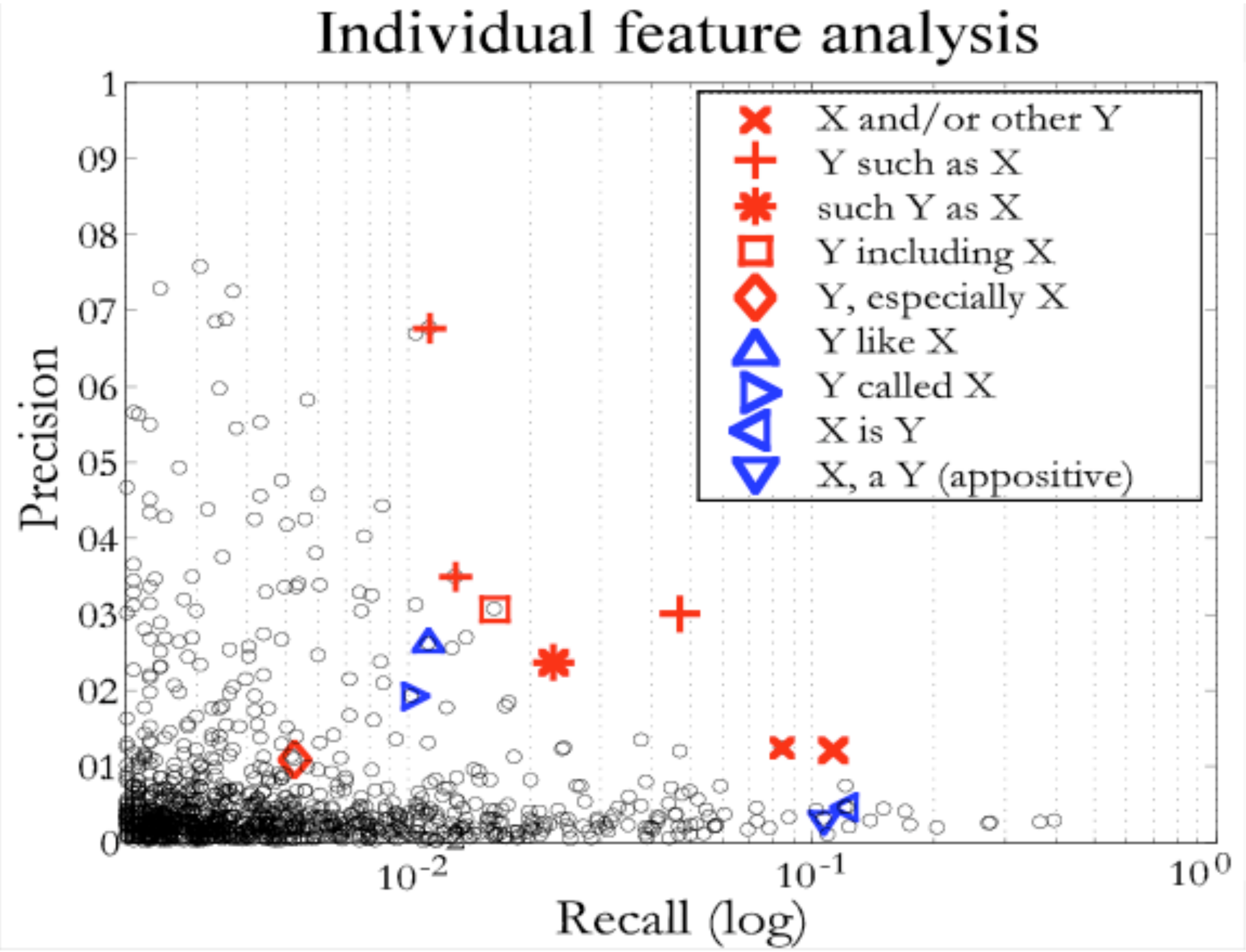
Coordinate Terms (taxonomic sisters)

$Y \underset{C}{\square} X$ if *X* and *Y* possess a common hypernym, i.e. $\exists Z$ such that "*X* and *Y* are both kinds of *Z*."

$$horse \underset{C}{\square} dog \underset{C}{\square} cat$$



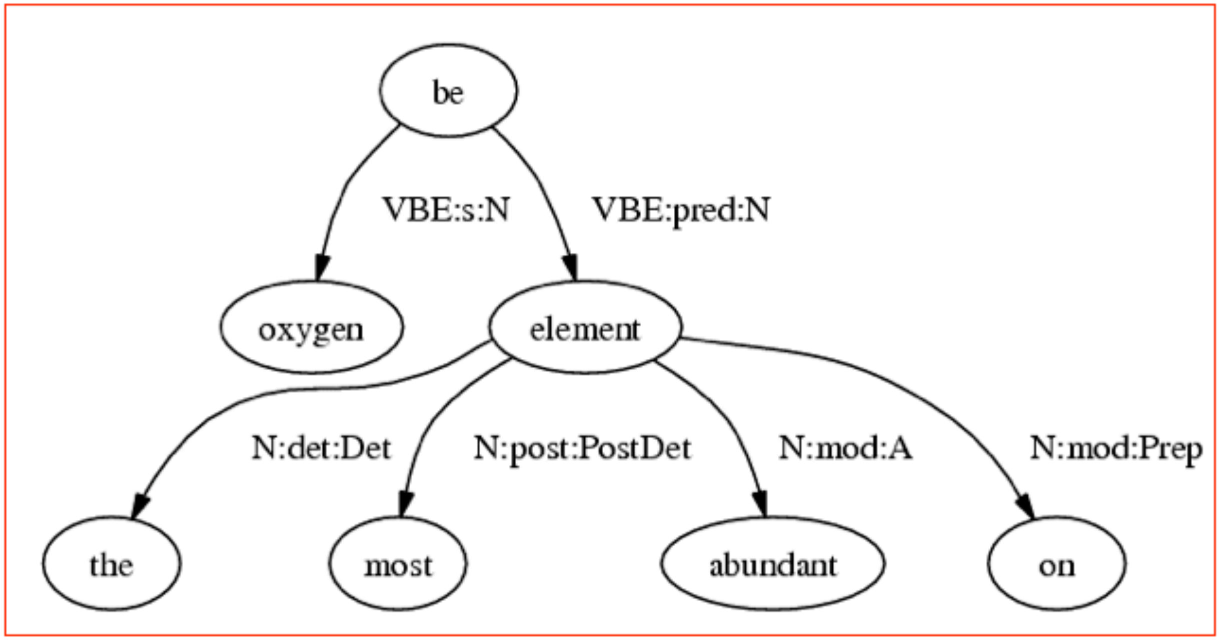
entity
organism
person
animal
vermin
mammal
horse
dog
cat
cattle
bird
chicken
duck
fish
herring
salmon
trout
reptile
turtle
snake
lizard
alligator



- Precision/recall for 69,592 classifiers (one per feature)
- Classifier *f* classifies noun pair ***x*** as hypernym iff $x_f > 0$
- **In red:** patterns originally proposed in (Hearst, 1992)

“Oxygen is the most abundant element on the moon.”

Dependency Graph:

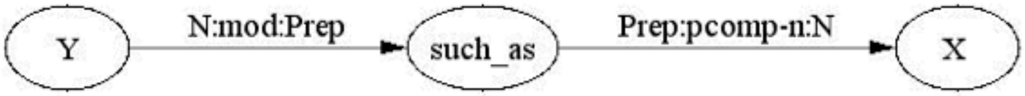


Dependency Paths (for “oxygen / element”):

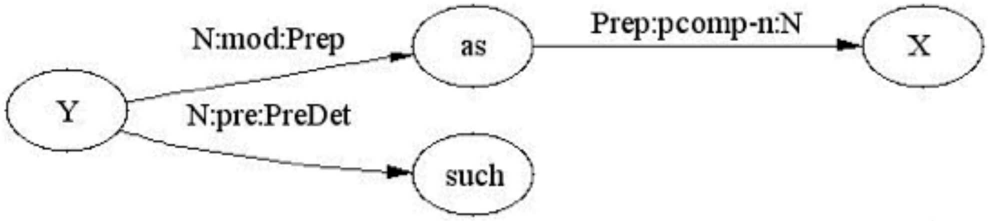
-N:s:VBE, “be” VBE:pred:N
-N:s:VBE, “be” VBE:pred:N,(the,Det:det:N)
-N:s:VBE, “be” VBE:pred:N,(most,PostDet:post:N)
-N:s:VBE, “be” VBE:pred:N,(abundant,A:mod:N)
-N:s:VBE, “be” VBE:pred:N,(on,Prep:mod:N)

Rediscovering Hearst's Patterns

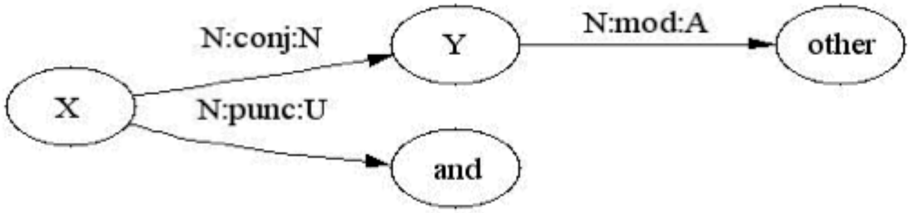
Y such as X…



Such Y as X…



X… and other Y



Proposed in (Hearst, 1992) and used in (Caraballo, 2001), (Widdows, 2003), and others – but what about the rest of the lexico-syntactic pattern space?

Example: Using the "Y called X" Pattern for Hypernym Acquisition

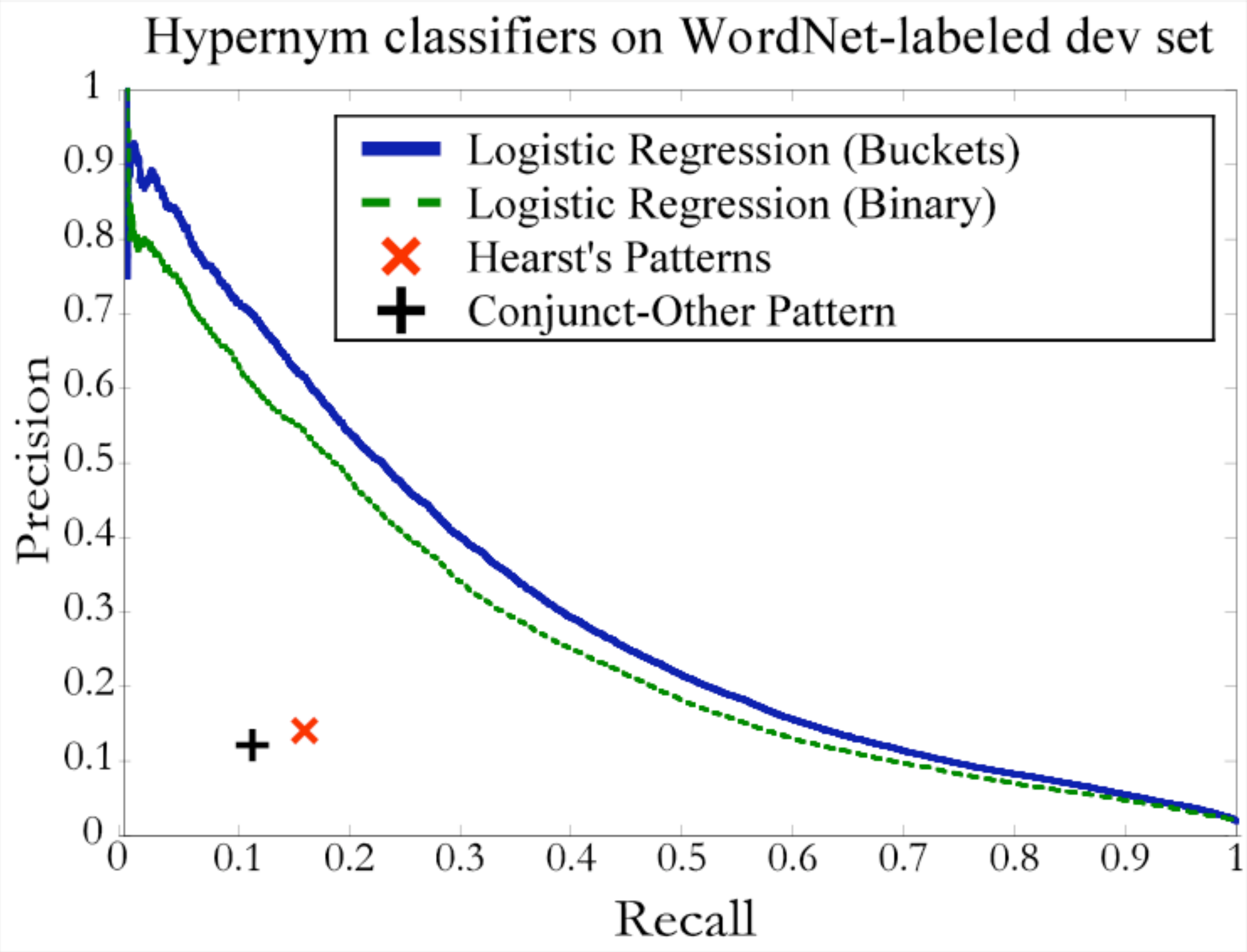
MINIPAR path: -N:desc:V,call,call,-V:vrel:N → "<hypernym> 'called' <hyponym>"

None of the following links are contained in WordNet (or the training set, by extension).

Hyponym	Hypernym	Sentence Fragment
efflorescence	**condition**	…and a **condition called efflorescence**…
'neal_inc	**company**	…The **company**, now **called O'Neal Inc.**…
hat_creek_outfit	**ranch**	…run a small **ranch called** the **Hat Creek Outfit**.
tardive_dyskinesia	**problem**	... irreversible **problem called tardive dyskinesia**…
hiv-1	**aids_virus**	…infected by the **AIDS virus**, **called HIV-1**.
bateau_mouche	**attraction**	…sightseeing **attraction called** the **Bateau Mouche**...
kibbutz_malkiyya	**collective_farm**	…Israeli **collective farm called Kibbutz Malkiyya**…

Type of Noun Pair	Count	Example Pair
NE: Person	7	"John F. Kennedy / president", "Marlin Fitzwater / spokesman"
NE: Place	7	"Diamond Bar / city", "France / place"
NE: Company	2	"American Can / company", "Simmons / company"
NE: Other	1	"Is Elvis Alive / book"
Not Named Entity:	9	"earthquake / disaster", "soybean / crop"

A better hypernym classifier



- 10-fold cross validation on the WordNet-labeled data
- **Conclusion:** 70,000 features are more powerful than 6

VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations

Timothy Chklovski and Patrick Pantel

Why Detect Semantic Rels between Verbs?

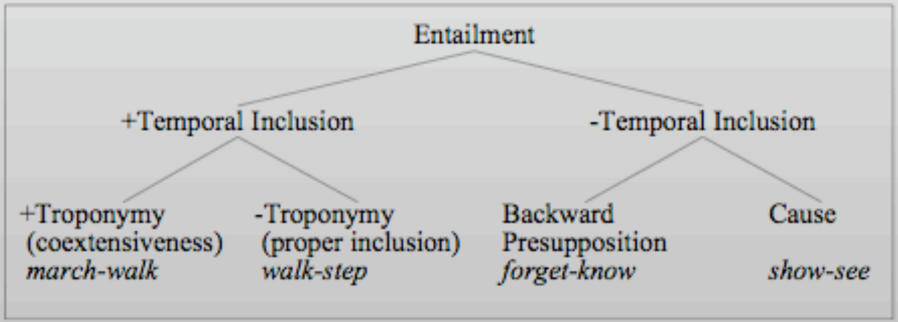
- So that we can
  - Understand the relationship when it's not stated
    - Napoleon ***fought*** and ***won*** the battle
    - During the holidays, people ***wrap*** and ***unwrap*** presents
    - Soldiers prefer to avoid getting ***wounded*** and ***killed***
  - Use the relationship when summarizing across documents (e.g. same event, preceding event)
    - The board ***considered*** the offer of $3B
    - The board ***accepted*** the offer $3.8B
    - The board ***okayed*** the offer of approximately $4B
  - Determine if two people have similar views on and event
    - "I ***nudged*** him."
    - "He ***shoved*** me."
- Hard to do manually

Why use Web? Motivating Intuition

- Small collections are tough: Semantics is often implied (Lenat, Chklovski)
- The Web's **10^{12}** is a lot of words
- So, Use small bits of more detailed text to help with mass of general text
  - Patterns issued to a search engine and their correlation

Relevant Work

- Levin’s classes (similarity)
  - 3200 verbs in 191 classes
- PropBank
  - 4,659 framesets (1.4 framesets per verb)
- VerbNet
  - 191 coarse-grained groupings (with overlap)
- FrameNet
- WordNet
  - troponomy
  - antonymy
  - entailment
  - cause



Entailment

+Temporal Inclusion

-Temporal Inclusion

+Troponymy (coextensiveness) *march-walk*

-Troponymy (proper inclusion) *walk-step*

Backward Presupposition *forget-know*

Cause *show-see*

Fellbaum’s (1998) entailment hierarchy.

VerbOcean: Web-based Extraction of Verb Relations

- VerbOcean is a network of verb relations
  - Currently, over 3400 nodes with on average 13 relations per verb
- Detected relation types are:
  - similarity
  - strength
  - antonymy
  - enablement
  - temporal precedence (happens-before)
- Download from http://semantics.isi.edu/ocean/

Approach

- Three stages:
  - Identify pairs of highly associated verbs co-occurring on the Web with sufficient frequency using DIRT (Lin and Pantel 2001)
  - For each verb pair
    - test patterns associated with each semantic relation
      - E.g. Temporal Precedence:

        “to *X* and then *Y*”, “*X*ed and then *Y*ed”
    - calculate a score for each possible semantic relation
  - Compare the strengths of the individual semantic relations and output a consistent set as the final output
    - prefer the most specific and then strongest relations

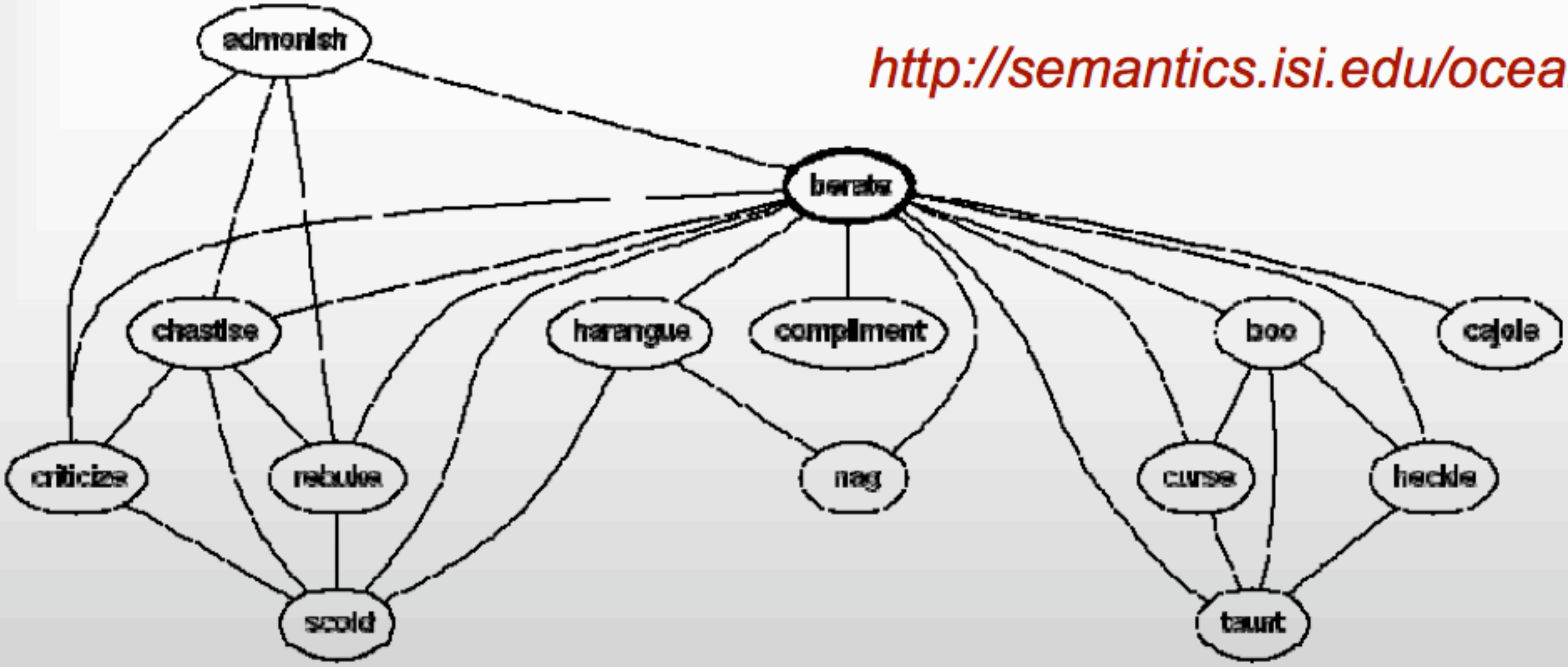
Lexical Patterns

Semantic Relation	*Surface Patterns*	*Example*
similarity (4)	X ie Y Xed and Yed	*"She **heckled and taunted** the comedian."*
strength (8)	X even Y Xed even Yed Xed and even Yed not just Xed but Yed	*"He not **just harassed, but terrorized** her."*
enablement (4)	Xed * by Ying the Xed * by Ying or to X * by Ying the	*"She **saved the document by clicking the** button."*
antonymy (7)	either X or Y either Xs or Ys Xed * but Yed	*"There's something about Mary: you will **either love or hate** her."*
happens-before (12)	to X and then Y Xed * and then Yed to X and later Y to X and subsequently Y Xed and subsequently Yed	*"He **designed the prototype and then patented** it."*

Lexical Patterns Match...

- Refined to decrease capturing wrong parts of speech or incorrect semantic relations
  - Xed * by Ying the; Xed * by Ying or
    - "… waved at by parking guard …"
    - "… encouraged further by sailing lessons …"

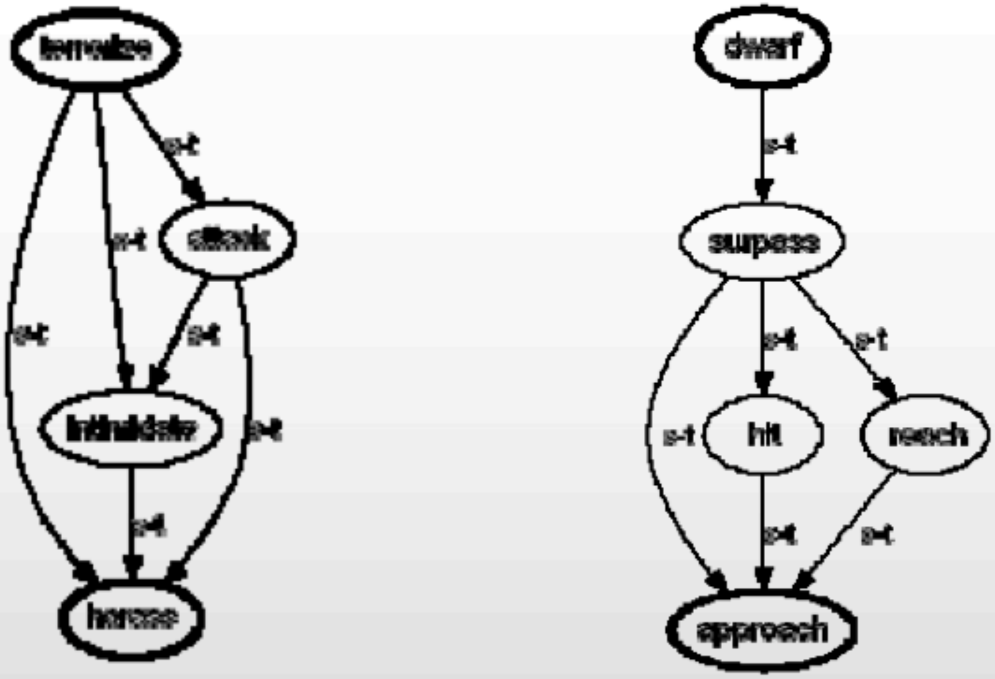
VerbOcean – Similarity



http://semantics.isi.edu/ocean/

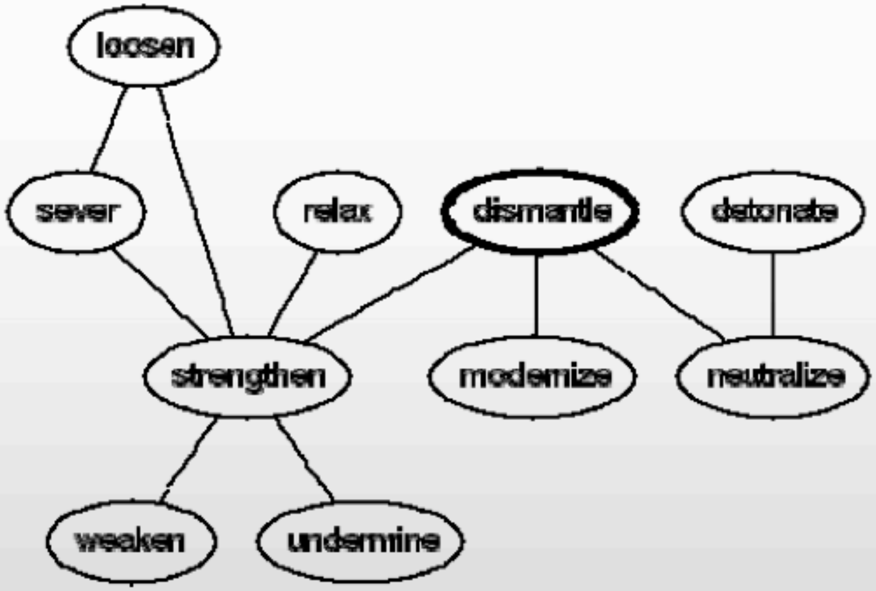
- Verbs that are similar or related
  - e.g. boo - heckle

VerbOcean – Strength



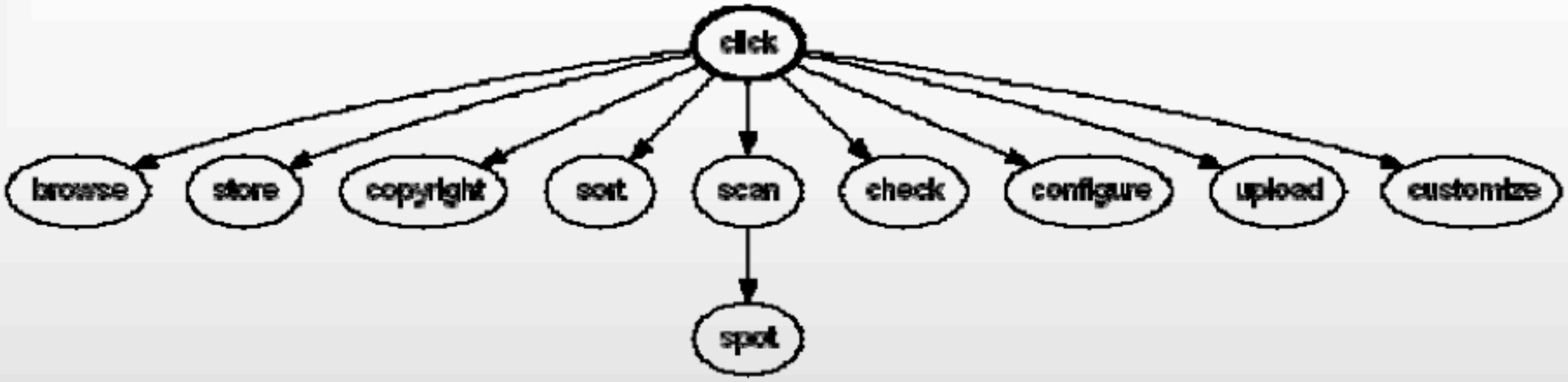
- Similar verbs that denote a more intense, thorough, comprehensive or absolute action
  - e.g. change-of-state verbs that denote a more complete change (shock → startle)

VerbOcean – Antonymy



- Semantic opposition
  - switching thematic roles associated with the verb (buy – sell)
  - stative verbs (live – die)
  - sibling verbs which share a parent (walk – run)
  - restitutive opposition: antonymy + happens-before (damage - repair)

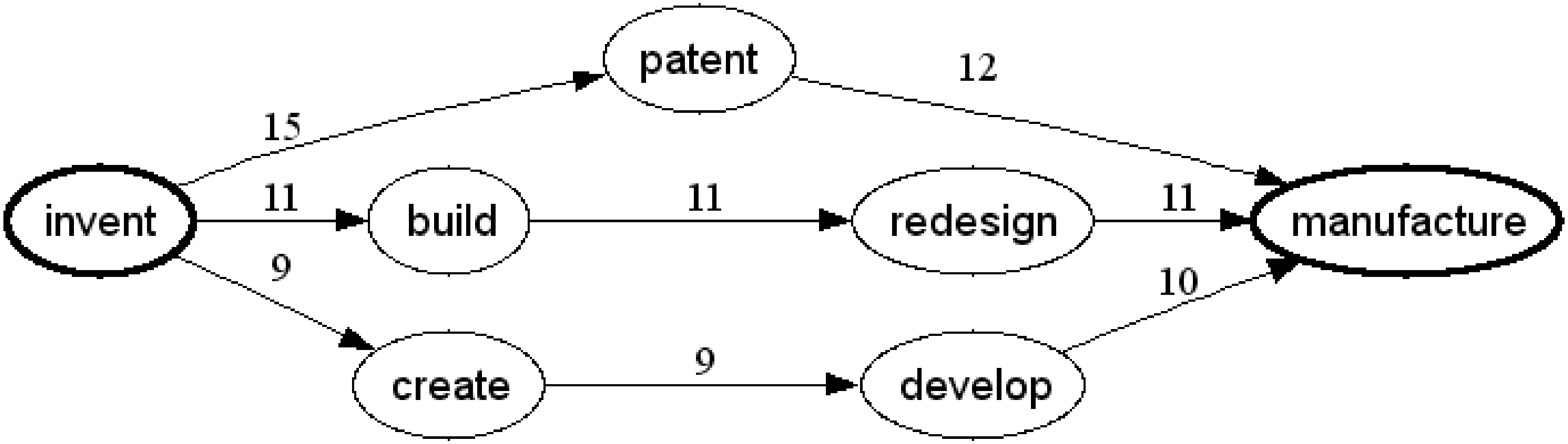
VerbOcean – Enablement



- Holds between two verbs V_1 and V_2 when the pair can be glossed as "V_1 is accomplished by V_2" (assess - review)

Appendix. Sample relations extracted by our system.

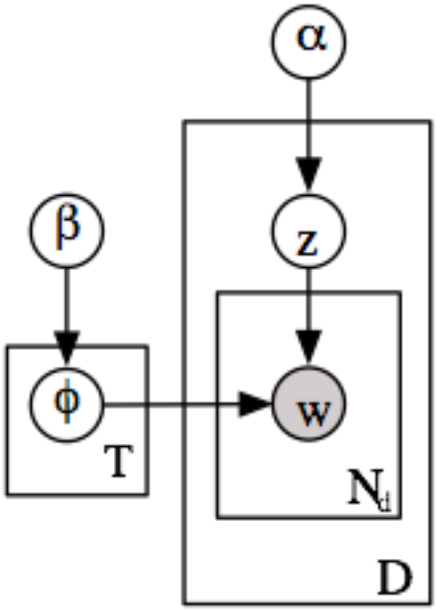
SEMANTIC RELATION	EXAMPLES	SEMANTIC RELATION	EXAMPLES	SEMANTIC RELATION	EXAMPLES
similarity	maximize :: enhance produce :: create reduce :: restrict	enablement	assess :: review accomplish :: complete double-click :: click	happens before	detain :: prosecute enroll :: graduate schedule :: reschedule
strength	permit :: authorize surprise :: startle startle :: shock	antonymy	assemble :: dismantle regard :: condemn roast :: fry		



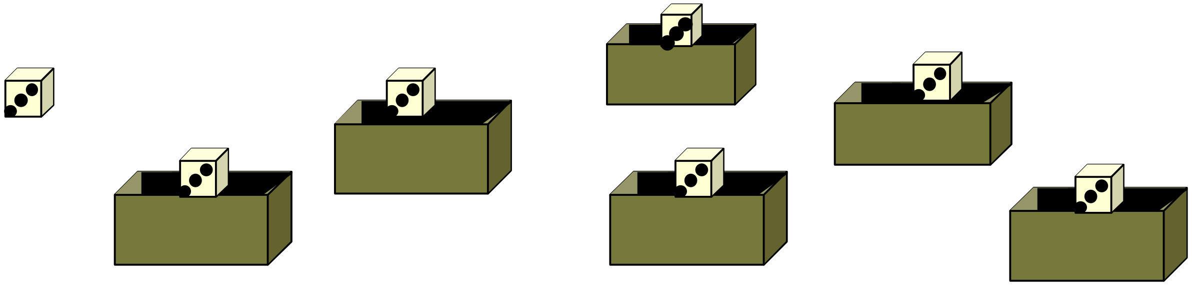
Topic Models

Unsupervised Models of
Word Co-occurrences

A Probabilistic Approach

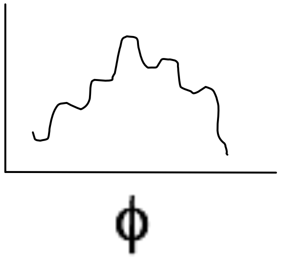


- Define a probabilistic generative model for documents.



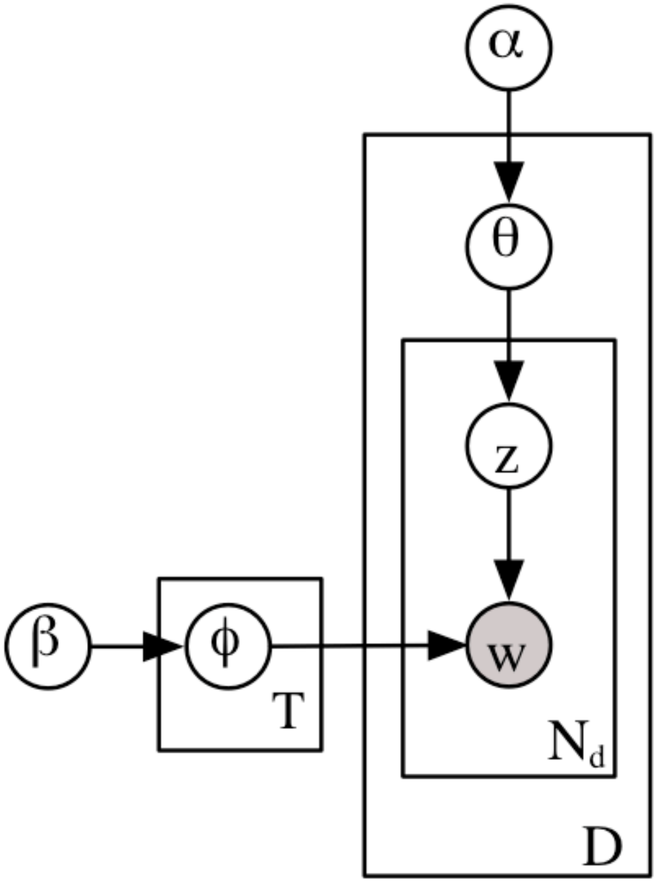
- Learn the parameters of this model by fitting them to the data and a prior.

$$\phi^* = \arg\max_{\phi} p(\phi|D_1 D_2 \ldots) = p(D_1 D_2 \ldots|\phi)\, p(\phi)$$



Clustering words into topics with Latent Dirichlet Allocation

[Blei, Ng, Jordan 2003]



Generative Process:	**Example:**
For each document:	
Sample a distribution over topics, θ	70% Iraq war 30% US election
For each word in doc	
Sample a topic, *z*	Iraq war
Sample a word from the topic, *w*	“bombing”

Example topics
induced from a large collection of text

DISEASE
BACTERIA
DISEASES
GERMS
FEVER
CAUSE
CAUSED
SPREAD
VIRUSES
INFECTION
VIRUS
MICROORGANISMS
PERSON
INFECTIOUS
COMMON
CAUSING
SMALLPOX
BODY
INFECTIONS
CERTAIN

WATER
FISH
SEA
SWIM
SWIMMING
POOL
LIKE
SHELL
SHARK
TANK
SHELLS
SHARKS
DIVING
DOLPHINS
SWAM
LONG
SEAL
DIVE
DOLPHIN
UNDERWATER

MIND
WORLD
DREAM
DREAMS
THOUGHT
IMAGINATION
MOMENT
THOUGHTS
OWN
REAL
LIFE
IMAGINE
SENSE
CONSCIOUSNESS
STRANGE
FEELING
WHOLE
BEING
MIGHT
HOPE

STORY
STORIES
TELL
CHARACTER
CHARACTERS
AUTHOR
READ
TOLD
SETTING
TALES
PLOT
TELLING
SHORT
FICTION
ACTION
TRUE
EVENTS
TELLS
TALE
NOVEL

FIELD
MAGNETIC
MAGNET
WIRE
NEEDLE
CURRENT
COIL
POLES
IRON
COMPASS
LINES
CORE
ELECTRIC
DIRECTION
FORCE
MAGNETS
BE
MAGNETISM
POLE
INDUCED

SCIENCE
STUDY
SCIENTISTS
SCIENTIFIC
KNOWLEDGE
WORK
RESEARCH
CHEMISTRY
TECHNOLOGY
MANY
MATHEMATICS
BIOLOGY
FIELD
PHYSICS
LABORATORY
STUDIES
WORLD
SCIENTIST
STUDYING
SCIENCES

BALL
GAME
TEAM
FOOTBALL
BASEBALL
PLAYERS
PLAY
FIELD
PLAYER
BASKETBALL
COACH
PLAYED
PLAYING
HIT
TENNIS
TEAMS
GAMES
SPORTS
BAT
TERRY

JOB
WORK
JOBS
CAREER
EXPERIENCE
EMPLOYMENT
OPPORTUNITIES
WORKING
TRAINING
SKILLS
CAREERS
POSITIONS
FIND
POSITION
FIELD
OCCUPATIONS
REQUIRE
OPPORTUNITY
EARN
ABLE

[Tennenbaum et al]

Example topics
induced from a large collection of text

DISEASE
BACTERIA
DISEASES
GERMS
FEVER
CAUSE
CAUSED
SPREAD
VIRUSES
INFECTION
VIRUS
MICROORGANISMS
PERSON
INFECTIOUS
COMMON
CAUSING
SMALLPOX
BODY
INFECTIONS
CERTAIN

WATER
FISH
SEA
SWIM
SWIMMING
POOL
LIKE
SHELL
SHARK
TANK
SHELLS
SHARKS
DIVING
DOLPHINS
SWAM
LONG
SEAL
DIVE
DOLPHIN
UNDERWATER

MIND
WORLD
DREAM
DREAMS
THOUGHT
IMAGINATION
MOMENT
THOUGHTS
OWN
REAL
LIFE
IMAGINE
SENSE
CONSCIOUSNESS
STRANGE
FEELING
WHOLE
BEING
MIGHT
HOPE

STORY
STORIES
TELL
CHARACTER
CHARACTERS
AUTHOR
READ
TOLD
SETTING
TALES
PLOT
TELLING
SHORT
FICTION
ACTION
TRUE
EVENTS
TELLS
TALE
NOVEL

FIELD
MAGNETIC
MAGNET
WIRE
NEEDLE
CURRENT
COIL
POLES
IRON
COMPASS
LINES
CORE
ELECTRIC
DIRECTION
FORCE
MAGNETS
BE
MAGNETISM
POLE
INDUCED

SCIENCE
STUDY
SCIENTISTS
SCIENTIFIC
KNOWLEDGE
WORK
RESEARCH
CHEMISTRY
TECHNOLOGY
MANY
MATHEMATICS
BIOLOGY
FIELD
PHYSICS
LABORATORY
STUDIES
WORLD
SCIENTIST
STUDYING
SCIENCES

BALL
GAME
TEAM
FOOTBALL
BASEBALL
PLAYERS
PLAY
FIELD
PLAYER
BASKETBALL
COACH
PLAYED
PLAYING
HIT
TENNIS
TEAMS
GAMES
SPORTS
BAT
TERRY

JOB
WORK
JOBS
CAREER
EXPERIENCE
EMPLOYMENT
OPPORTUNITIES
WORKING
TRAINING
SKILLS
CAREERS
POSITIONS
FIND
POSITION
FIELD
OCCUPATIONS
REQUIRE
OPPORTUNITY
EARN
ABLE

[Tennenbaum et al]

Collocations

- An expression consisting of two or more words that correspond to some conventional way of saying things.
- Characterized by limited *compositionality*.
  - *compositional*: meaning of expression can be predicted by meaning of its parts.
  - “dynamic programming”, “hidden Markov model”
  - “weapons of mass destruction”
  - “kick the bucket”, “hear it through the grapevine”

Topics Modeling Phrases

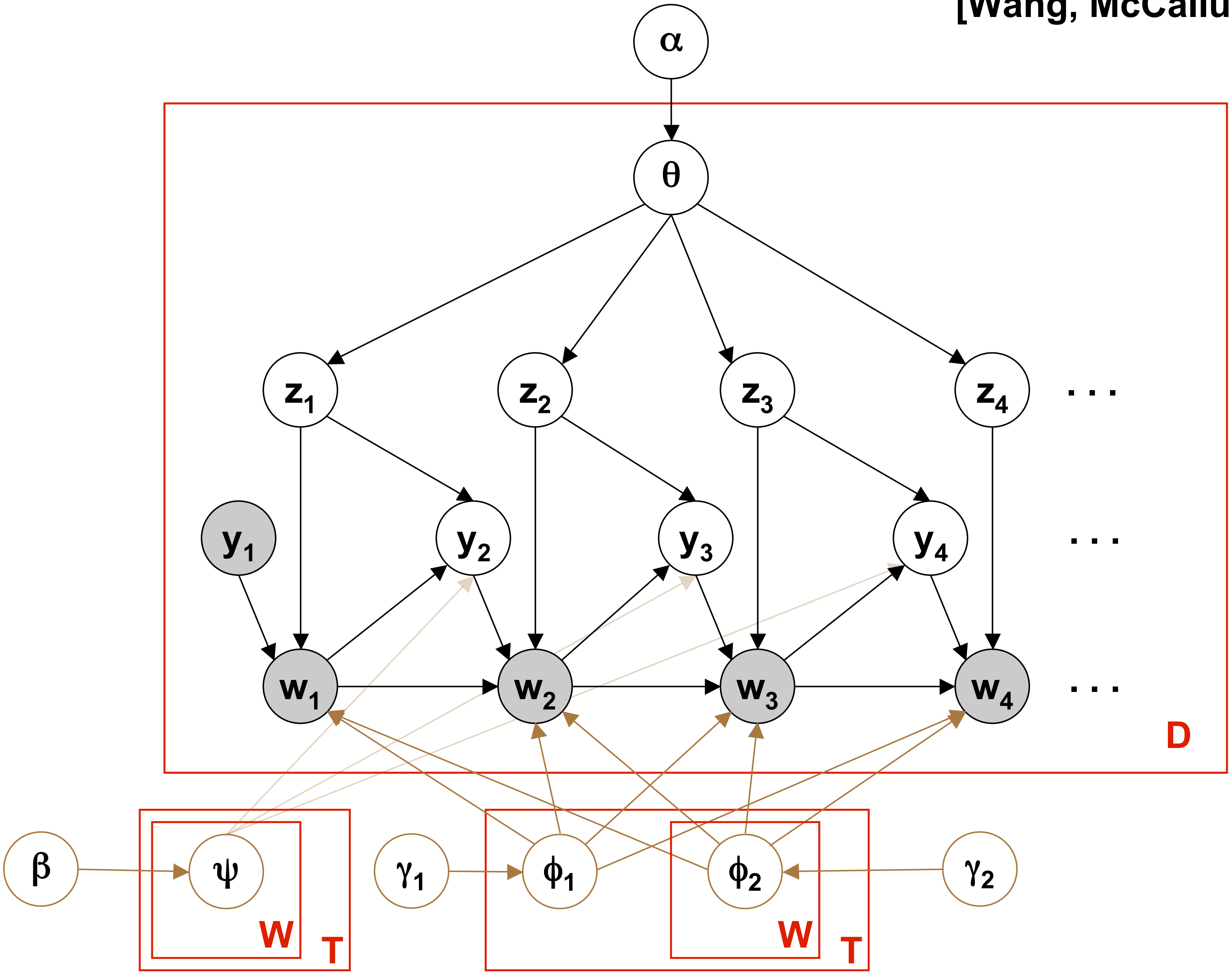
- Topics based only on unigrams often difficult to interpret

- Topic discovery itself is confused because important meaning / distinctions carried by phrases.

- Significant opportunity to provide improved language models to ASR, MT, IR, etc.

Topical N-gram Model

[Wang, McCallum 2005]



LDA Topic

LDA	Topical N-grams
algorithms	genetic algorithms
algorithm	genetic algorithm
genetic	evolutionary computation
problems	evolutionary algorithms
efficient	fitness function

Topic Comparison

LDA	Topical N-grams (2)	Topical N-grams (1)
learning	reinforcement learning	policy
optimal	optimal policy	action
reinforcement	dynamic programming	states
state	optimal control	actions
problems	function approximator	function
policy	prioritized sweeping	reward
dynamic	finite-state controller	control
action	learning system	agent
programming	reinforcement learning rl	q-learning
actions	function approximators	optimal
function	markov decision problems	goal
markov	markov decision processes	learning
methods	local search	space
decision	state-action pair	step
rl	markov decision process	environment
continuous	belief states	system
spaces	stochastic policy	problem
step	action selection	steps
policies	upright position	sutton
planning	reinforcement learning methods	policies

Topic Comparison

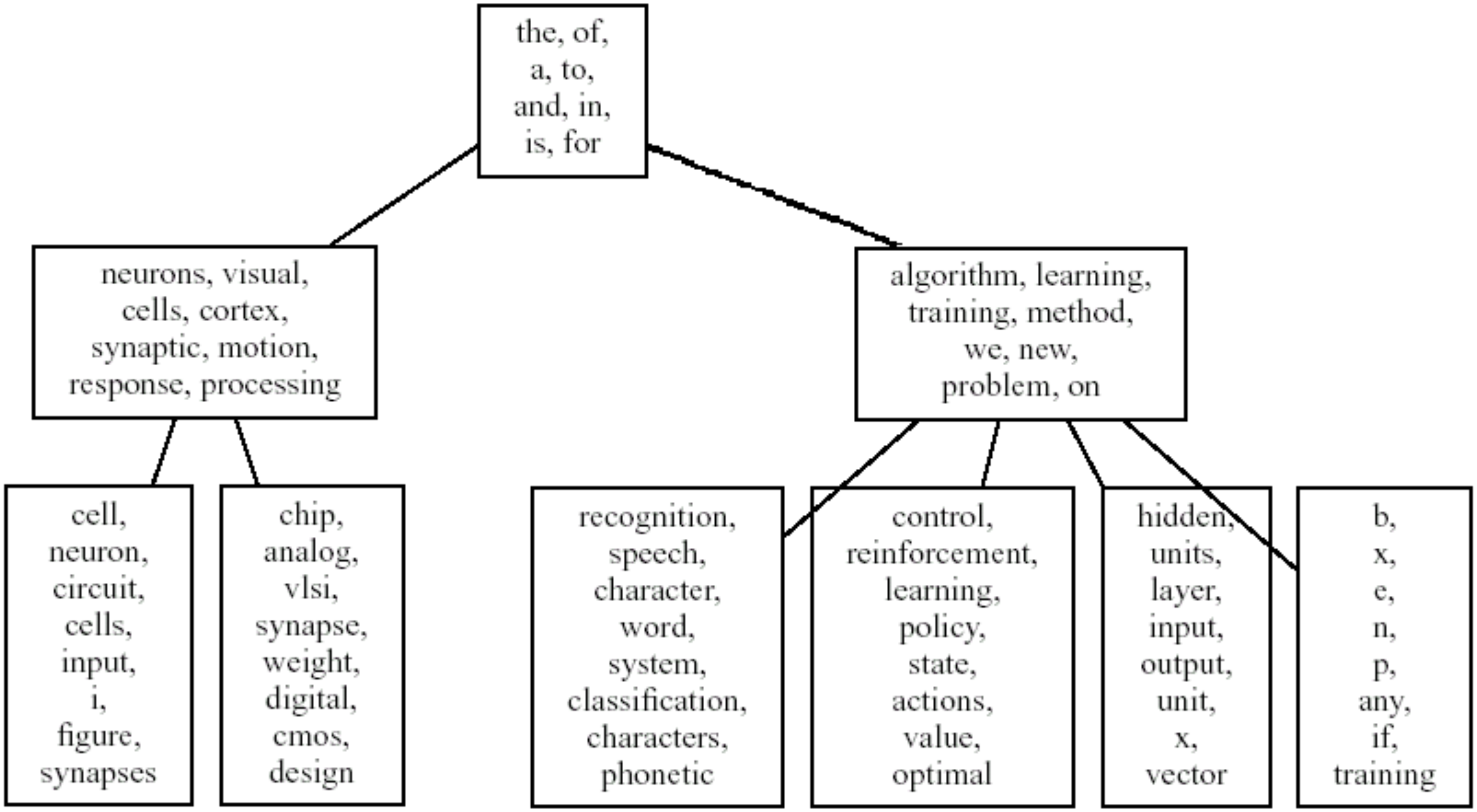
LDA	Topical N-grams (2)	Topical N-grams (1)
motion	receptive field	motion
visual	spatial frequency	response
field	temporal frequency	direction
position	visual motion	cells
figure	motion energy	stimulus
direction	tuning curves	figure
fields	horizontal cells	contrast
eye	motion detection	velocity
location	preferred direction	model
retina	visual processing	responses
receptive	area mt	stimuli
velocity	visual cortex	moving
vision	light intensity	cell
moving	directional selectivity	intensity
system	high contrast	population
flow	motion detectors	image
edge	spatial phase	center
center	moving stimuli	tuning
light	decision strategy	complex
local	visual stimuli	directions

Topic Comparison

LDA	Topical N-grams (2)	Topical N-grams (1)
word	speech recognition	speech
system	training data	word
recognition	neural network	training
hmm	error rates	system
speech	neural net	recognition
training	hidden markov model	hmm
performance	feature vectors	speaker
phoneme	continuous speech	performance
words	training procedure	phoneme
context	continuous speech recognition	acoustic
systems	gamma filter	words
frame	hidden control	context
trained	speech production	systems
speaker	neural nets	frame
sequence	input representation	trained
speakers	output layers	sequence
mlp	training algorithm	phonetic
frames	test set	speakers
segmentation	speech frames	mlp
models	speaker dependent	hybrid

Unsupervised learning of topic hierarchies

(Blei, Griffiths, Jordan & Tenenbaum, NIPS 2003)



the, of, a, to, and, in, is, for

neurons, visual, cells, cortex, synaptic, motion, response, processing

algorithm, learning, training, method, we, new, problem, on

cell, neuron, circuit, cells, input, i, figure, synapses

chip, analog, vlsi, synapse, weight, digital, cmos, design

recognition, speech, character, word, system, classification, characters, phonetic

control, reinforcement, learning, policy, state, actions, value, optimal

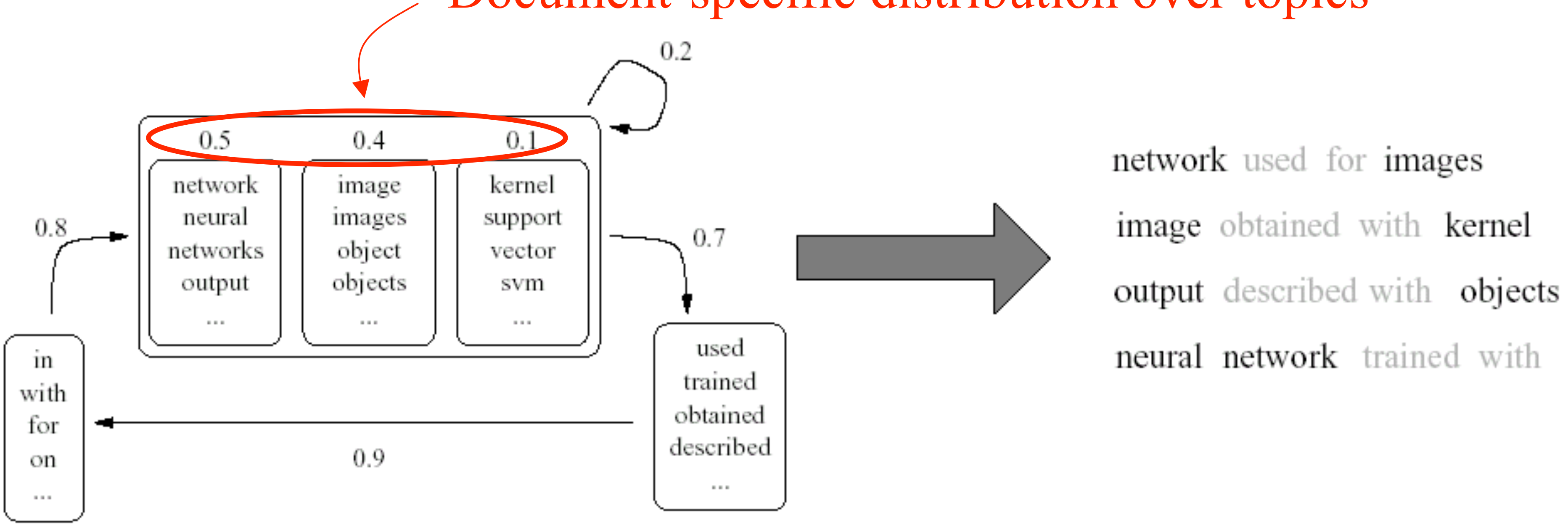
hidden, units, layer, input, output, unit, x, vector

b, x, e, n, p, any, if, training

Joint models of syntax and semantics (Griffiths, Steyvers, Blei & Tenenbaum, NIPS 2004)

- Embed topics model inside an *n*th order Hidden Markov Model:

Document-specific distribution over topics



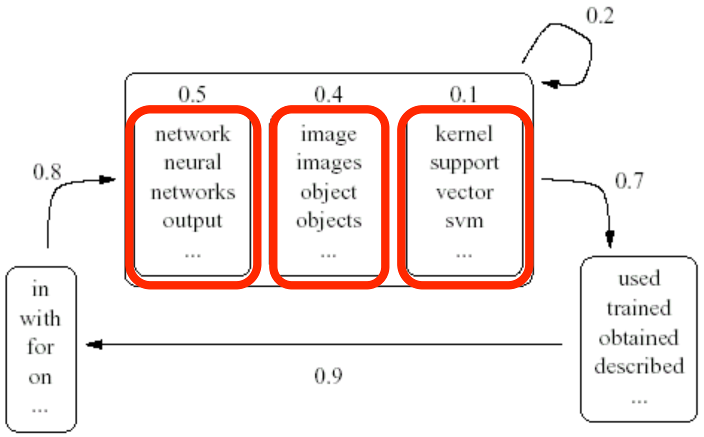
network used for images

image obtained with kernel

output described with objects

neural network trained with

Semantic classes



FOOD
FOODS
BODY
NUTRIENTS
DIET
FAT
SUGAR
ENERGY
MILK
EATING
FRUITS
VEGETABLES
WEIGHT
FATS
NEEDS
CARBOHYDRATES
VITAMINS
CALORIES
PROTEIN
MINERALS

MAP
NORTH
EARTH
SOUTH
POLE
MAPS
EQUATOR
WEST
LINES
EAST
AUSTRALIA
GLOBE
POLES
HEMISPHERE
LATITUDE
PLACES
LAND
WORLD
COMPASS
CONTINENTS

DOCTOR
PATIENT
HEALTH
HOSPITAL
MEDICAL
CARE
PATIENTS
NURSE
DOCTORS
MEDICINE
NURSING
TREATMENT
NURSES
PHYSICIAN
HOSPITALS
DR
SICK
ASSISTANT
EMERGENCY
PRACTICE

BOOK
BOOKS
READING
INFORMATION
LIBRARY
REPORT
PAGE
TITLE
SUBJECT
PAGES
GUIDE
WORDS
MATERIAL
ARTICLE
ARTICLES
WORD
FACTS
AUTHOR
REFERENCE
NOTE

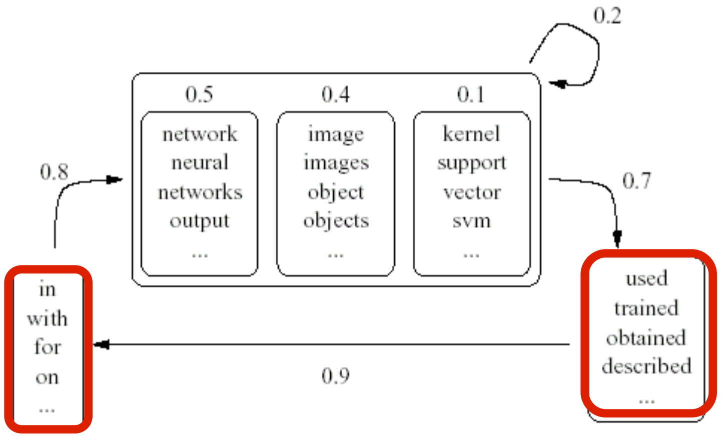
GOLD
IRON
SILVER
COPPER
METAL
METALS
STEEL
CLAY
LEAD
ADAM
ORE
ALUMINUM
MINERAL
MINE
STONE
MINERALS
POT
MINING
MINERS
TIN

BEHAVIOR
SELF
INDIVIDUAL
PERSONALITY
RESPONSE
SOCIAL
EMOTIONAL
LEARNING
FEELINGS
PSYCHOLOGISTS
INDIVIDUALS
PSYCHOLOGICAL
EXPERIENCES
ENVIRONMENT
HUMAN
RESPONSES
BEHAVIORS
ATTITUDES
PSYCHOLOGY
PERSON

CELLS
CELL
ORGANISMS
ALGAE
BACTERIA
MICROSCOPE
MEMBRANE
ORGANISM
FOOD
LIVING
FUNGI
MOLD
MATERIALS
NUCLEUS
CELLED
STRUCTURES
MATERIAL
STRUCTURE
GREEN
MOLDS

PLANTS
PLANT
LEAVES
SEEDS
SOIL
ROOTS
FLOWERS
WATER
FOOD
GREEN
SEED
STEMS
FLOWER
STEM
LEAF
ANIMALS
ROOT
POLLEN
GROWING
GROW

Syntactic classes



SAID
ASKED
THOUGHT
TOLD
SAYS
MEANS
CALLED
CRIED
SHOWS
ANSWERED
TELLS
REPLIED
SHOUTED
EXPLAINED
LAUGHED
MEANT
WROTE
SHOWED
BELIEVED
WHISPERED

THE
HIS
THEIR
YOUR
HER
ITS
MY
OUR
THIS
THESE
A
AN
THAT
NEW
THOSE
EACH
MR
ANY
MRS
ALL

MORE
SUCH
LESS
MUCH
KNOWN
JUST
BETTER
RATHER
GREATER
HIGHER
LARGER
LONGER
FASTER
EXACTLY
SMALLER
SOMETHING
BIGGER
FEWER
LOWER
ALMOST

ON
AT
INTO
FROM
WITH
THROUGH
OVER
AROUND
AGAINST
ACROSS
UPON
TOWARD
UNDER
ALONG
NEAR
BEHIND
OFF
ABOVE
DOWN
BEFORE

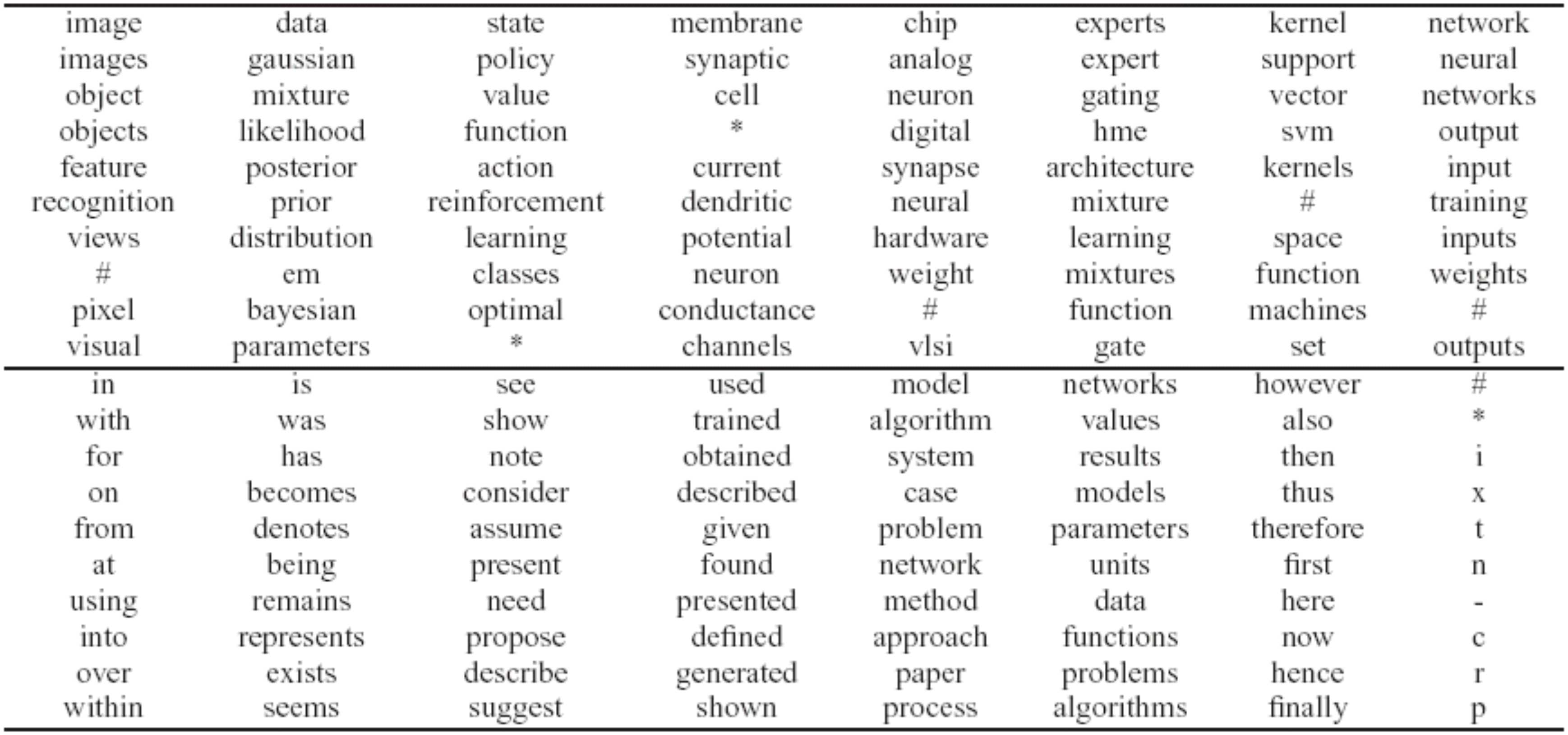
GOOD
SMALL
NEW
IMPORTANT
GREAT
LITTLE
LARGE
*
BIG
LONG
HIGH
DIFFERENT
SPECIAL
OLD
STRONG
YOUNG
COMMON
WHITE
SINGLE
CERTAIN

ONE
SOME
MANY
TWO
EACH
ALL
MOST
ANY
THREE
THIS
EVERY
SEVERAL
FOUR
FIVE
BOTH
TEN
SIX
MUCH
TWENTY
EIGHT

HE
YOU
THEY
I
SHE
WE
IT
PEOPLE
EVERYONE
OTHERS
SCIENTISTS
SOMEONE
WHO
NOBODY
ONE
SOMETHING
ANYONE
EVERYBODY
SOME
THEN

BE
MAKE
GET
HAVE
GO
TAKE
DO
FIND
USE
SEE
HELP
KEEP
GIVE
LOOK
COME
WORK
MOVE
LIVE
EAT
BECOME

Corpus-specific factorization (NIPS)



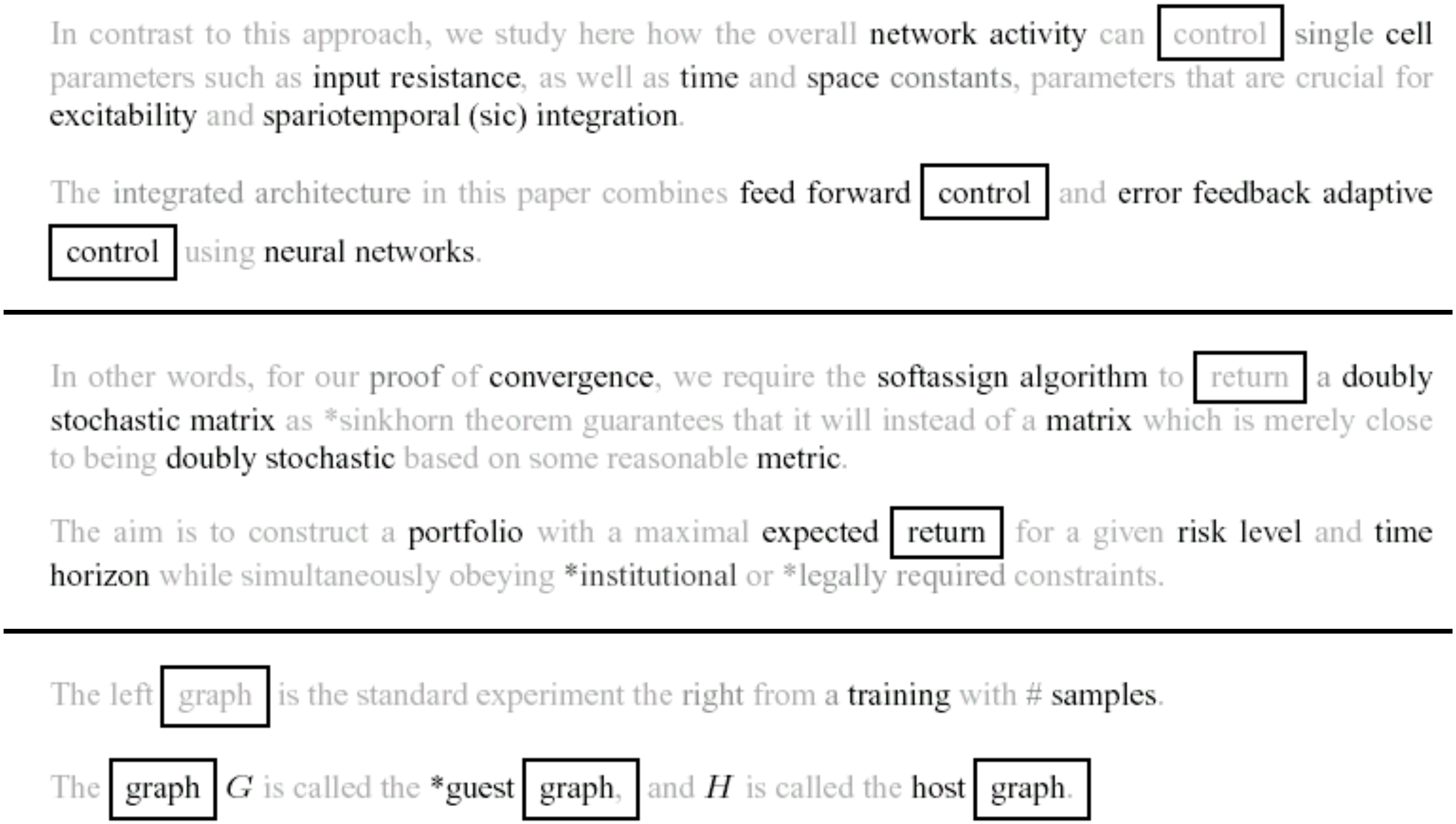
Semantics	image	data	state	membrane	chip	experts	kernel	network
	images	gaussian	policy	synaptic	analog	expert	support	neural
	object	mixture	value	cell	neuron	gating	vector	networks
	objects	likelihood	function	*	digital	hme	svm	output
	feature	posterior	action	current	synapse	architecture	kernels	input
	recognition	prior	reinforcement	dendritic	neural	mixture	#	training
	views	distribution	learning	potential	hardware	learning	space	inputs
	#	em	classes	neuron	weight	mixtures	function	weights
	pixel	bayesian	optimal	conductance	#	function	machines	#
	visual	parameters	*	channels	vlsi	gate	set	outputs
Syntax	in	is	see	used	model	networks	however	#
	with	was	show	trained	algorithm	values	also	*
	for	has	note	obtained	system	results	then	i
	on	becomes	consider	described	case	models	thus	x
	from	denotes	assume	given	problem	parameters	therefore	t
	at	being	present	found	network	units	first	n
	using	remains	need	presented	method	data	here	-
	into	represents	propose	defined	approach	functions	now	c
	over	exists	describe	generated	paper	problems	hence	r
	within	seems	suggest	shown	process	algorithms	finally	p

Syntactic classes in PNAS

5	8	14	25	26	30	33
IN	ARE	THE	SUGGEST	LEVELS	RESULTS	BEEN
FOR	WERE	THIS	INDICATE	NUMBER	ANALYSIS	MAY
ON	WAS	ITS	SUGGESTING	LEVEL	DATA	CAN
BETWEEN	IS	THEIR	SUGGESTS	RATE	STUDIES	COULD
DURING	WHEN	AN	SHOWED	TIME	STUDY	WELL
AMONG	REMAIN	EACH	REVEALED	CONCENTRATIONS	FINDINGS	DID
FROM	REMAINS	ONE	SHOW	VARIETY	EXPERIMENTS	DOES
UNDER	REMAINED	ANY	DEMONSTRATE	RANGE	OBSERVATIONS	DO
WITHIN	PREVIOUSLY	INCREASED	INDICATING	CONCENTRATION	HYPOTHESIS	MIGHT
THROUGHOUT	BECOME	EXOGENOUS	PROVIDE	DOSE	ANALYSES	SHOULD
THROUGH	BECAME	OUR	SUPPORT	FAMILY	ASSAYS	WILL
TOWARD	BEING	RECOMBINANT	INDICATES	SET	POSSIBILITY	WOULD
INTO	BUT	ENDOGENOUS	PROVIDES	FREQUENCY	MICROSCOPY	MUST
AT	GIVE	TOTAL	INDICATED	SERIES	PAPER	CANNOT
INVOLVING	MERE	PURIFIED	DEMONSTRATED	AMOUNTS	WORK	REMAINED
AFTER	APPEARED	TILE	SHOWS	RATES	EVIDENCE	ALSO
ACROSS	APPEAR	FULL	SO	CLASS	FINDING	THEY
AGAINST	ALLOWED	CHRONIC	REVEAL	VALUES	MUTAGENESIS	BECOME
WHEN	NORMALLY	ANOTHER	DEMONSTRATES	AMOUNT	OBSERVATION	MAG
ALONG	EACH	EXCESS	SUGGESTED	SITES	MEASUREMENTS	LIKELY

Semantic highlighting

Darker words are more likely to have been generated from the topic-based “semantics” module:



In contrast to this approach, we study here how the overall network activity can control single cell parameters such as input resistance, as well as time and space constants, parameters that are crucial for excitability and spariotemporal (sic) integration.

The integrated architecture in this paper combines feed forward control and error feedback adaptive control using neural networks.

In other words, for our proof of convergence, we require the softassign algorithm to return a doubly stochastic matrix as *sinkhorn theorem guarantees that it will instead of a matrix which is merely close to being doubly stochastic based on some reasonable metric.

The aim is to construct a portfolio with a maximal expected return for a given risk level and time horizon while simultaneously obeying *institutional or *legally required constraints.

The left graph is the standard experiment the right from a training with # samples.

The graph G is called the *guest graph, and H is called the host graph.

Social Network Analysis with Links *and Text*

Role Discovery

Group Discovery

Trend Discovery

Community Discovery

Impact Measurement

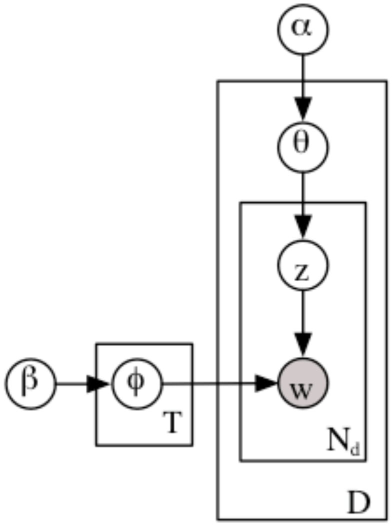
From LDA to Author-Recipient-Topic

(ART)

Latent Dirichlet Allocation

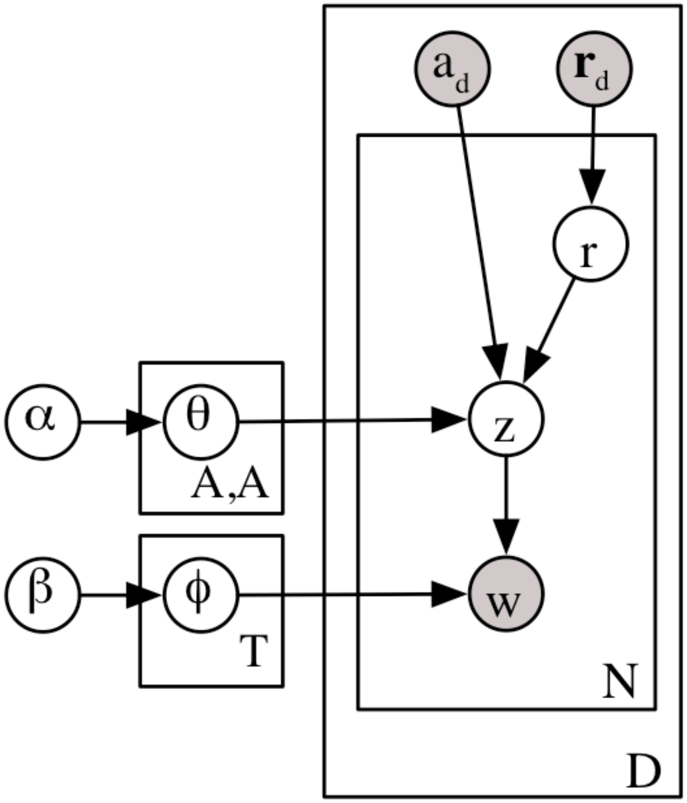
(LDA)

[Blei, Ng, Jordan, 2003]



Inference and Estimation

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



Gibbs Sampling:

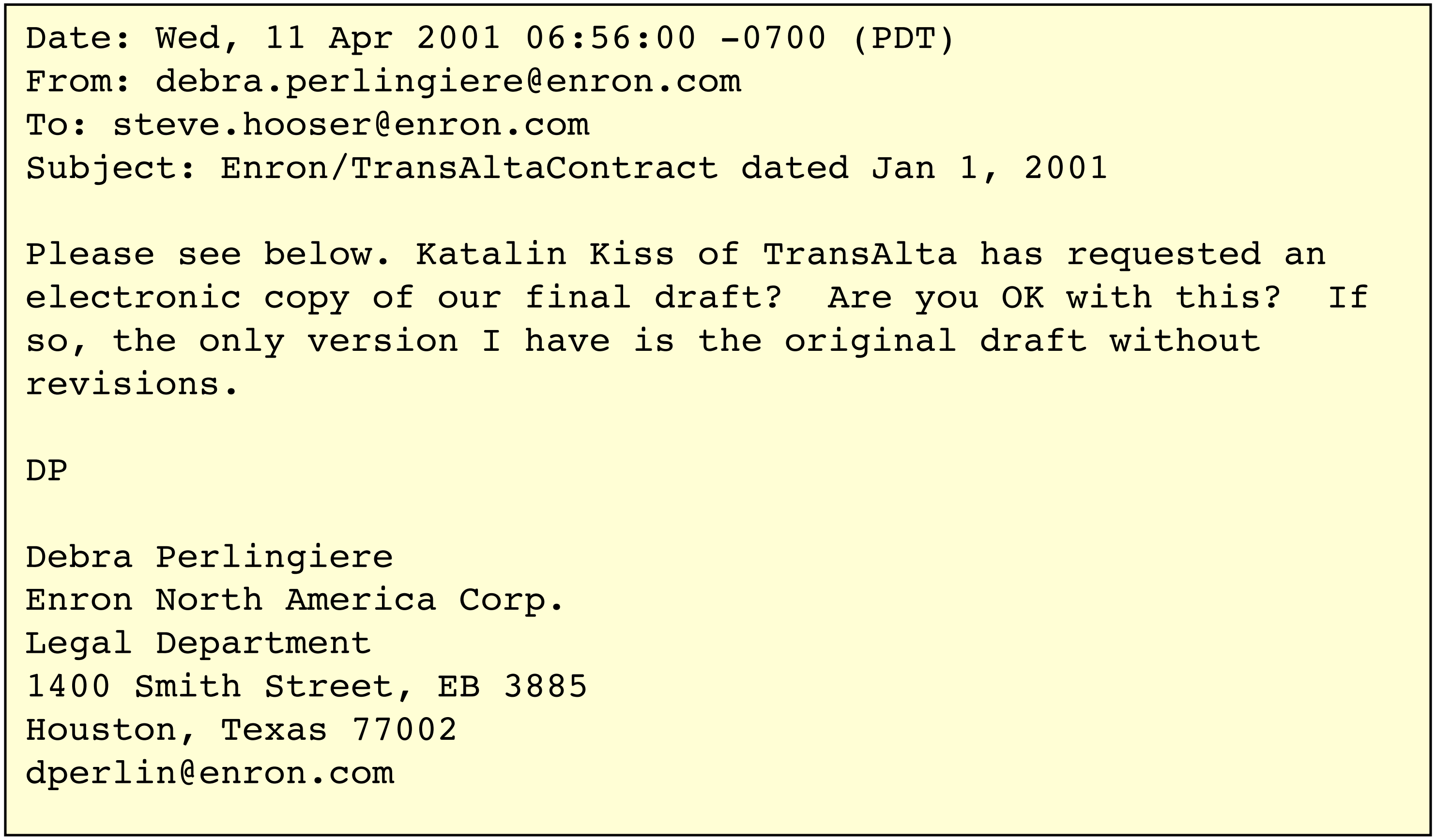
- Easy to implement
- Reasonably fast

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

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

Enron Email Corpus

- 250k email messages
- 23k people



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

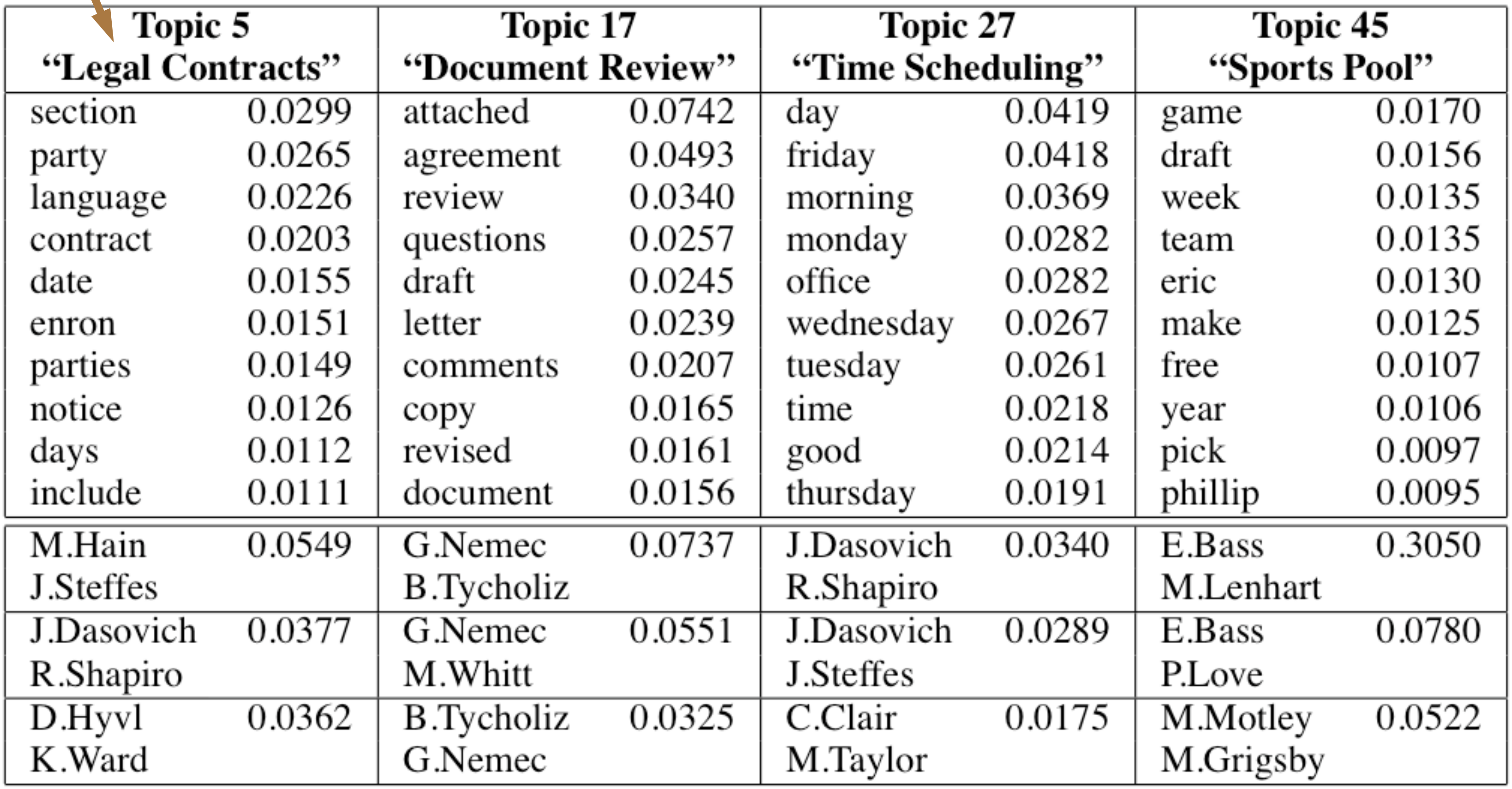
Please see below. Katalin Kiss of TransAlta has requested an electronic copy of our final draft? Are you OK with this? If so, the only version I have is the original draft without revisions.

DP

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

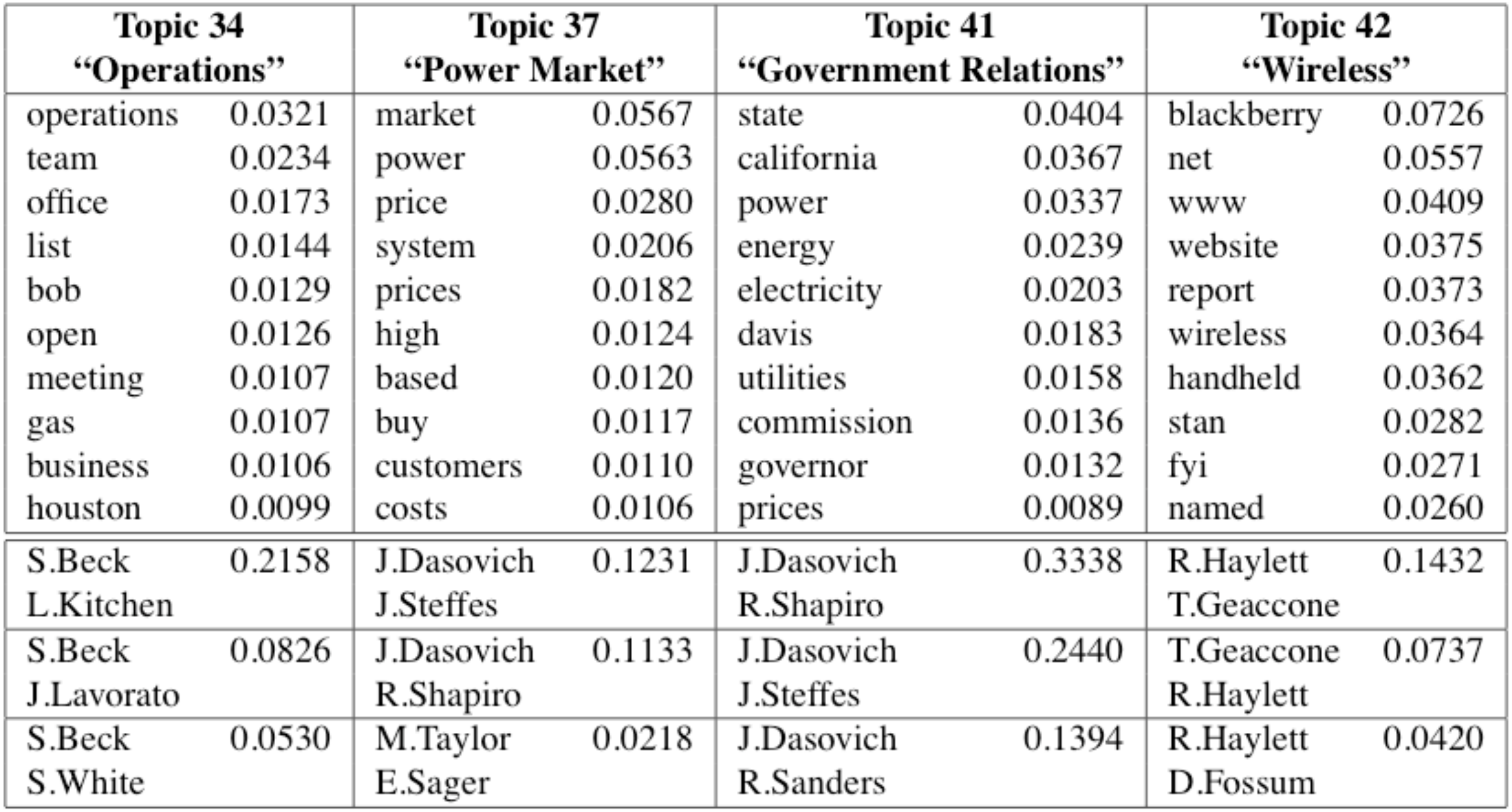
Topics, and prominent senders / receivers discovered by ART

Topic names, by hand



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

Topics, and prominent senders / receivers discovered by ART



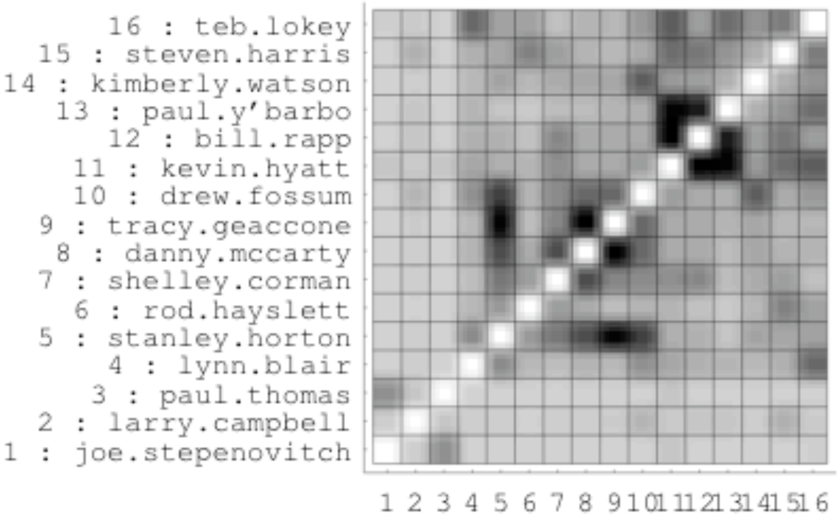
Topic 34 “Operations”		Topic 37 “Power Market”		Topic 41 “Government Relations”		Topic 42 “Wireless”	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck 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”

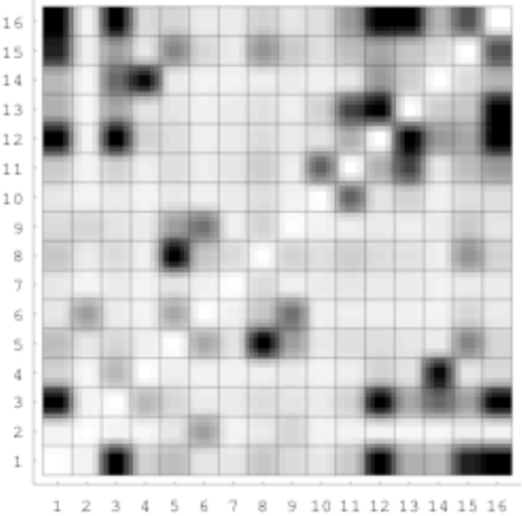
Dasovich = “Government Relations Executive”
Shapiro = “Vice President of Regulatory Affairs”
Steffes = “Vice President of Government Affairs”

Comparing Role Discovery

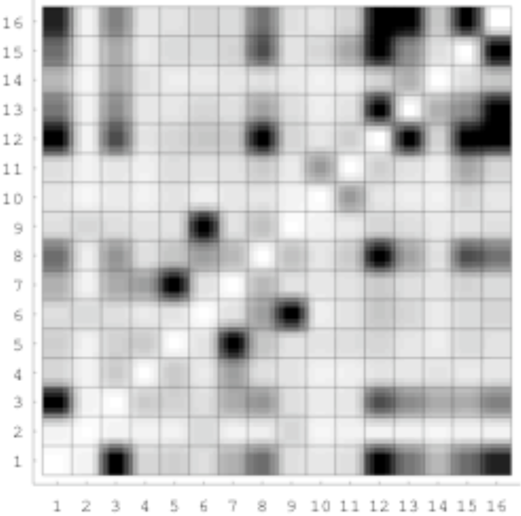
Traditional SNA



ART



Author-Topic



connection strength (A,B) =

distribution over recipients

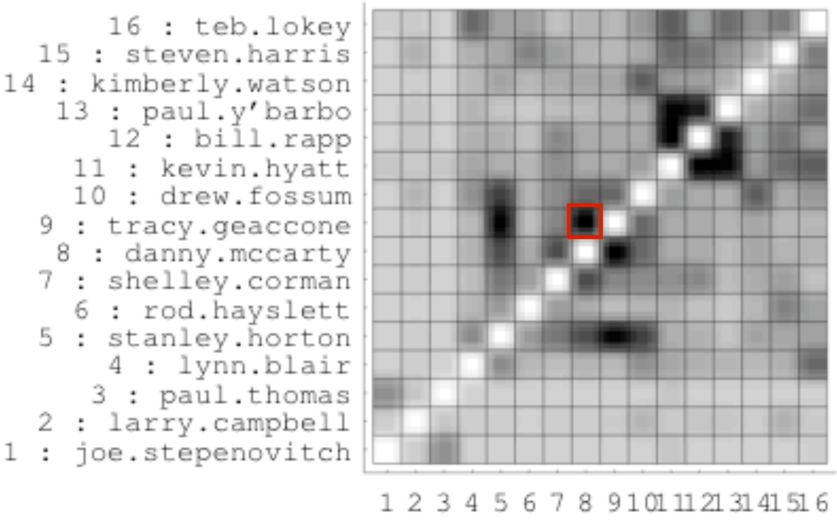
distribution over authored topics

distribution over authored topics

Comparing Role Discovery

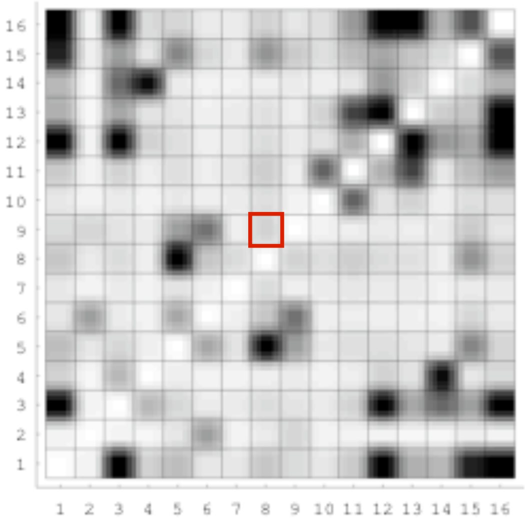
Tracy Geaconne ⇔ Dan McCarty

Traditional SNA



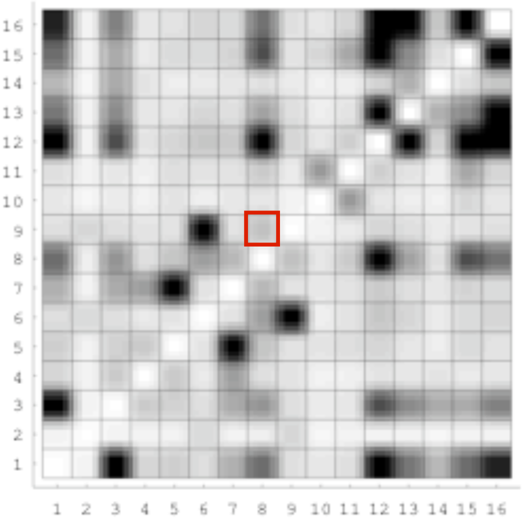
Similar roles

ART



Different roles

Author-Topic

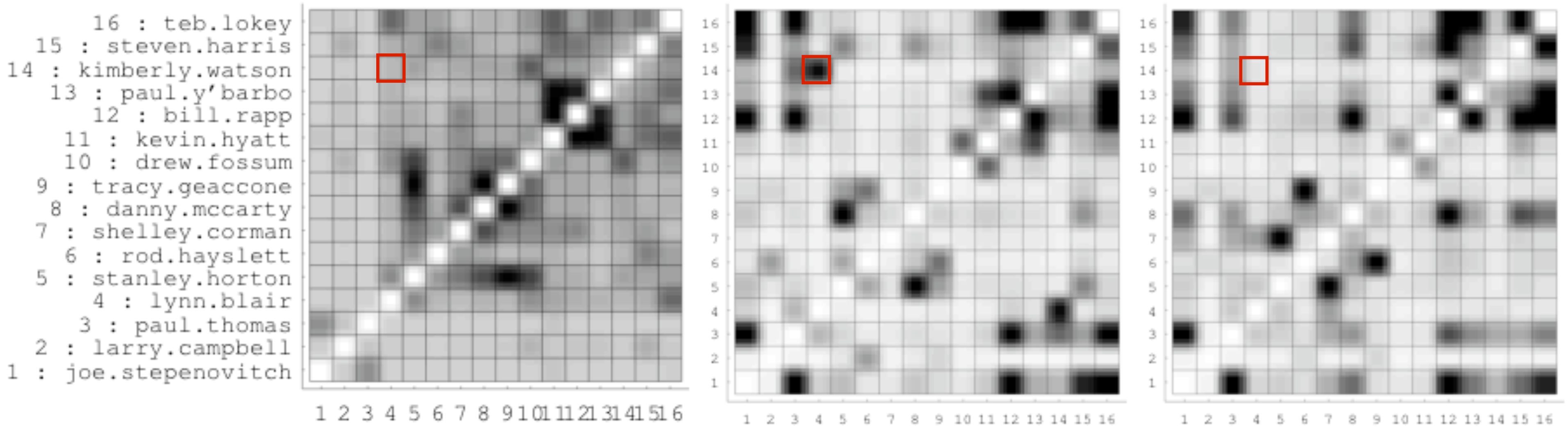


Different roles

Geaconne = “Secretary”
McCarty = “Vice President”

Comparing Role Discovery

Lynn Blair ⇔ Kimberly Watson



Traditional SNA — Different roles

ART — Very similar

Author-Topic — Very different

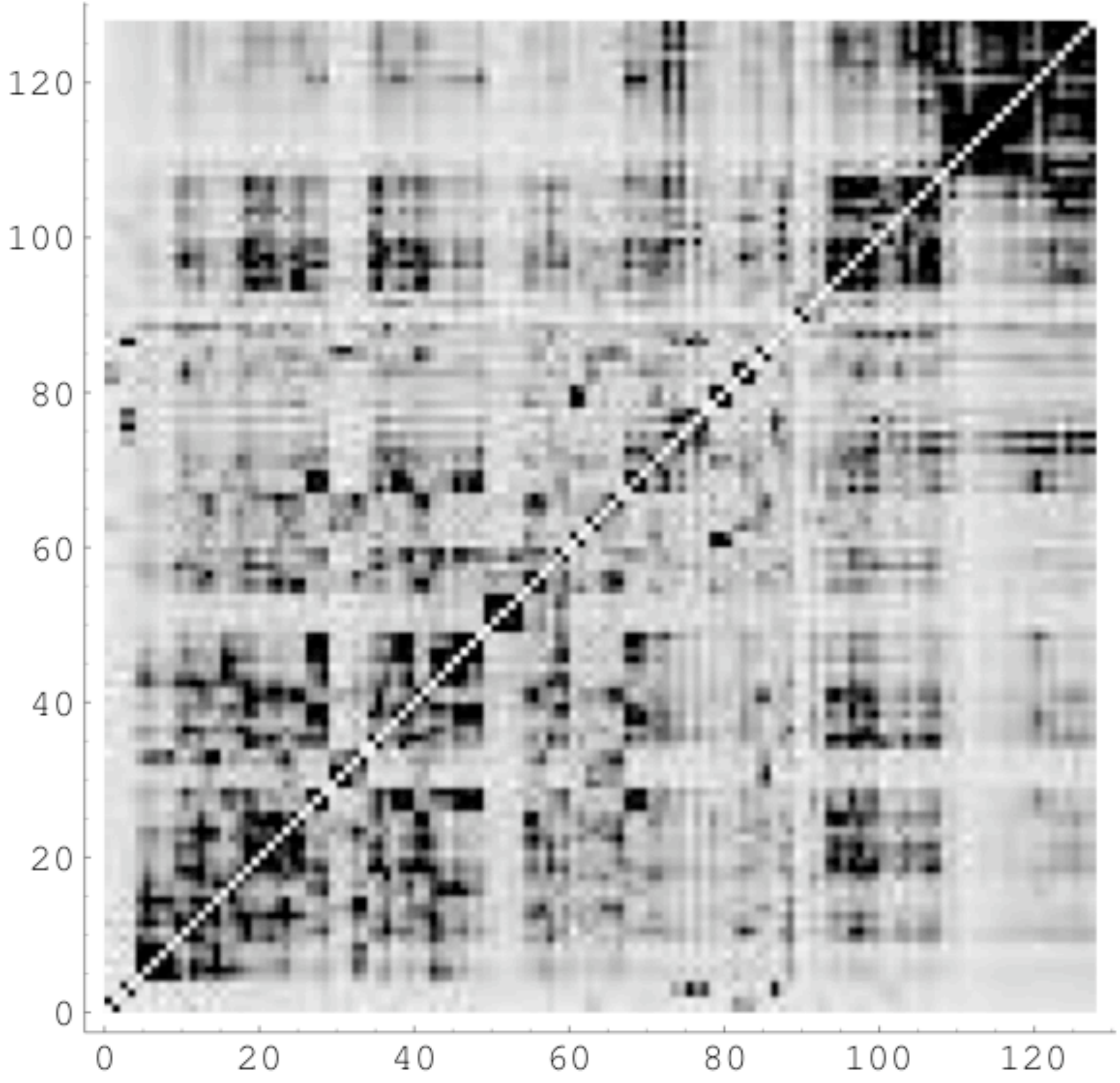
Blair = "Gas pipeline logistics"
Watson = "Pipeline facilities planning"

McCallum Email Corpus 2004

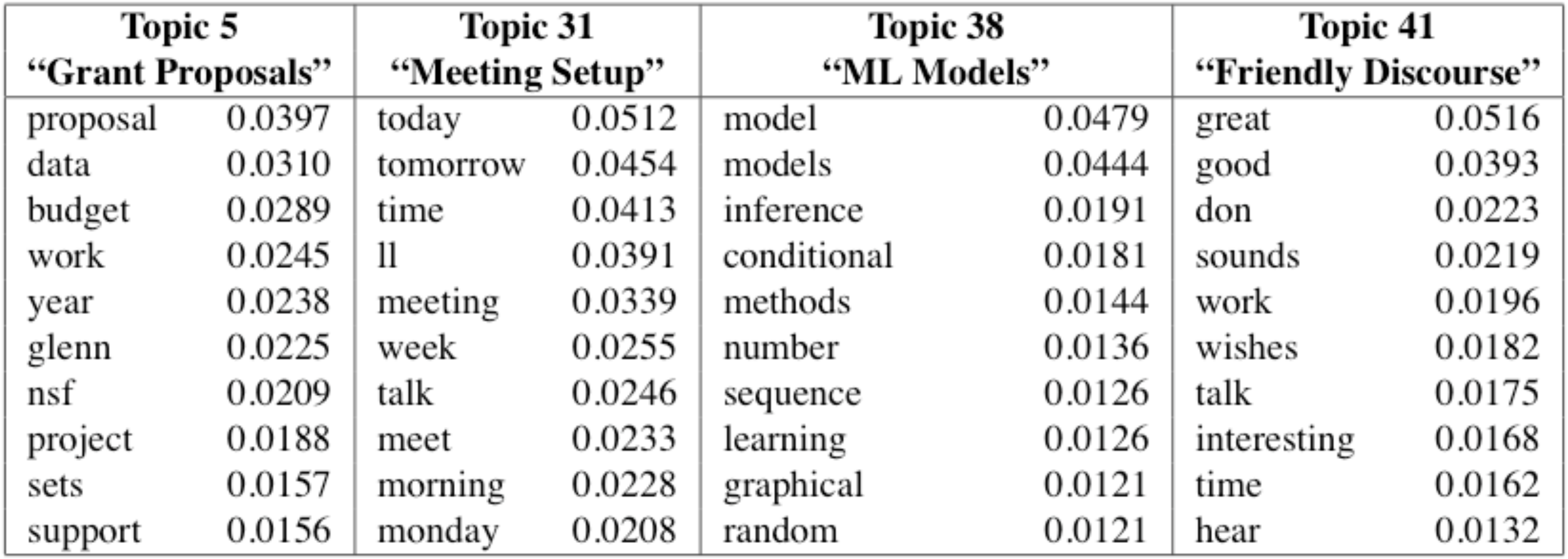
- January - October 2004
- 23k email messages
- 825 people

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

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

NIPS registration receipt.
CALO registration receipt.

Thanks,
Kate
```

McCallum Email Blockstructure



Four most prominent topics in discussions with ____?



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

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

Two most prominent topics in discussions with ____?

Topic 1

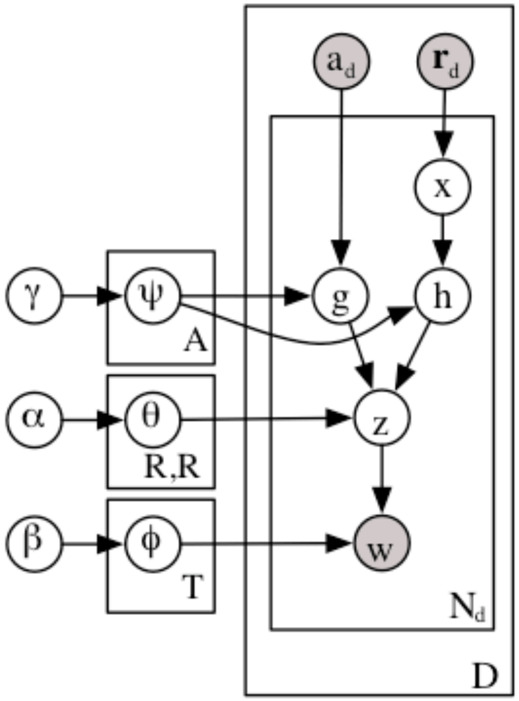
Words	Prob
love	0.030514
house	0.015402
	0.013659
time	0.012351
great	0.011334
hope	0.011043
dinner	0.00959
saturday	0.009154
left	0.009154
ll	0.009009
	0.008282
visit	0.008137
evening	0.008137
stay	0.007847
bring	0.007701
weekend	0.007411
road	0.00712
sunday	0.006829
kids	0.006539
flight	0.006539

Topic 2

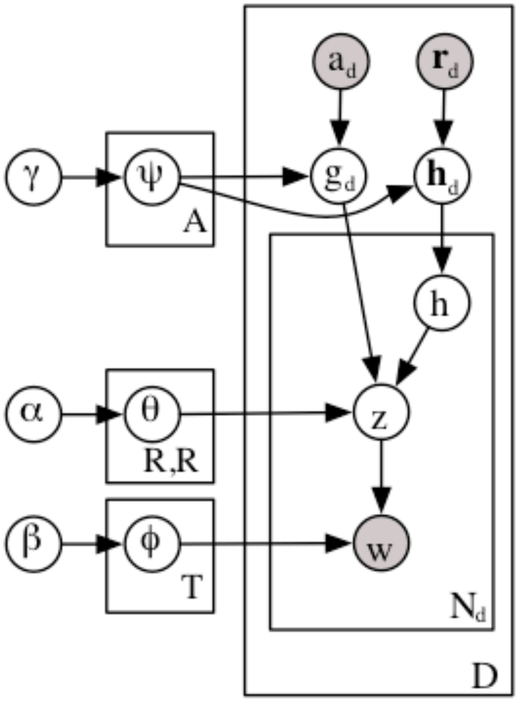
Words	Prob
today	0.051152
tomorrow	0.045393
time	0.041289
ll	0.039145
meeting	0.033877
week	0.025484
talk	0.024626
meet	0.023279
morning	0.022789
monday	0.020767
back	0.019358
call	0.016418
free	0.015621
home	0.013967
won	0.013783
day	0.01311
hope	0.012987
leave	0.012987
office	0.012742
tuesday	0.012558

Role-Author-Recipient-Topic Models

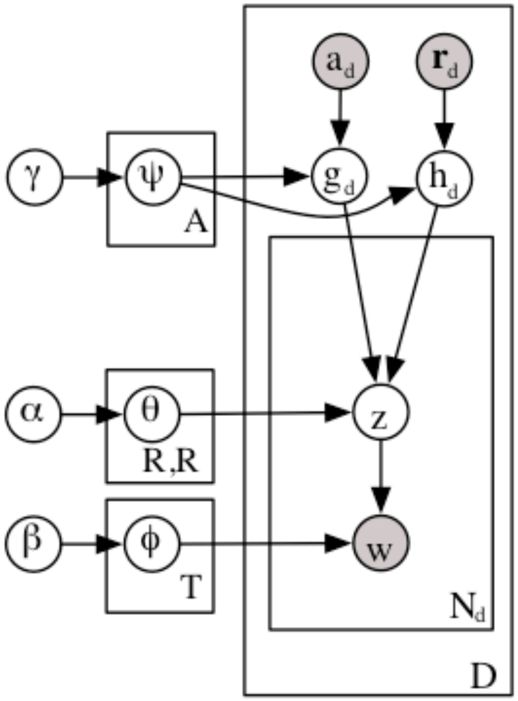
Role-Author-Recipient-Topic Model 1 (RART1)



Role-Author-Recipient-Topic Model 2 (RART2)



Role-Author-Recipient-Topic Model 3 (RART3)



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

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

Roles for `allan` (James Allan)

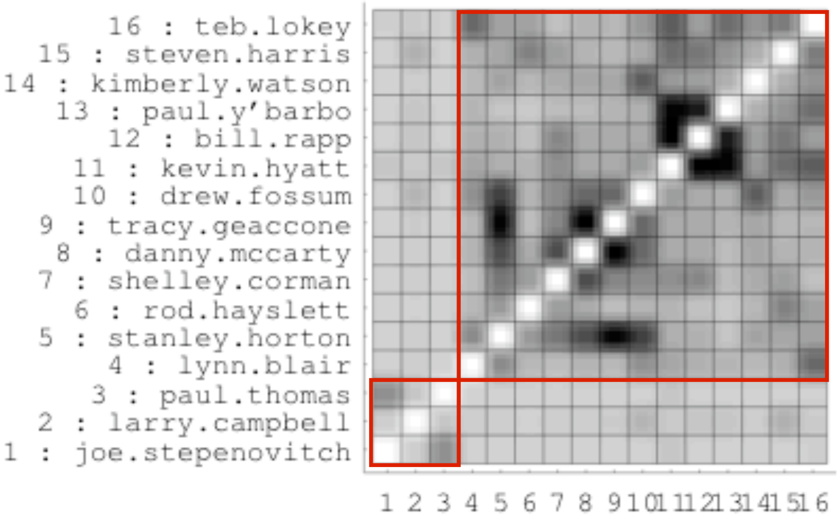
Role #3	I.T. support
Role #2	Natural Language researcher

Roles for `pereira` (Fernando Pereira)

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

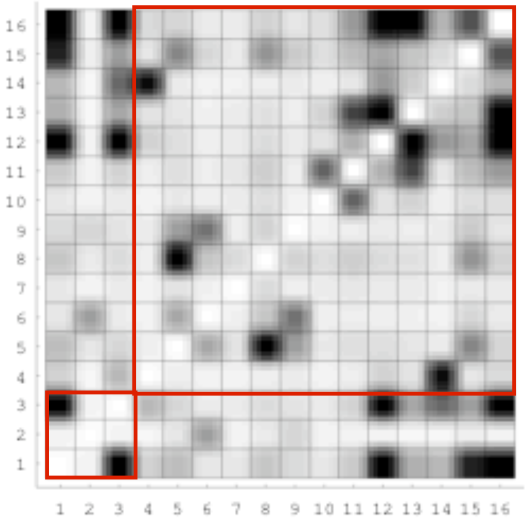
ART: Roles but not Groups

Traditional SNA



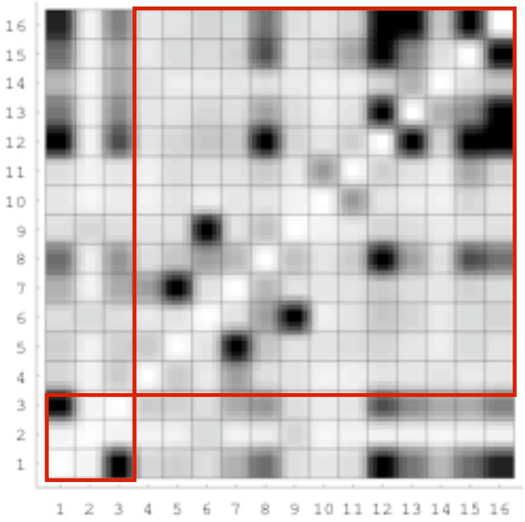
Block structured

ART



Not

Author-Topic



Not

Enron TransWestern Division

Social Network Analysis with Links *and Text*

Role Discovery

Group Discovery

Trend Discovery

Community Discovery

Impact Measurement

Groups and Topics

- Input:
  - Observed relations between people
  - Attributes on those relations (text, or categorical)

- Output:
  - Attributes clustered into “topics”
  - Groups of people---varying depending on topic

Adjacency Matrix Representing Relations

Student Roster	**Academic Admiration**
Adams	Acad(A, B) Acad(C, B)
Bennett	Acad(A, D) Acad(C, D)
Carter	Acad(B, E) Acad(D, E)
Davis	Acad(B, F) Acad(D, F)
Edwards	Acad(E, A) Acad(F, A)
Frederking	Acad(E, C) Acad(F, C)

	A	B	C	D	E	F
A		🟩		🟩		
B					🟩	🟩
C		🟩		🟩		
D					🟩	🟩
E	🟩		🟩			
F	🟩		🟩			

		A	B	C	D	E	F
		G1	**G2**	**G1**	**G2**	**G3**	**G3**
A	**G1**		🟩		🟩		
B	**G2**					🟩	🟩
C	**G1**		🟩		🟩		
D	**G2**					🟩	🟩
E	**G3**	🟩		🟩			
F	**G3**	🟩		🟩			

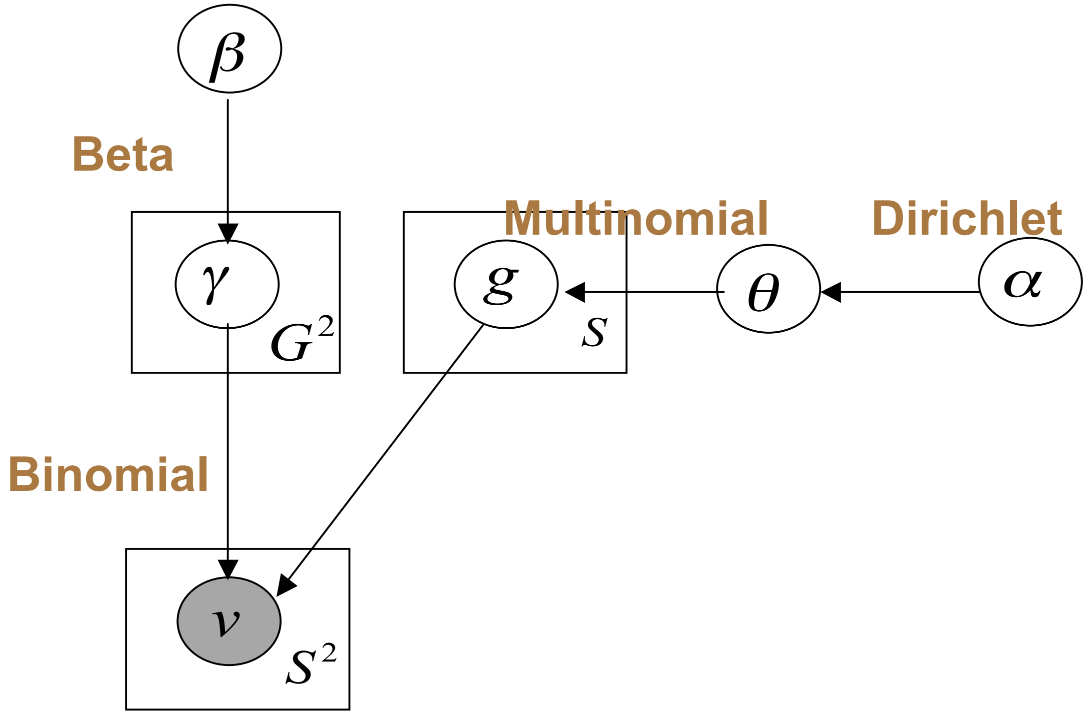
		A	C	B	D	E	F
		G1	**G1**	**G2**	**G2**	**G3**	**G3**
A	**G1**			🟩	🟩		
C	**G1**			🟩	🟩		
B	**G2**					🟩	🟩
D	**G2**					🟩	🟩
E	**G3**	🟩	🟩				
F	**G3**	🟩	🟩				

Group Model: Partitioning Entities into Groups

Stochastic Blockstructures for Relations
[Nowicki, Snijders 2001]

S: number of entities

G: number of groups



Enhanced with arbitrary number of groups in [Kemp, Griffiths, Tenenbaum 2004]

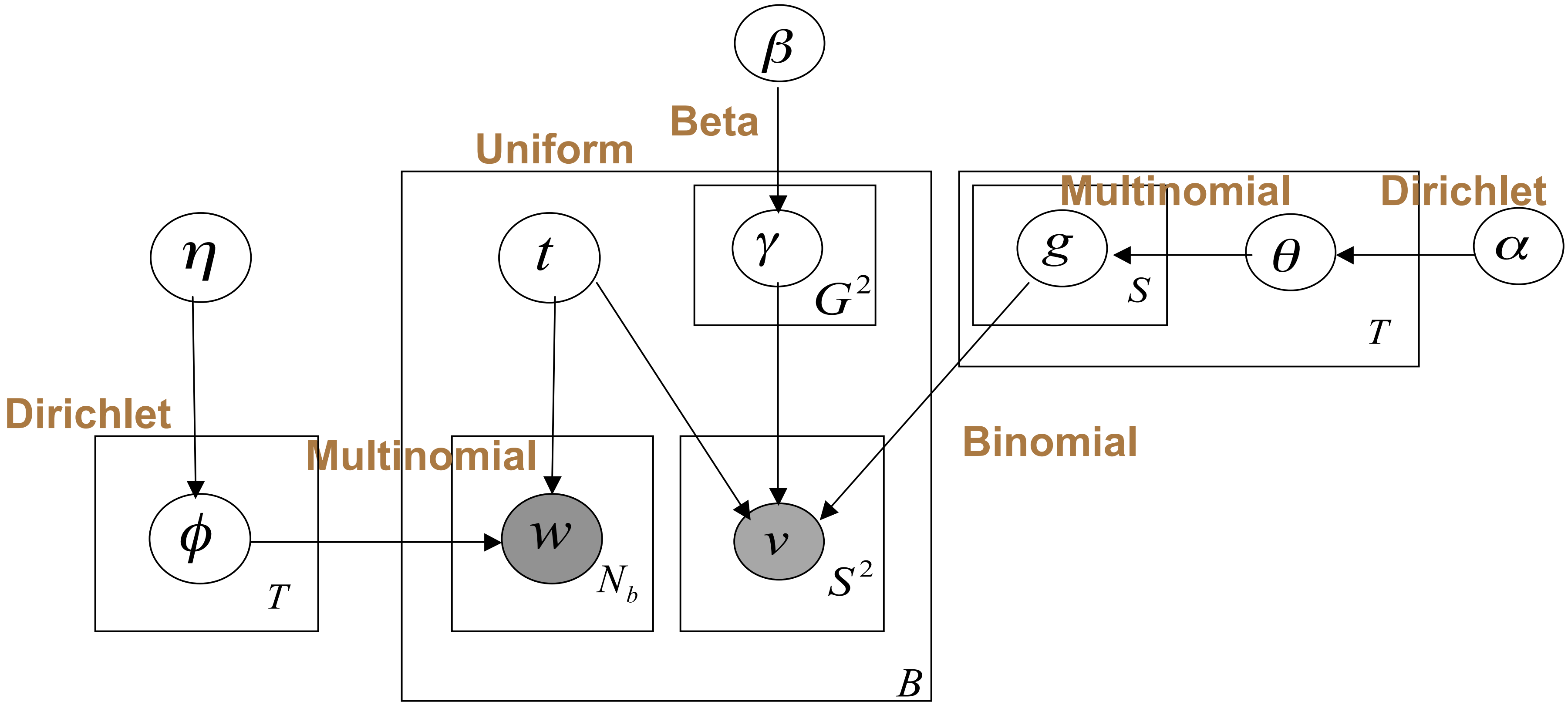
Two Relations with Different Attributes

Student Roster	Academic Admiration	Social Admiration
Adams	Acad(A, B) Acad(C, B)	Soci(A, B) Soci(A, D) Soci(A, F)
Bennett	Acad(A, D) Acad(C, D)	Soci(B, A) Soci(B, C) Soci(B, E)
Carter	Acad(B, E) Acad(D, E)	Soci(C, B) Soci(C, D) Soci(C, F)
Davis	Acad(B, F) Acad(D, F)	Soci(D, A) Soci(D, C) Soci(D, E)
Edwards	Acad(E, A) Acad(F, A)	Soci(E, B) Soci(E, D) Soci(E, F)
Frederking	Acad(E, C) Acad(F, C)	Soci(F, A) Soci(F, C) Soci(F, E)

		A	C	B	D	E	F
		G1	**G1**	**G2**	**G2**	**G3**	**G3**
A	**G1**			🟩	🟩		
C	**G1**			🟩	🟩		
B	**G2**					🟩	🟩
D	**G2**					🟩	🟩
E	**G3**	🟩	🟩				
F	**G3**	🟩	🟩				

		A	C	E	B	D	F
		G1	**G1**	**G1**	**G2**	**G2**	**G2**
A	**G1**				🟩	🟩	🟩
C	**G1**				🟩	🟩	🟩
E	**G1**				🟩	🟩	🟩
B	**G2**	🟩	🟩	🟩			
D	**G2**	🟩	🟩	🟩			
F	**G2**	🟩	🟩	🟩			

The Group-Topic Model:
Discovering Groups and Topics Simultaneously



Dataset #1:

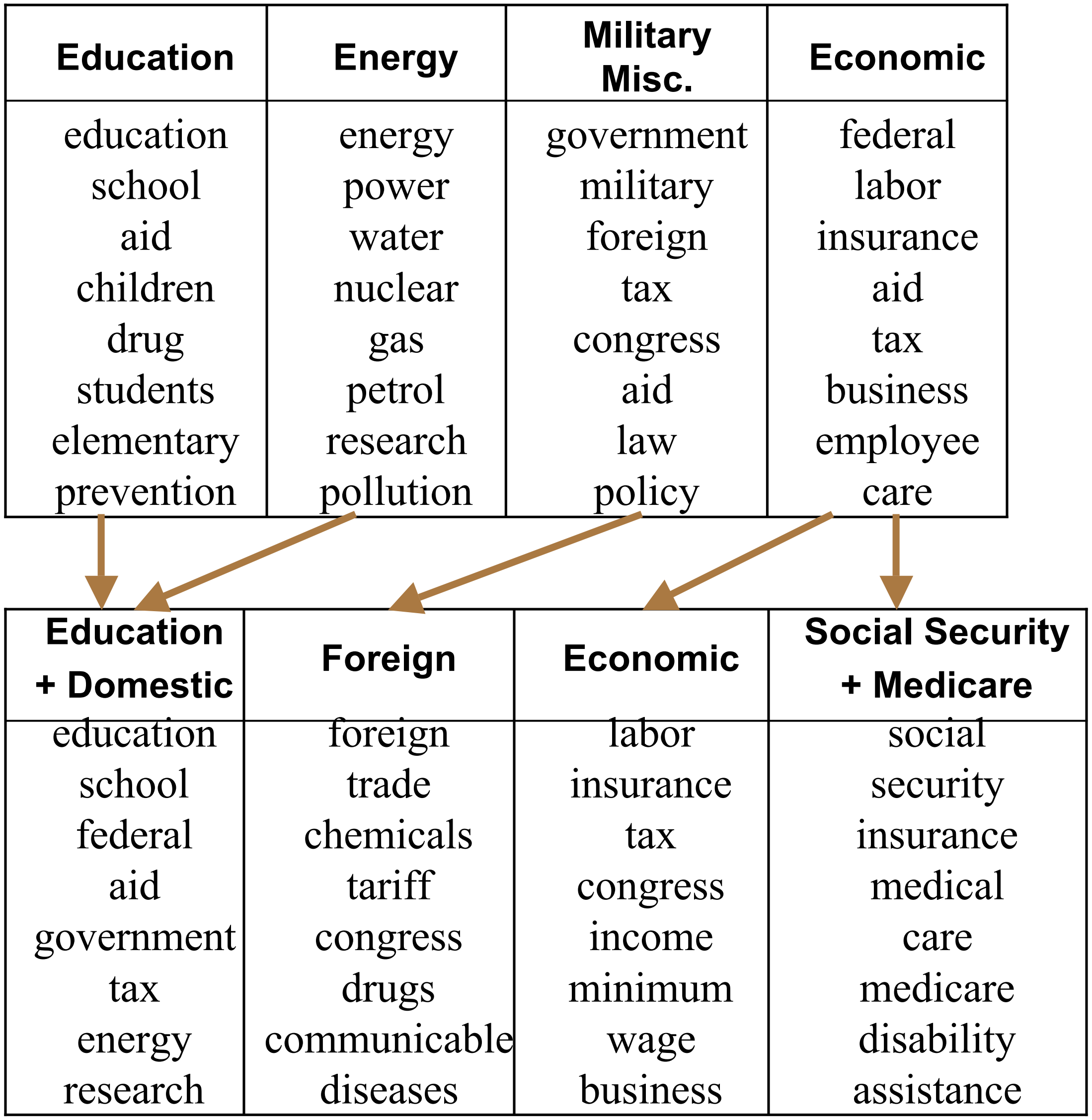
U.S. Senate

- **16 years of voting records in the US Senate (1989 – 2005)**
- **a Senator may respond *Yea* or *Nay* to a resolution**
- **3423 resolutions with text attributes (index terms)**
- **191 Senators in total across 16 years**

S.543
Title: An Act to reform Federal deposit insurance, protect the deposit insurance funds, recapitalize the Bank Insurance Fund, improve supervision and regulation of insured depository institutions, and for other purposes.
Sponsor: Sen Riegle, Donald W., Jr. [MI] (introduced 3/5/1991) Cosponsors (2)
Latest Major Action: 12/19/1991 Became Public Law No: 102-242.
Index terms: Banks and banking Accounting Administrative fees Cost control Credit Deposit insurance Depressed areas and other 110 terms

Adams (D-WA), **Nay** Akaka (D-HI), **Yea** Bentsen (D-TX), **Yea** Biden (D-DE), **Yea** Bond (R-MO), **Yea** Bradley (D-NJ), **Nay** Conrad (D-ND), **Nay ……**

Topics Discovered (U.S. Senate)



Mixture of Unigrams

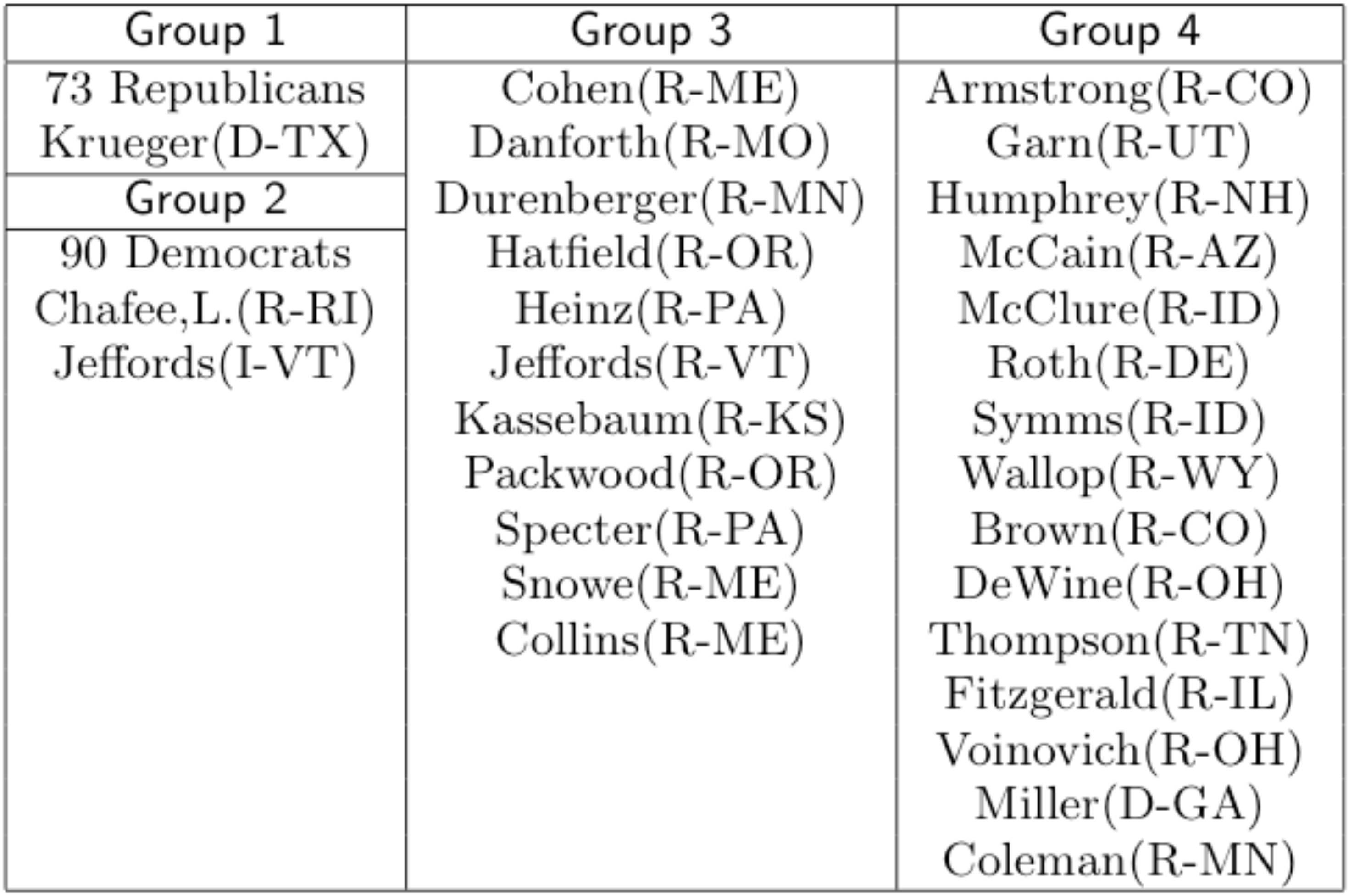
Education	Energy	Military Misc.	Economic
education	energy	government	federal
school	power	military	labor
aid	water	foreign	insurance
children	nuclear	tax	aid
drug	gas	congress	tax
students	petrol	aid	business
elementary	research	law	employee
prevention	pollution	policy	care

Group-Topic Model

Education + Domestic	Foreign	Economic	Social Security + Medicare
education	foreign	labor	social
school	trade	insurance	security
federal	chemicals	tax	insurance
aid	tariff	congress	medical
government	congress	income	care
tax	drugs	minimum	medicare
energy	communicable	wage	disability
research	diseases	business	assistance

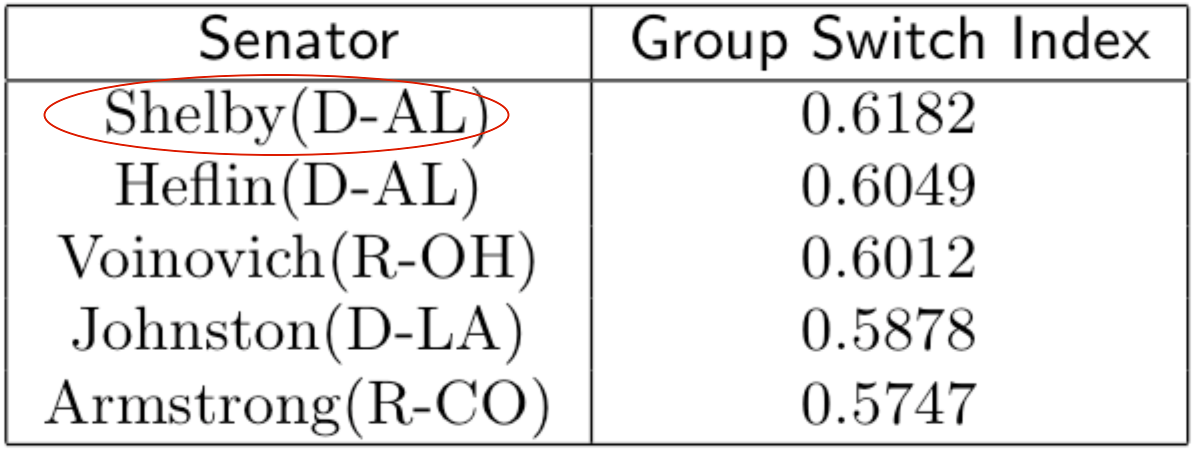
Groups Discovered (US Senate)

Groups from topic ***Education + Domestic***



Group 1	Group 3	Group 4
73 Republicans	Cohen(R-ME)	Armstrong(R-CO)
Krueger(D-TX)	Danforth(R-MO)	Garn(R-UT)
Group 2	Durenberger(R-MN)	Humphrey(R-NH)
90 Democrats	Hatfield(R-OR)	McCain(R-AZ)
Chafee,L.(R-RI)	Heinz(R-PA)	McClure(R-ID)
Jeffords(I-VT)	Jeffords(R-VT)	Roth(R-DE)
	Kassebaum(R-KS)	Symms(R-ID)
	Packwood(R-OR)	Wallop(R-WY)
	Specter(R-PA)	Brown(R-CO)
	Snowe(R-ME)	DeWine(R-OH)
	Collins(R-ME)	Thompson(R-TN)
		Fitzgerald(R-IL)
		Voinovich(R-OH)
		Miller(D-GA)
		Coleman(R-MN)

Senators Who Change Coalition the most Dependent on Topic



Senator	Group Switch Index
Shelby(D-AL)	0.6182
Heflin(D-AL)	0.6049
Voinovich(R-OH)	0.6012
Johnston(D-LA)	0.5878
Armstrong(R-CO)	0.5747

e.g. Senator Shelby (D-AL) votes
with the Republicans on Economic
with the Democrats on Education + Domestic
with a small group of maverick Republicans on Social Security + Medicaid

Dataset #2:

The UN General Assembly

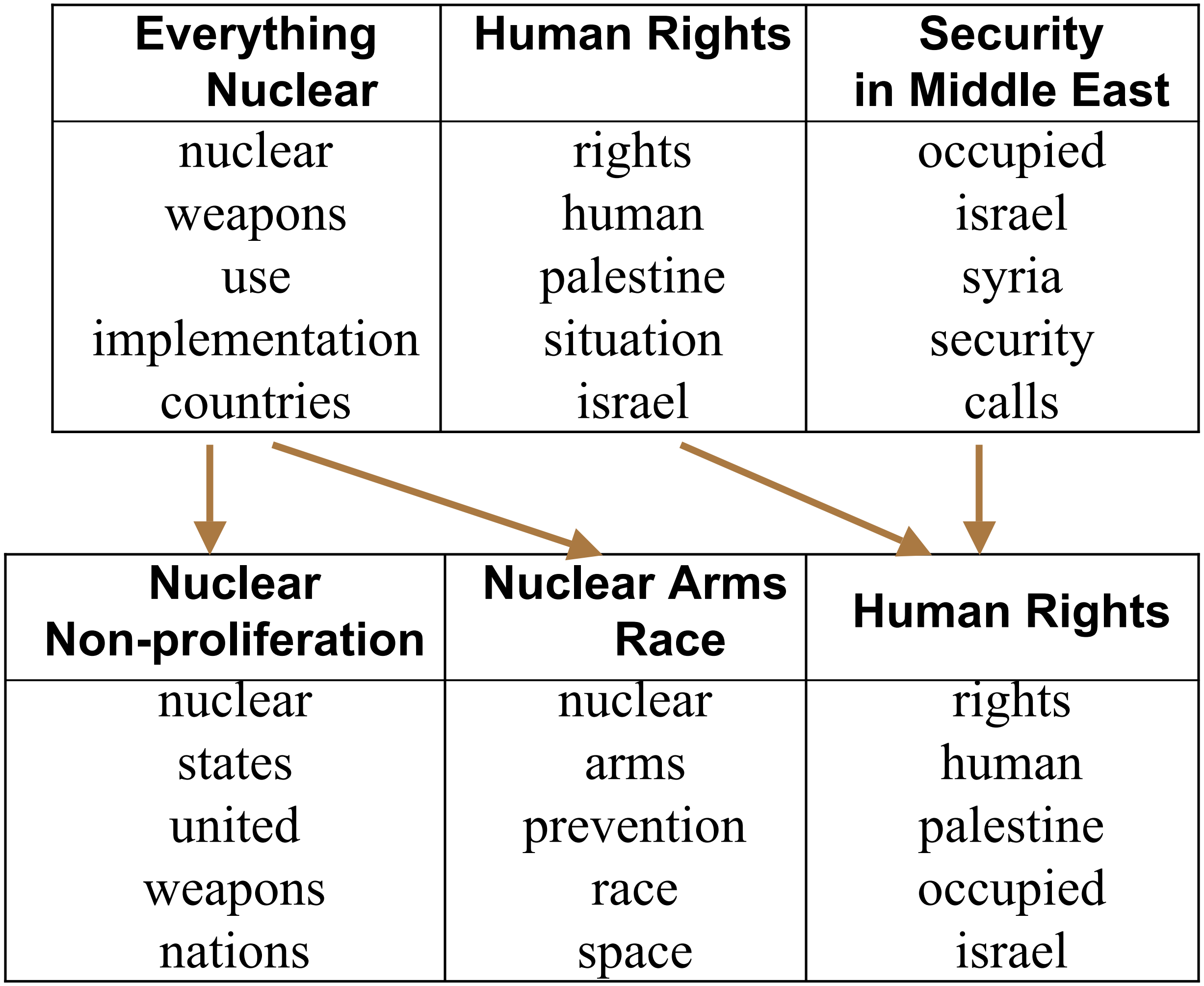
- Voting records of the UN General Assembly (1990 - 2003)
- A country may choose to vote *Yes*, *No* or *Abstain*
- 931 resolutions with text attributes (titles)
- 192 countries in total
- Also experiments later with resolutions from 1960-2003

Vote on Permanent Sovereignty of Palestinian People, 87th plenary meeting

The draft resolution on permanent sovereignty of the Palestinian people in the occupied Palestinian territory, including Jerusalem, and of the Arab population in the occupied Syrian Golan over their natural resources (document A/54/591) was adopted by a recorded vote of 145 in favour to 3 against with 6 abstentions:

In favour: Afghanistan, Argentina, Belgium, Brazil, Canada, China, France, Germany, India, Japan, Mexico, Netherlands, New Zealand, Pakistan, Panama, Russian Federation, South Africa, Spain, Turkey, and other 126 countries.
Against: Israel, Marshall Islands, United States.
Abstain: Australia, Cameroon, Georgia, Kazakhstan, Uzbekistan, Zambia.

Topics Discovered (UN)



Mixture of Unigrams

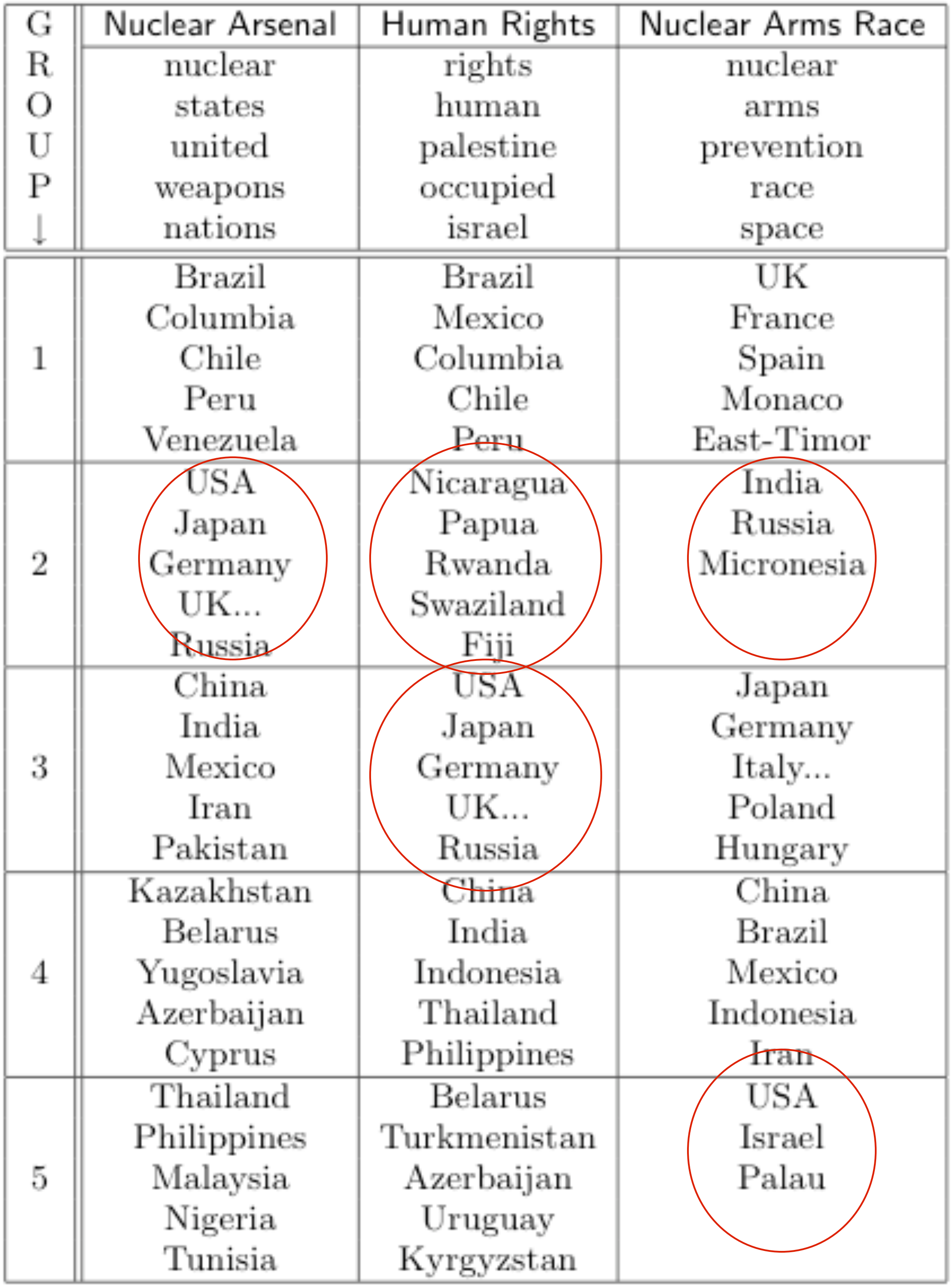
Everything Nuclear	Human Rights	Security in Middle East
nuclear	rights	occupied
weapons	human	israel
use	palestine	syria
implementation	situation	security
countries	israel	calls

Group-Topic Model

Nuclear Non-proliferation	Nuclear Arms Race	Human Rights
nuclear	nuclear	rights
states	arms	human
united	prevention	palestine
weapons	race	occupied
nations	space	israel

Groups Discovered (UN)

The countries list for each group are ordered by their 2005 GDP (PPP) and only 5 countries are shown in groups that have more than 5 members.



G	Nuclear Arsenal	Human Rights	Nuclear Arms Race
R	nuclear	rights	nuclear
O	states	human	arms
U	united	palestine	prevention
P	weapons	occupied	race
↓	nations	israel	space
1	Brazil	Brazil	UK
	Columbia	Mexico	France
	Chile	Columbia	Spain
	Peru	Chile	Monaco
	Venezuela	Peru	East-Timor
2	USA	Nicaragua	India
	Japan	Papua	Russia
	Germany	Rwanda	Micronesia
	UK...	Swaziland	
	Russia	Fiji	
3	China	USA	Japan
	India	Japan	Germany
	Mexico	Germany	Italy...
	Iran	UK...	Poland
	Pakistan	Russia	Hungary
4	Kazakhstan	China	China
	Belarus	India	Brazil
	Yugoslavia	Indonesia	Mexico
	Azerbaijan	Thailand	Indonesia
	Cyprus	Philippines	Iran
5	Thailand	Belarus	USA
	Philippines	Turkmenistan	Israel
	Malaysia	Azerbaijan	Palau
	Nigeria	Uruguay	
	Tunisia	Kyrgyzstan	

Groups and Topics, Trends over Time (UN)

Time Period	Topic 1	Topic 2	Topic 3	Group distributions for Topic 3				
				Group 1	Group2	Group3	Group4	Group5
60-75	Nuclear operative general nuclear power	Procedure committee amendment assembly deciding	Africa Indep. calling right africa self	India Indonesia Iran Thailand Philippines	USA Japan UK France Italy	Argentina Colombia Chile Venezuela Dominican	USSR Poland Hungary Bulgaria Belarus	Turkey
65-80	Independence territories independence self colonial	Finance budget appropriation contribution income	Weapons nuclear UN international weapons	Cuba Albania	India Indonesia Pakistan Saudi Egypt	Algeria Iraq Syria Libya Afganistan	USSR Poland Hungary Bulgaria Belarus	USA Japan UK France Italy
70-85	N. Weapons nuclear international UN human	Israel israel measures hebron expelling	Rights africa territories south right	Mexico Indonesia Iran Thailand Philippines	China	USA Japan UK France Italy	Brazil Turkey Argentina Colombia Chile	India USSR Poland Vietnam Hungary
75-90	Rights south africa israel rights	Israel/Pal. israel arab occupied palestine	Disarmament UN international nuclear disarmament	Mexico Indonesia Iran Thailand Philippines	USA Japan UK France USSR	Algeria Vietnam Iraq Syria Libya	China Brazil Argentina Colombia Chile	India
80-95	Disarmament nuclear US disarmament international	Conflict need israel palestine secretary	Pal. Rights rights palestine israel occupied	USA Israel	China India Russia Spain Hungary	Japan UK France Italy Canada	Guatemala St Vincent Dominican	Malawi
85-00	Weapons nuclear weapons use international	Rights rights human fundamental freedoms	Israel/Pal. israeli palestine occupied disarmament	Poland Czech R. Hungary Bulgaria Albania	China India Brazil Mexico Indonesia	USA Japan UK France Italy	Russia Argentina Ukraine Belarus Malta	Cameroon Congo Ivory C. Liberia

Social Network Analysis with Links *and Text*

Role Discovery

Group Discovery

Trend Discovery

Community Discovery

Impact Measurement

Groups and Topics, Trends over Time (UN)

Time				Group distributions for Topic 3				
Period	Topic 1	Topic 2	Topic 3	Group 1	Group2	Group3	Group4	Group5
60-75	Nuclear	Procedure	Africa Indep.	India	USA	Argentina	USSR	Turkey
	operative	committee	calling	Indonesia	Japan	Colombia	Poland	
	general	amendment	right	Iran	UK	Chile	Hungary	
	nuclear	assembly	africa	Thailand	France	Venezuela	Bulgaria	
	power	deciding	self	Philippines	Italy	Dominican	Belarus	
65-80	Independence	Finance	Weapons	Cuba	India	Algeria	USSR	USA
	territories	budget	nuclear	Albania	Indonesia	Iraq	Poland	Japan
	independence	appropriation	UN		Pakistan	Syria	Hungary	UK
	self	contribution	international		Saudi	Libya	Bulgaria	France
	colonial	income	weapons		Egypt	Afganistan	Belarus	Italy
70-85	N. Weapons	Israel	Rights	Mexico	China	USA	Brazil	India
	nuclear	israel	africa	Indonesia		Japan	Turkey	USSR
	international	measures	territories	Iran		UK	Argentina	Poland
	UN	hebron	south	Thailand		France	Colombia	Vietnam
	human	expelling	right	Philippines		Italy	Chile	Hungary
75-90	Rights	Israel/Pal.	Disarmament	Mexico	USA	Algeria	China	India
	south	israel	UN	Indonesia	Japan	Vietnam	Brazil	
	africa	arab	international	Iran	UK	Iraq	Argentina	
	israel	occupied	nuclear	Thailand	France	Syria	Colombia	
	rights	palestine	disarmament	Philippines	USSR	Libya	Chile	
80-95	Disarmament	Conflict	Pal. Rights	USA	China	Japan	Guatemala	Malawi
	nuclear	need	rights	Israel	India	UK	St Vincent	
	US	israel	palestine		Russia	France	Dominican	
	disarmament	palestine	israel		Spain	Italy		
	international	secretary	occupied		Hungary	Canada		
85-00	Weapons	Rights	Israel/Pal.	Poland	China	USA	Russia	Cameroon
	nuclear	rights	israeli	Czech R.	India	Japan	Argentina	Congo
	weapons	human	palestine	Hungary	Brazil	UK	Ukraine	Ivory C.
	use	fundamental	occupied	Bulgaria	Mexico	France	Belarus	Liberia
	international	freedoms	disarmament	Albania	Indonesia	Italy	Malta	

Want to Model Trends over Time

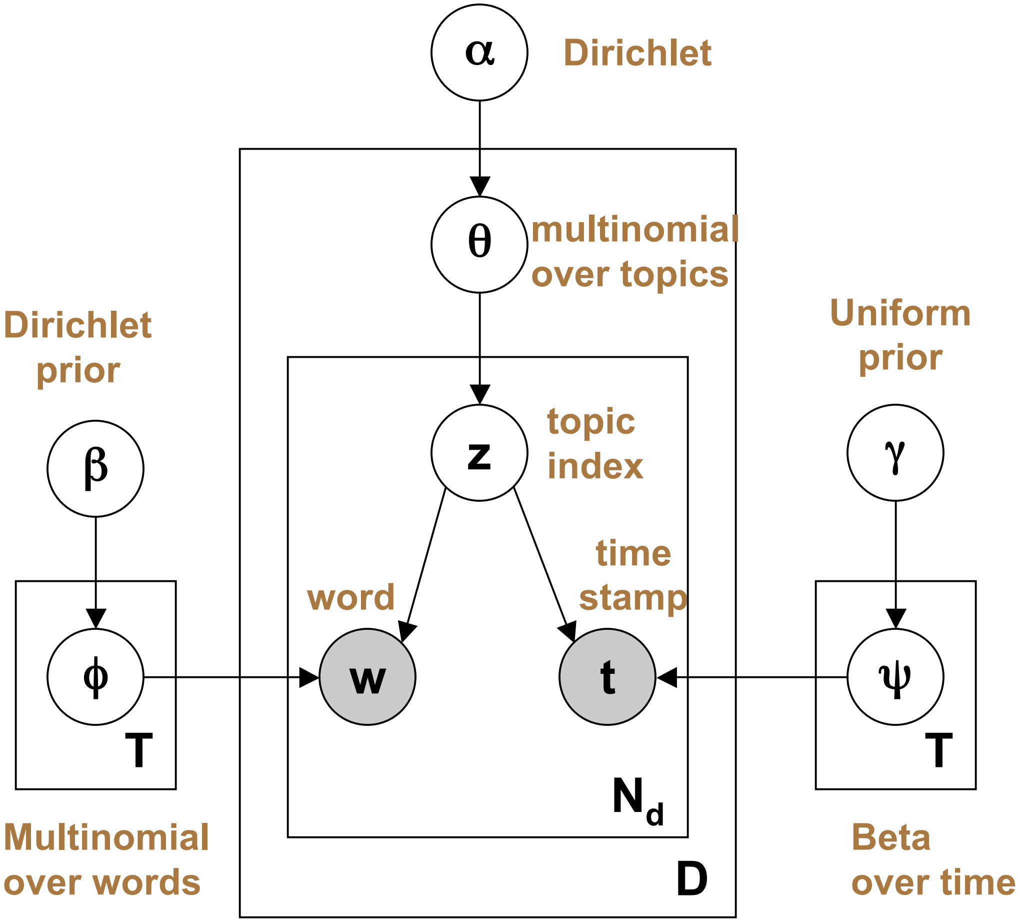
- Pattern appears only briefly
  - Capture its statistics in focused way
  - Don’t confuse it with patterns elsewhere in time

- Is prevalence of topic growing or waning?

- How do roles, groups, influence shift over time?

Topics over Time (TOT)

[Wang, McCallum, KDD 2006]



State of the Union Address

208 Addresses delivered between January 8, 1790 and January 29, 2002.

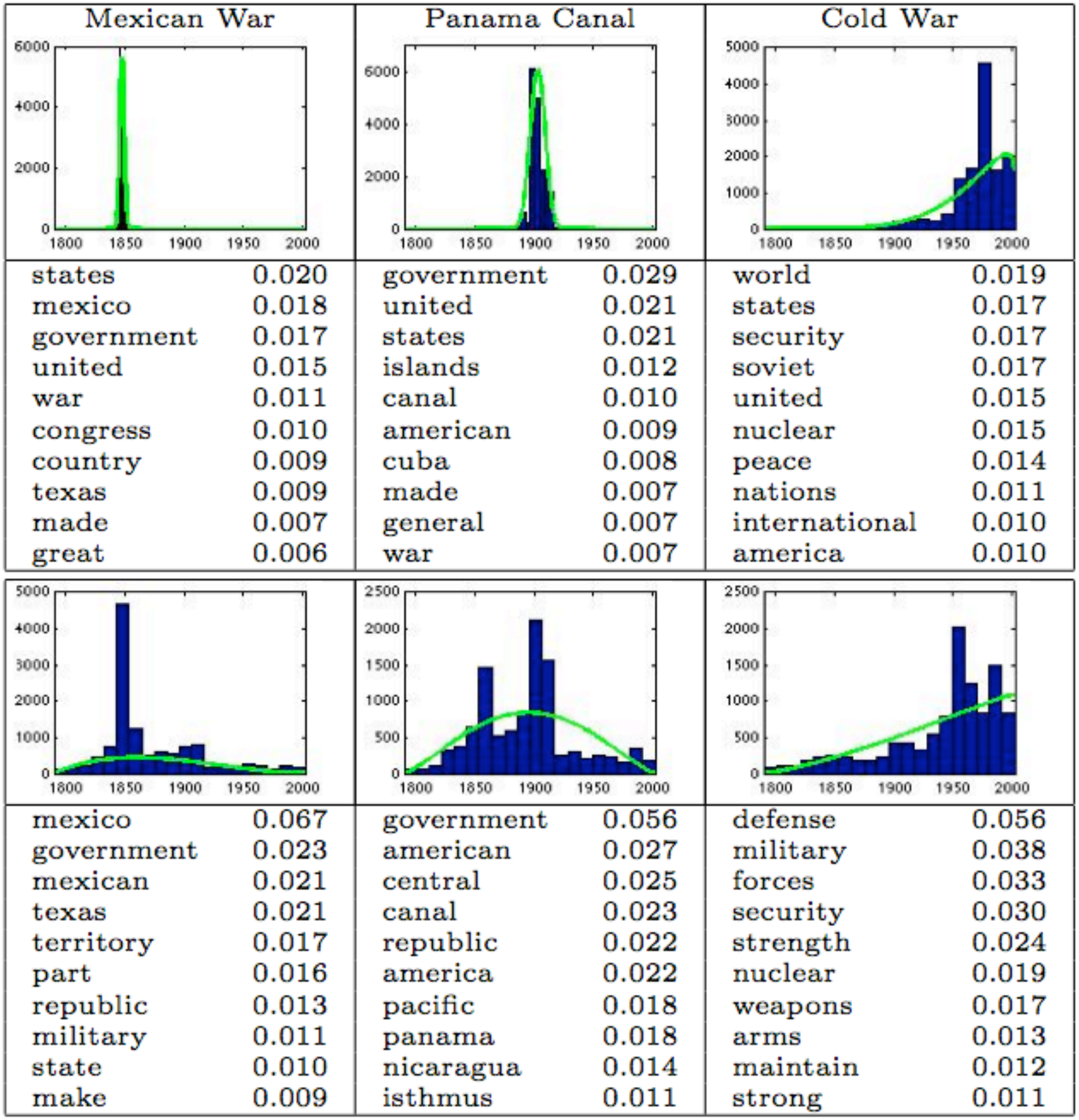
To increase the number of documents, we split the addresses into paragraphs and treated them as ‘documents’. One-line paragraphs were excluded. Stopping was applied.

- **17156 ‘documents’**
- **21534 words**
- **669,425 tokens**

> Our scheme of taxation, by means of which this needless surplus is taken from the people and put into the public Treasury, consists of a tariff or duty levied upon importations from abroad and internal-revenue taxes levied upon the consumption of tobacco and spirituous and malt liquors. It must be conceded that none of the things subjected to internal-revenue taxation are, strictly speaking, necessaries. There appears to be no just complaint of this taxation by the consumers of these articles, and there seems to be nothing so well able to bear the burden without hardship to any portion of the people.

1910

Comparing TOT against LDA



Mexican War		Panama Canal		Cold War	
states	0.020	government	0.029	world	0.019
mexico	0.018	united	0.021	states	0.017
government	0.017	states	0.021	security	0.017
united	0.015	islands	0.012	soviet	0.017
war	0.011	canal	0.010	united	0.015
congress	0.010	american	0.009	nuclear	0.015
country	0.009	cuba	0.008	peace	0.014
texas	0.009	made	0.007	nations	0.011
made	0.007	general	0.007	international	0.010
great	0.006	war	0.007	america	0.010

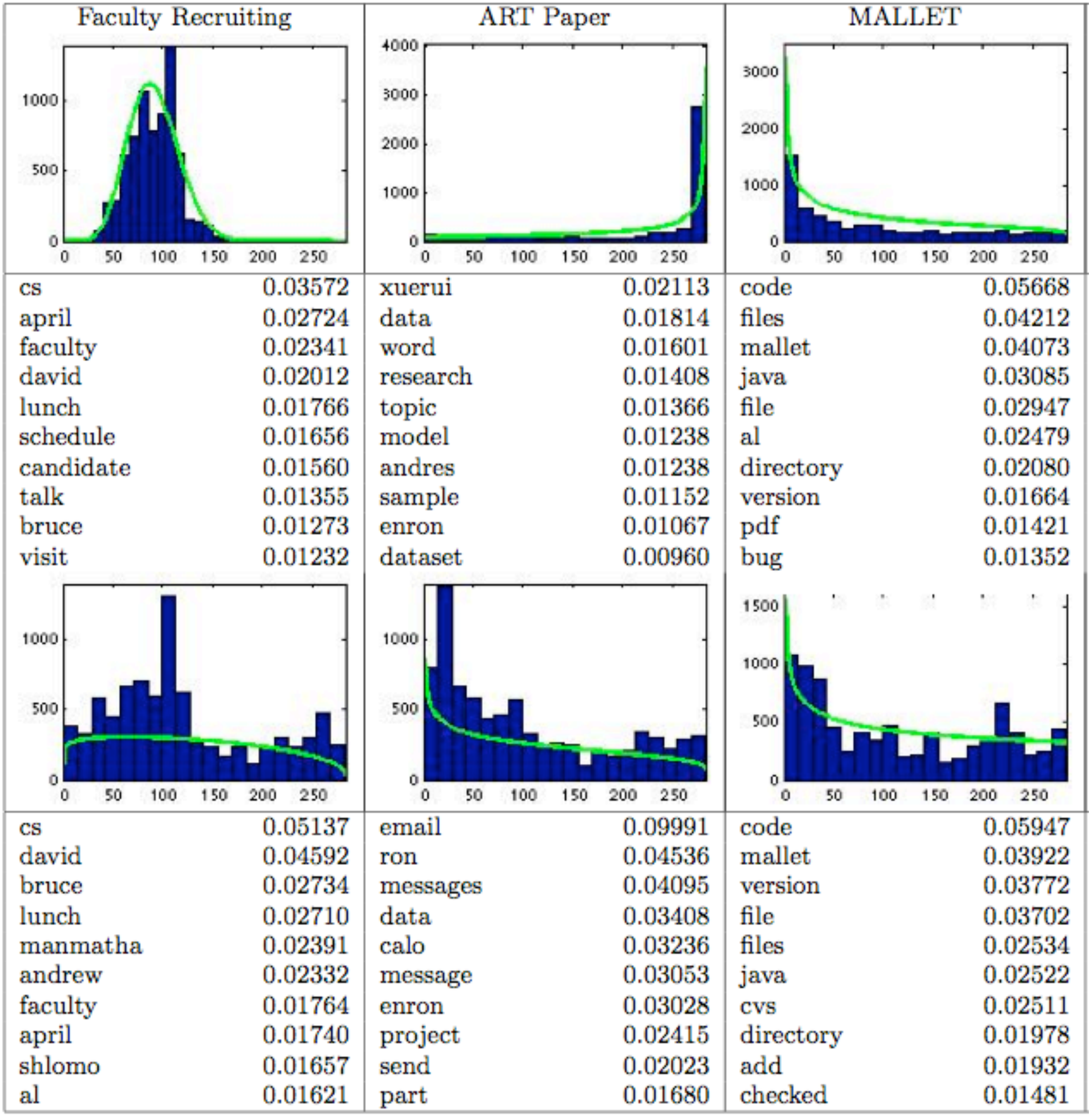
mexico	0.067	government	0.056	defense	0.056
government	0.023	american	0.027	military	0.038
mexican	0.021	central	0.025	forces	0.033
texas	0.021	canal	0.023	security	0.030
territory	0.017	republic	0.022	strength	0.024
part	0.016	america	0.022	nuclear	0.019
republic	0.013	pacific	0.018	weapons	0.017
military	0.011	panama	0.018	arms	0.013
state	0.010	nicaragua	0.014	maintain	0.012
make	0.009	isthmus	0.011	strong	0.011

TOT

versus

LDA

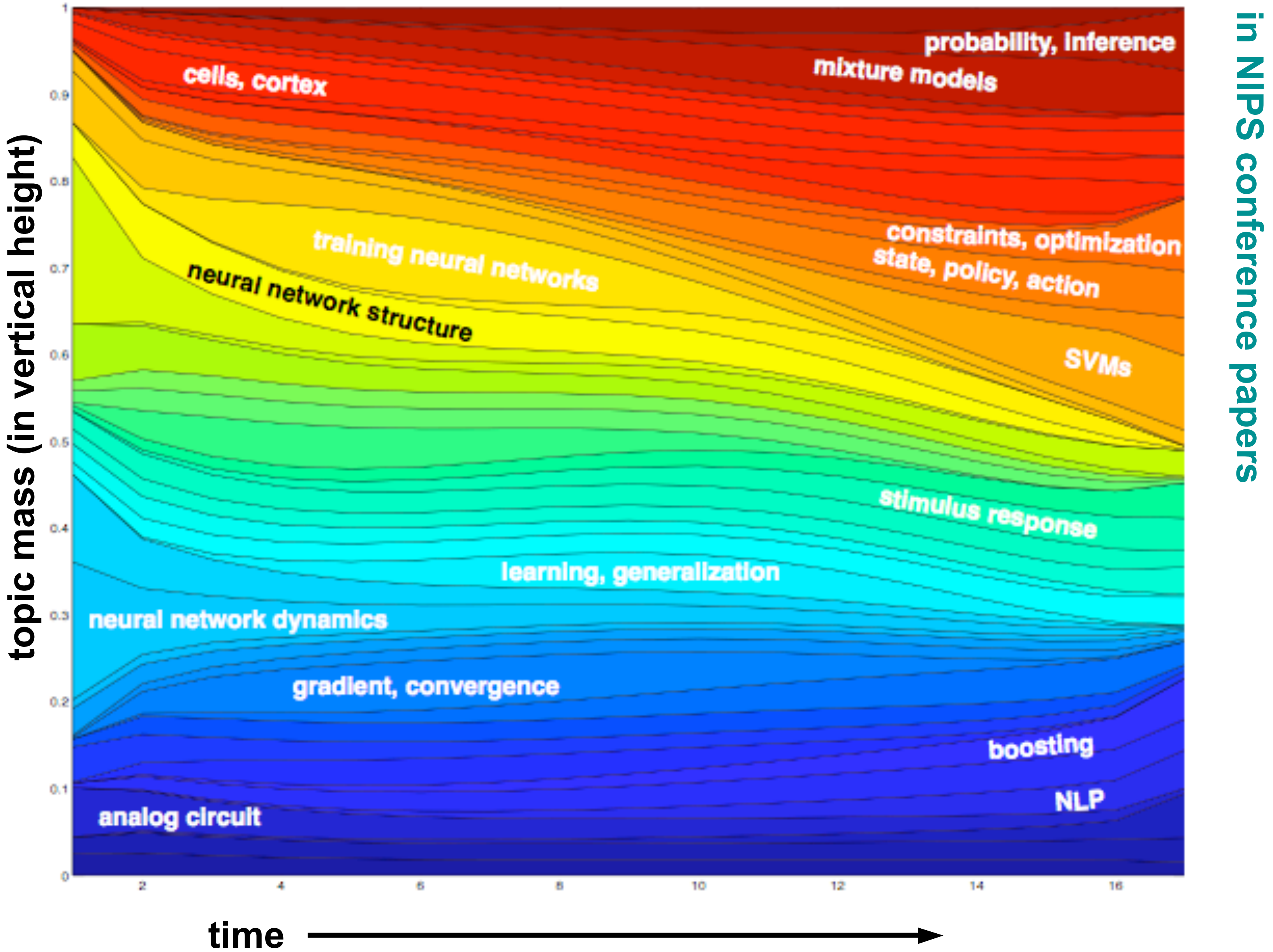
on my email



Faculty Recruiting		ART Paper		MALLET	
cs	0.03572	xuerui	0.02113	code	0.05668
april	0.02724	data	0.01814	files	0.04212
faculty	0.02341	word	0.01601	mallet	0.04073
david	0.02012	research	0.01408	java	0.03085
lunch	0.01766	topic	0.01366	file	0.02947
schedule	0.01656	model	0.01238	al	0.02479
candidate	0.01560	andres	0.01238	directory	0.02080
talk	0.01355	sample	0.01152	version	0.01664
bruce	0.01273	enron	0.01067	pdf	0.01421
visit	0.01232	dataset	0.00960	bug	0.01352

cs	0.05137	email	0.09991	code	0.05947
david	0.04592	ron	0.04536	mallet	0.03922
bruce	0.02734	messages	0.04095	version	0.03772
lunch	0.02710	data	0.03408	file	0.03702
manmatha	0.02391	calo	0.03236	files	0.02534
andrew	0.02332	message	0.03053	java	0.02522
faculty	0.01764	enron	0.03028	cvs	0.02511
april	0.01740	project	0.02415	directory	0.01978
shlomo	0.01657	send	0.02023	add	0.01932
al	0.01621	part	0.01680	checked	0.01481

Topic Distributions Conditioned on Time



Social Network Analysis with Links *and Text*

Role Discovery

Group Discovery

Trend Discovery

Community Discovery

Impact Measurement

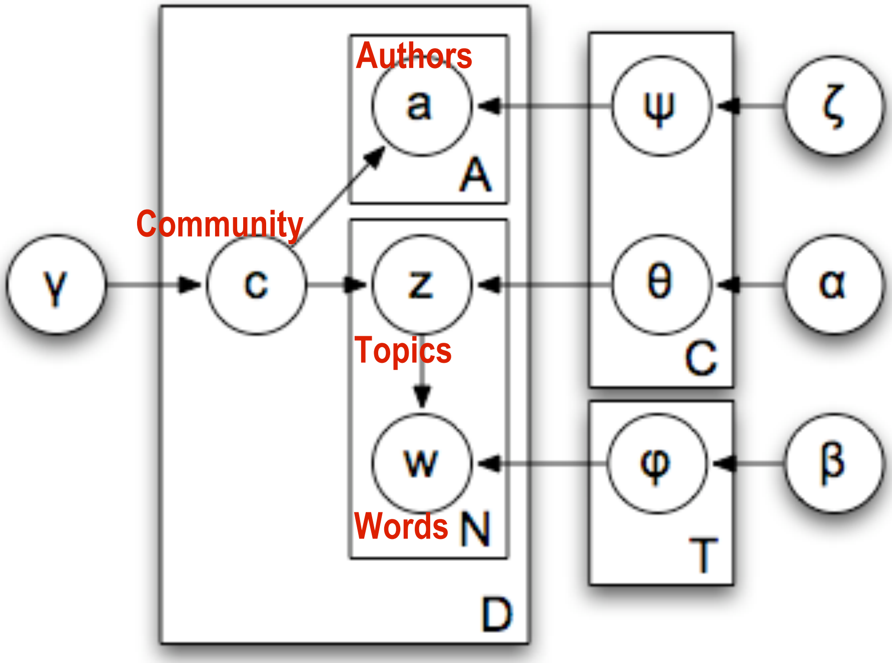
How do new links form in social networks?

1) Randomly *(Poisson graph)*
2) Pick someone popular *(Preferential attachment)*
3) Pick someone with mutual friends *(Adamic & Adar, Liben-Nowell & Kleinberg)*

4) Pick someone from one of your “communities” *(Mimno, Wallach & McCallum 2007)*

Can we find communities that help predict links?

A Community-based Generative Model for Text and Co-authorships

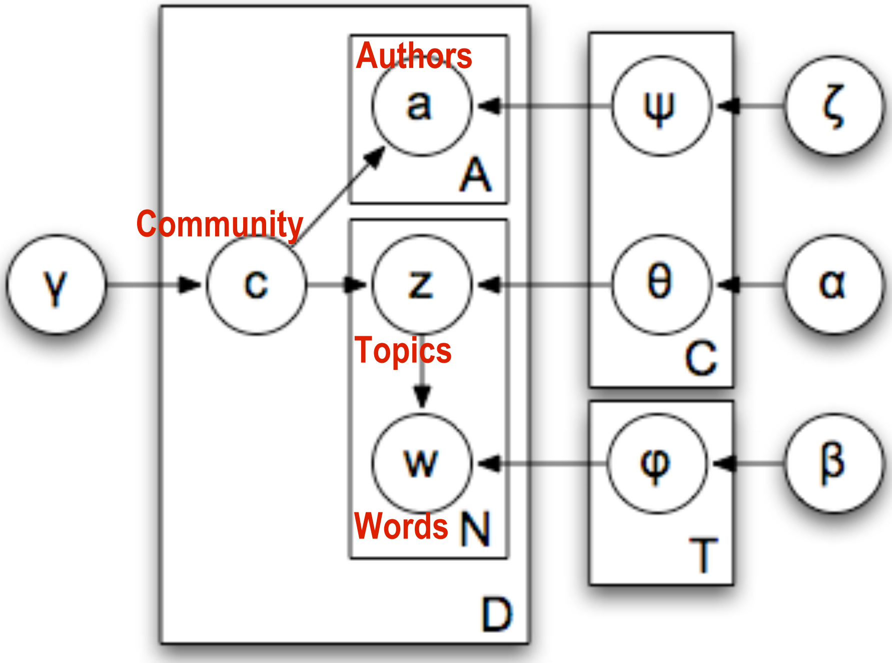


1) To generate a document, we first pick a community.

2) The community then determines the choice of authors and topics.

3) From topics, we pick words.

A Community-based Generative Model for Text and Co-authorships



Graphical Model can answer various queries!

$P(\text{author}_3 \mid \text{author}_1, \text{author}_2)$

$P(\text{author}_3 \mid \text{author}_1, \text{author}_2, \text{text})$

P(community | authors)

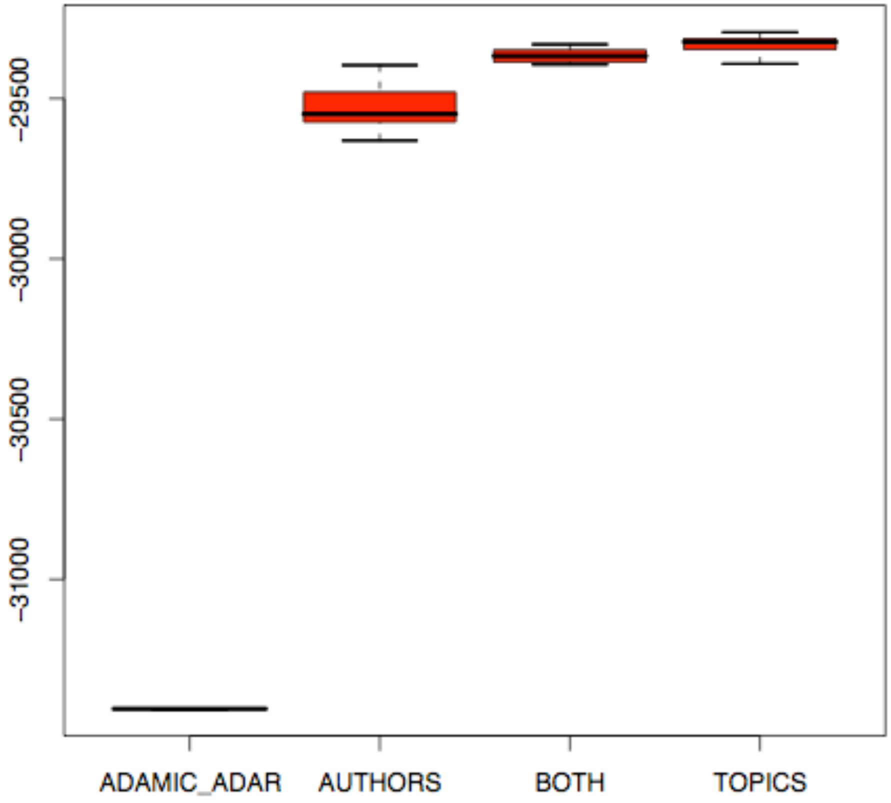
P (authors | community)

P (text | community)

P (text | authors)

Link Prediction

Probability of NIPS 2004-6 Co-authorships



(Preferential attachment is much worse, at -40,121.)

Community-Author View

Ng_A	features, feature, markov, sequence, models, conditional, label, function, set
Koller_D	number, results, paper, based, function, previous, resulting, introduction, general
Parr_R	policy, learning, action, states, function, reward, actions, optimal, mdp
Abbeel_P	control, controller, model, helicopter, system, neural, forward, learning, systems
Jordan_M	model, models, press, shows, figure, related, journal, underlying, correspond
Merzenich_M	present, effect, figure, references, important, increase, similar, addition, increased
Mel_B	learning, control, reinforcement, sutton, action, space, task, trajectory, methods

Jordan_M	propagation, belief, tree, nodes, node, approximation, variational, networks, boun
Jaakkola_T	number, results, paper, based, function, previous, resulting, introduction, general
Saul_L	theorem, case, proof, function, assume, set, section, algorithm, bound
Bach_F_R	field, boltzmann, approximations, exact, jordan, parameters, set, step, network
Singh_S	log, models, inference, variables, model, distribution, variational, parameters, matr
Wainwright_M	problem, algorithm, optimization, methods, solution, method, problems, proposed,
Nguyen_X	clustering, spectral, graph, matrix, cut, data, clusters, eigenvectors, normalized

Community-Author-Topic View

Griffiths_T_L	words, model, word, documents, document, text, topic, distribution, mixture
Singer_Y	suffix, algorithm, feature, adaptor, space, model, kernels, strings, natural
Blei_D	learning, category, naive, definition, estimation, single, figure, applied, obtain
Goldwater_S	set, labels, analysis, adclus, pmm, function, evaluation, problem, alphabet
Jordan_M	number, results, paper, based, function, previous, resulting, introduction, general
Johnson_M	prior, posterior, distribution, bayesian, likelihood, data, models, probability, model
Campbell_W	target, task, visual, figure, contrast, attention, search, orientation, discrimination

Jordan_M	propagation, belief, tree, nodes, node, approximation, variational, networks, bounc
Willsky_A	field, boltzmann, approximations, exact, jordan, parameters, set, step, network
Jaakkola_T	log, models, inference, variables, model, distribution, variational, parameters, matr
Saul_L	network, variables, node, inference, distribution, nodes, algorithm, message, tree
Wiegerinck_W	number, results, paper, based, function, previous, resulting, introduction, general
Kappen_H	theorem, case, proof, function, assume, set, section, algorithm, bound
Wainwright_M	mixture, data, gaussian, density, likelihood, parameters, distribution, model, functi

Kawato_M	control, motor, learning, arm, model, movement, feedback, movements, hand
Jordan_M	eye, vor, visual, desired, field, controller, force, cerebellum, vestibular
Barto_A	neural, data, activity, figure, firing, movement, motor, speech, dynamics
Vatikiotis	present, effect, figure, references, important, increase, similar, addition, increased

Social Network Analysis with Links *and Text*

Role Discovery

Group Discovery

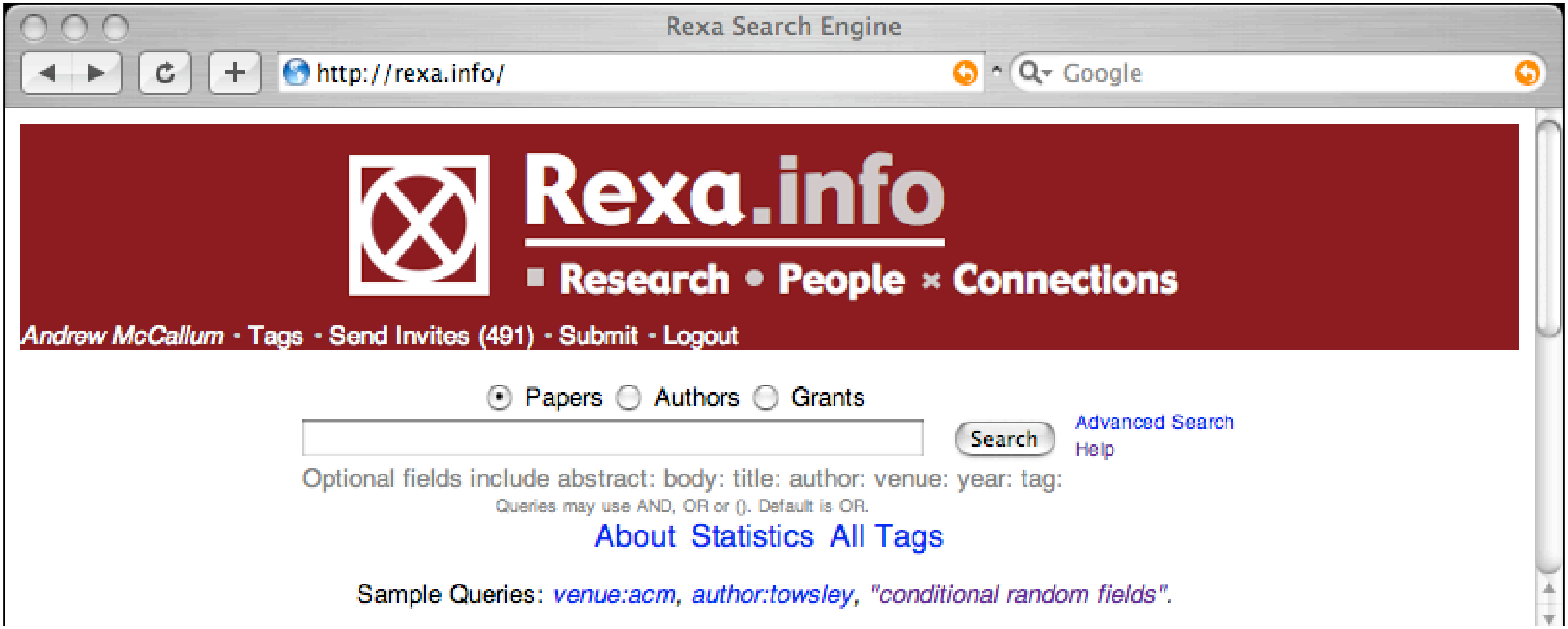
Trend Discovery

Community Discovery

Impact Measurement

Our Data

- Over 1.6 million research papers, gathered as part of *Rexa.info* portal.
- Cross linked references / citations.



Andrew McCallum • Tags • Send Invites (491) • Submit • Logout

◉ Papers ○ Authors ○ Grants

Search

Advanced Search

Help

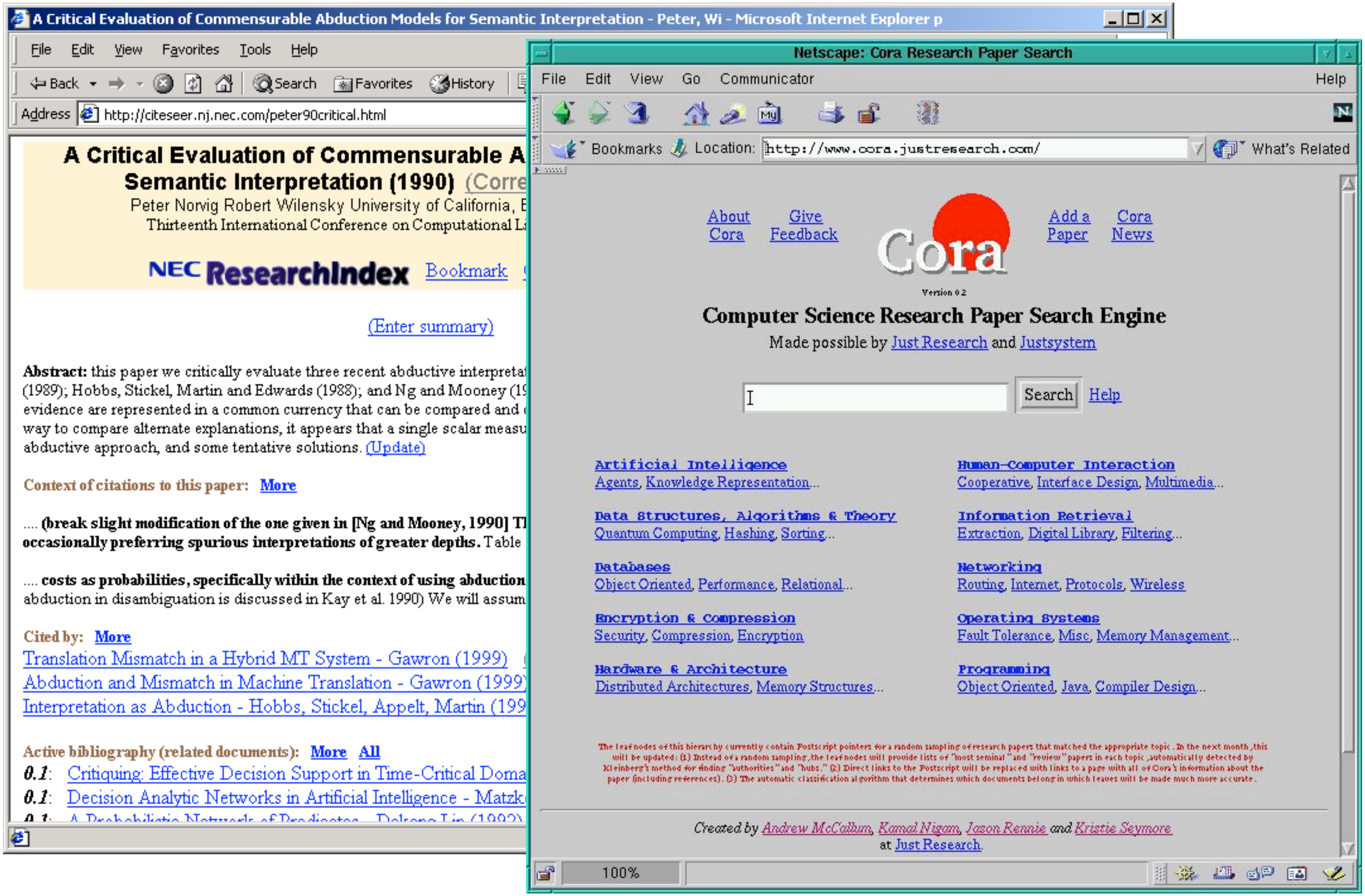
Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

About Statistics All Tags

Sample Queries: *venue:acm*, *author:towsley*, *"conditional random fields"*.

Previous Systems



A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://citeseer.nj.nec.com/peter90critical.html

A Critical Evaluation of Commensurable A
Semantic Interpretation (1990) (Corre

Peter Norvig Robert Wilensky University of California, E
Thirteenth International Conference on Computational Li

NEC ResearchIndex Bookmark

(Enter summary)

Abstract: this paper we critically evaluate three recent abductive interpretat (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (19 evidence are represented in a common currency that can be compared and way to compare alternate explanations, it appears that a single scalar measu abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: More

.... **(break slight modification of the one given in [Ng and Mooney, 1990] Tl occasionally preferring spurious interpretations of greater depths.** Table

.... **costs as probabilities, specifically within the context of using abduction** abduction in disambiguation is discussed in Kay et al. 1990) We will assum

Cited by: More

Translation Mismatch in a Hybrid MT System - Gawron (1999)
Abduction and Mismatch in Machine Translation - Gawron (1999)
Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin (199

Active bibliography (related documents): More All

0.1: Critiquing: Effective Decision Support in Time-Critical Doma
0.1: Decision Analytic Networks in Artificial Intelligence - Matzk
0.1: A Probabilistic Network of Predicates - Dekang Lin (1992)

Netscape: Cora Research Paper Search
File Edit View Go Communicator Help
Bookmarks Location: http://www.cora.justresearch.com/ What's Related
About Cora
Give Feedback
Cora
Version 0.2
Add a Paper
Cora News
Search Help
100%

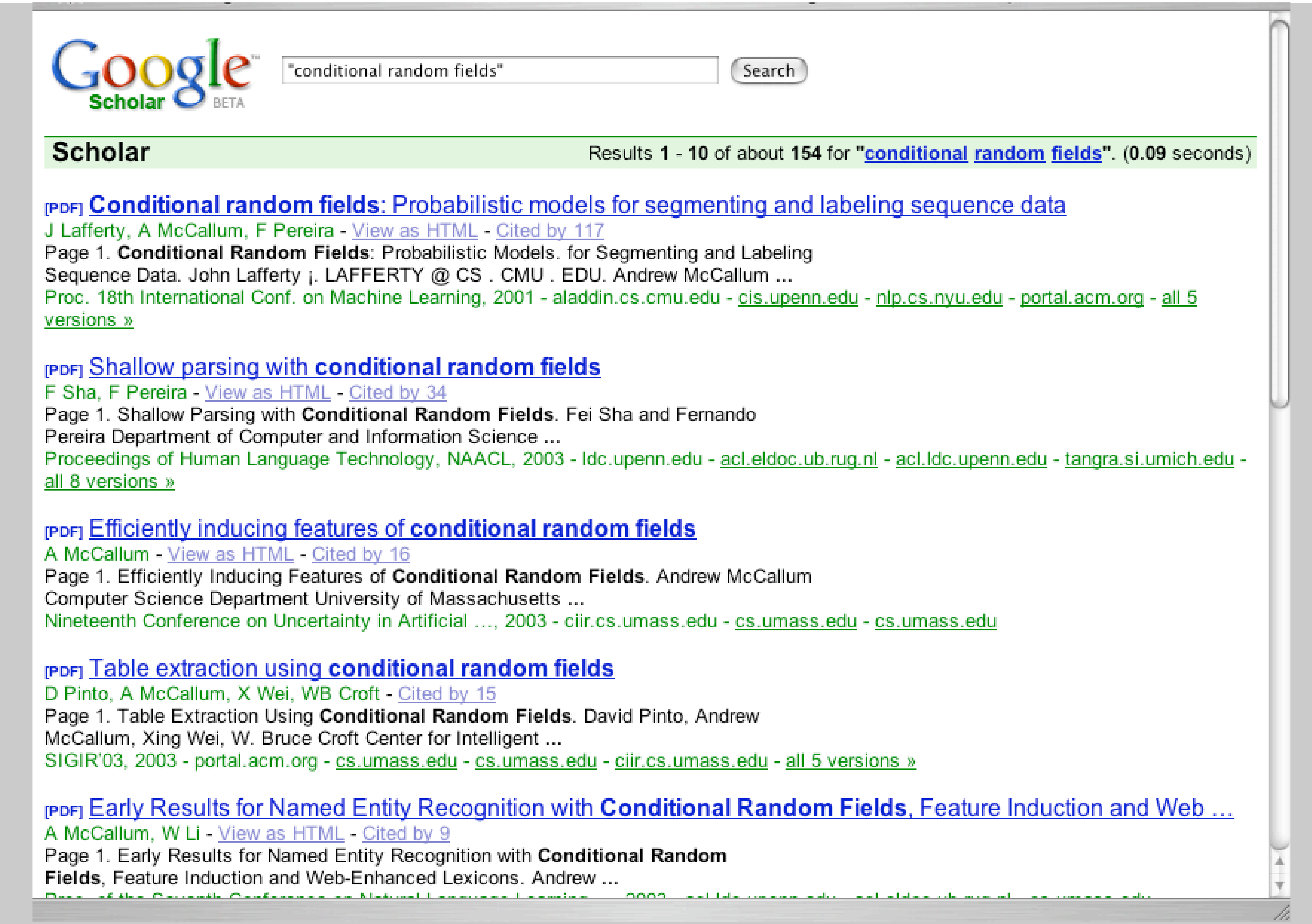
Computer Science Research Paper Search Engine

Made possible by Just Research and Justsystem

Artificial Intelligence Agents, Knowledge Representation...	**Human-Computer Interaction** Cooperative, Interface Design, Multimedia...
Data Structures, Algorithms & Theory Quantum Computing, Hashing, Sorting...	**Information Retrieval** Extraction, Digital Library, Filtering...
Databases Object Oriented, Performance, Relational...	**Networking** Routing, Internet, Protocols, Wireless
Encryption & Compression Security, Compression, Encryption	**Operating Systems** Fault Tolerance, Misc, Memory Management...
Hardware & Architecture Distributed Architectures, Memory Structures...	**Programming** Object Oriented, Java, Compiler Design...

The leaf nodes of this hierarchy currently contain Postscript pointers for a random sampling of research papers that matched the appropriate topic. In the next month, this will be updated: (1) Instead of a random sampling, the leaf nodes will provide lists of "most seminal" and "review" papers in each topic, automatically detected by Kleinberg's method for finding "authorities" and "hubs." (2) Direct links to the Postscript will be replaced with links to a page with all of Cora's information about the paper (including references). (3) The automatic classification algorithm that determines which documents belong in which leaves will be made much more accurate.

Created by Andrew McCallum, Kamal Nigam, Jason Rennie and Kristie Seymore
at Just Research.



Scholar Results **1** - **10** of about **154** for "**conditional random fields**". (**0.09** seconds)

[PDF] **Conditional random fields**: Probabilistic models for segmenting and labeling sequence data
J Lafferty, A McCallum, F Pereira - View as HTML - Cited by 117
Page 1. **Conditional Random Fields**: Probabilistic Models. for Segmenting and Labeling Sequence Data. John Lafferty ¡. LAFFERTY @ CS . CMU . EDU. Andrew McCallum **...**
Proc. 18th International Conf. on Machine Learning, 2001 - aladdin.cs.cmu.edu - cis.upenn.edu - nlp.cs.nyu.edu - portal.acm.org - all 5 versions »

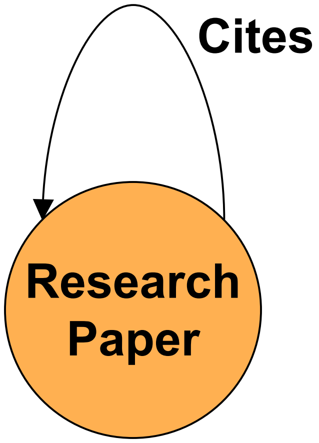
[PDF] Shallow parsing with **conditional random fields**
F Sha, F Pereira - View as HTML - Cited by 34
Page 1. Shallow Parsing with **Conditional Random Fields**. Fei Sha and Fernando Pereira Department of Computer and Information Science **...**
Proceedings of Human Language Technology, NAACL, 2003 - ldc.upenn.edu - acl.eldoc.ub.rug.nl - acl.ldc.upenn.edu - tangra.si.umich.edu - all 8 versions »

[PDF] Efficiently inducing features of **conditional random fields**
A McCallum - View as HTML - Cited by 16
Page 1. Efficiently Inducing Features of **Conditional Random Fields**. Andrew McCallum Computer Science Department University of Massachusetts **...**
Nineteenth Conference on Uncertainty in Artificial …, 2003 - ciir.cs.umass.edu - cs.umass.edu - cs.umass.edu

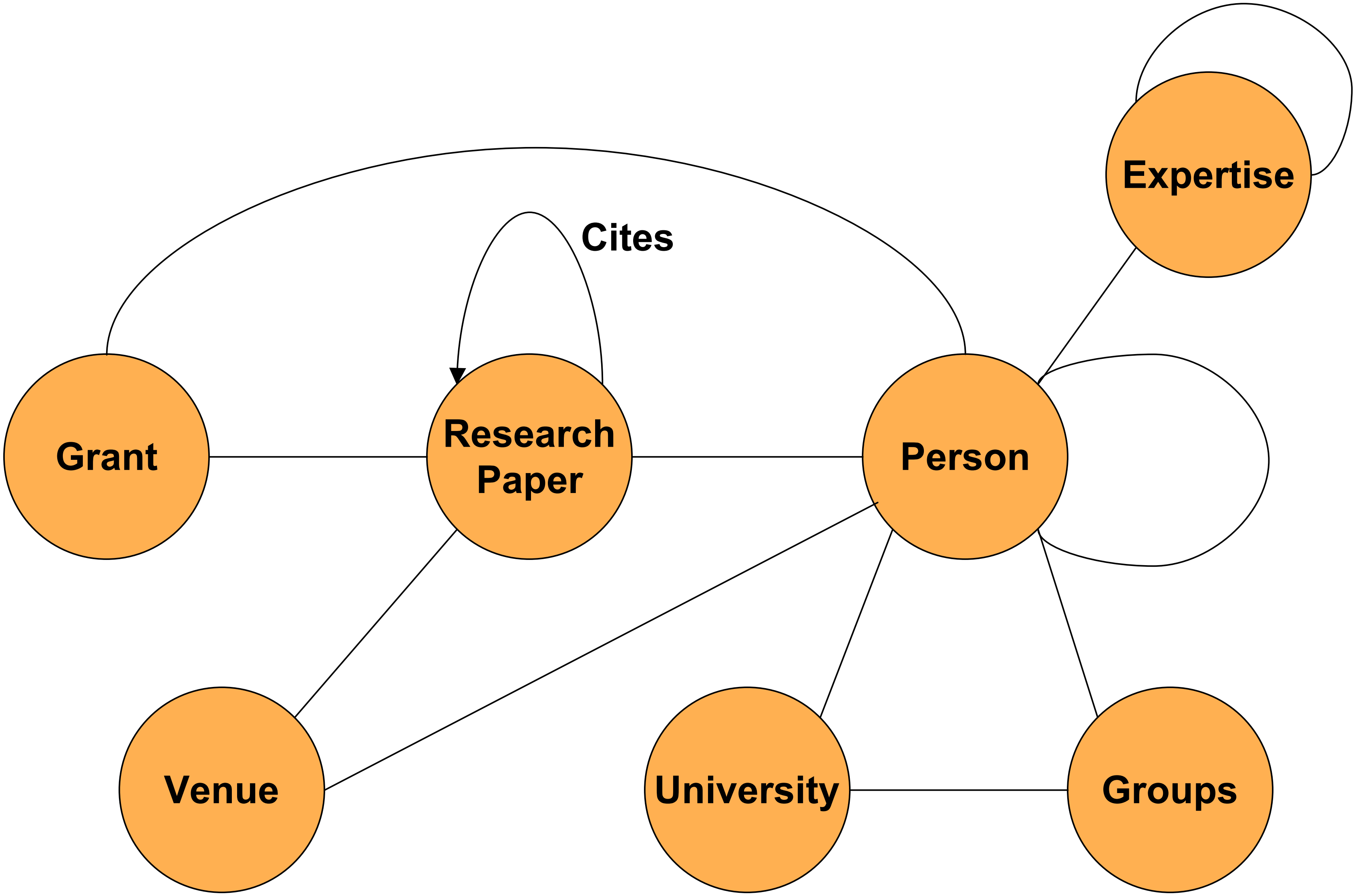
[PDF] Table extraction using **conditional random fields**
D Pinto, A McCallum, X Wei, WB Croft - Cited by 15
Page 1. Table Extraction Using **Conditional Random Fields**. David Pinto, Andrew McCallum, Xing Wei, W. Bruce Croft Center for Intelligent **...**
SIGIR'03, 2003 - portal.acm.org - cs.umass.edu - cs.umass.edu - ciir.cs.umass.edu - all 5 versions »

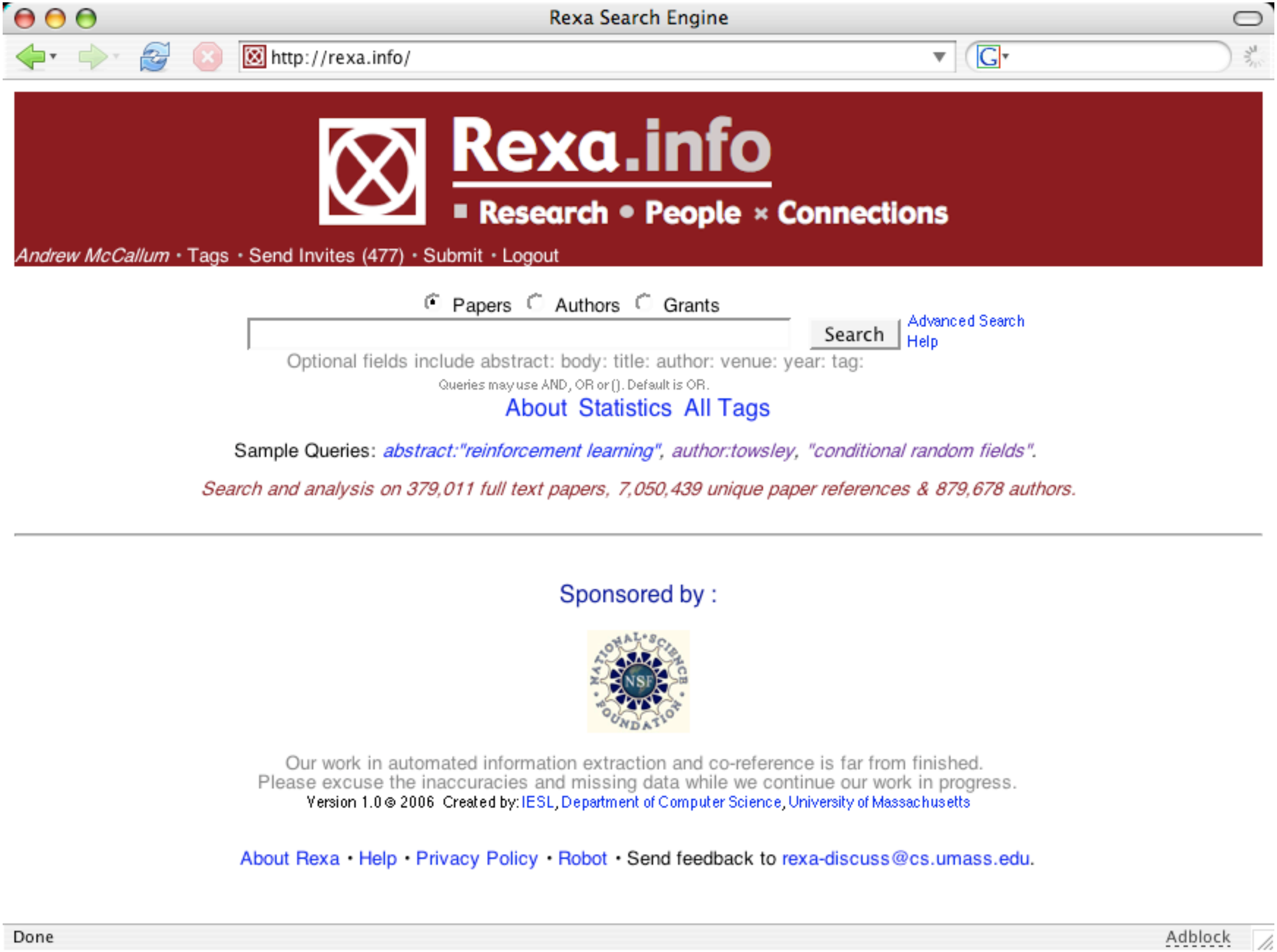
[PDF] Early Results for Named Entity Recognition with **Conditional Random Fields**, Feature Induction and Web …
A McCallum, W Li - View as HTML - Cited by 9
Page 1. Early Results for Named Entity Recognition with **Conditional Random Fields**, Feature Induction and Web-Enhanced Lexicons. Andrew **...**
Proc. of the Seventh Conference on Natural Language Learning …, 2003 - acl.ldc.upenn.edu - acl.eldoc.ub.rug.nl - cs.umass.edu

Previous Systems



More Entities and Relations





Andrew McCallum • Tags • Send Invites (477) • Submit • Logout

● Papers ○ Authors ○ Grants

Search

Advanced Search
Help

Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

About Statistics All Tags

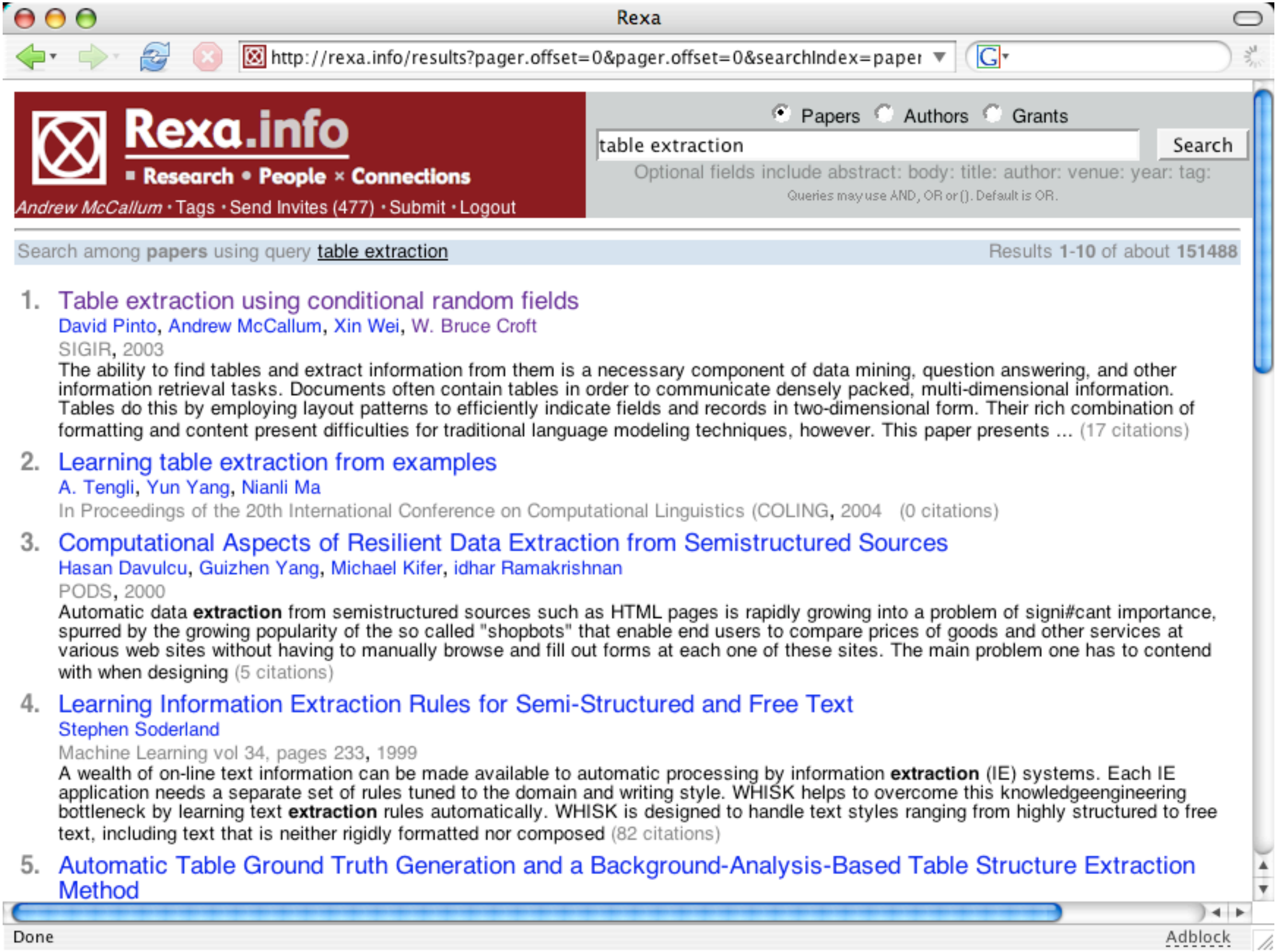
Sample Queries: *abstract:"reinforcement learning"*, *author:towsley*, *"conditional random fields"*.

Search and analysis on 379,011 full text papers, 7,050,439 unique paper references & 879,678 authors.

Sponsored by :

Our work in automated information extraction and co-reference is far from finished.
Please excuse the inaccuracies and missing data while we continue our work in progress.
Version 1.0 © 2006 Created by: IESL, Department of Computer Science, University of Massachusetts

About Rexa • Help • Privacy Policy • Robot • Send feedback to rexa-discuss@cs.umass.edu.


Rexa
http://rexa.info/results?pager.offset=0&pager.offset=0&searchIndex=paper
Rexa.info
Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
table extraction
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.
Done
Adblock

Search among **papers** using query table extraction Results **1-10** of about **151488**

1. Table extraction using conditional random fields
David Pinto, Andrew McCallum, Xin Wei, W. Bruce Croft
SIGIR, 2003
The ability to find tables and extract information from them is a necessary component of data mining, question answering, and other information retrieval tasks. Documents often contain tables in order to communicate densely packed, multi-dimensional information. Tables do this by employing layout patterns to efficiently indicate fields and records in two-dimensional form. Their rich combination of formatting and content present difficulties for traditional language modeling techniques, however. This paper presents ... (17 citations)

2. Learning table extraction from examples
A. Tengli, Yun Yang, Nianli Ma
In Proceedings of the 20th International Conference on Computational Linguistics (COLING, 2004 (0 citations)

3. Computational Aspects of Resilient Data Extraction from Semistructured Sources
Hasan Davulcu, Guizhen Yang, Michael Kifer, idhar Ramakrishnan
PODS, 2000
Automatic data **extraction** from semistructured sources such as HTML pages is rapidly growing into a problem of signi#cant importance, spurred by the growing popularity of the so called "shopbots" that enable end users to compare prices of goods and other services at various web sites without having to manually browse and fill out forms at each one of these sites. The main problem one has to contend with when designing (5 citations)

4. Learning Information Extraction Rules for Semi-Structured and Free Text
Stephen Soderland
Machine Learning vol 34, pages 233, 1999
A wealth of on-line text information can be made available to automatic processing by information **extraction** (IE) systems. Each IE application needs a separate set of rules tuned to the domain and writing style. WHISK helps to overcome this knowledgeengineering bottleneck by learning text **extraction** rules automatically. WHISK is designed to handle text styles ranging from highly structured to free text, including text that is neither rigidly formatted nor composed (82 citations)

5. Automatic Table Ground Truth Generation and a Background-Analysis-Based Table Structure Extraction Method

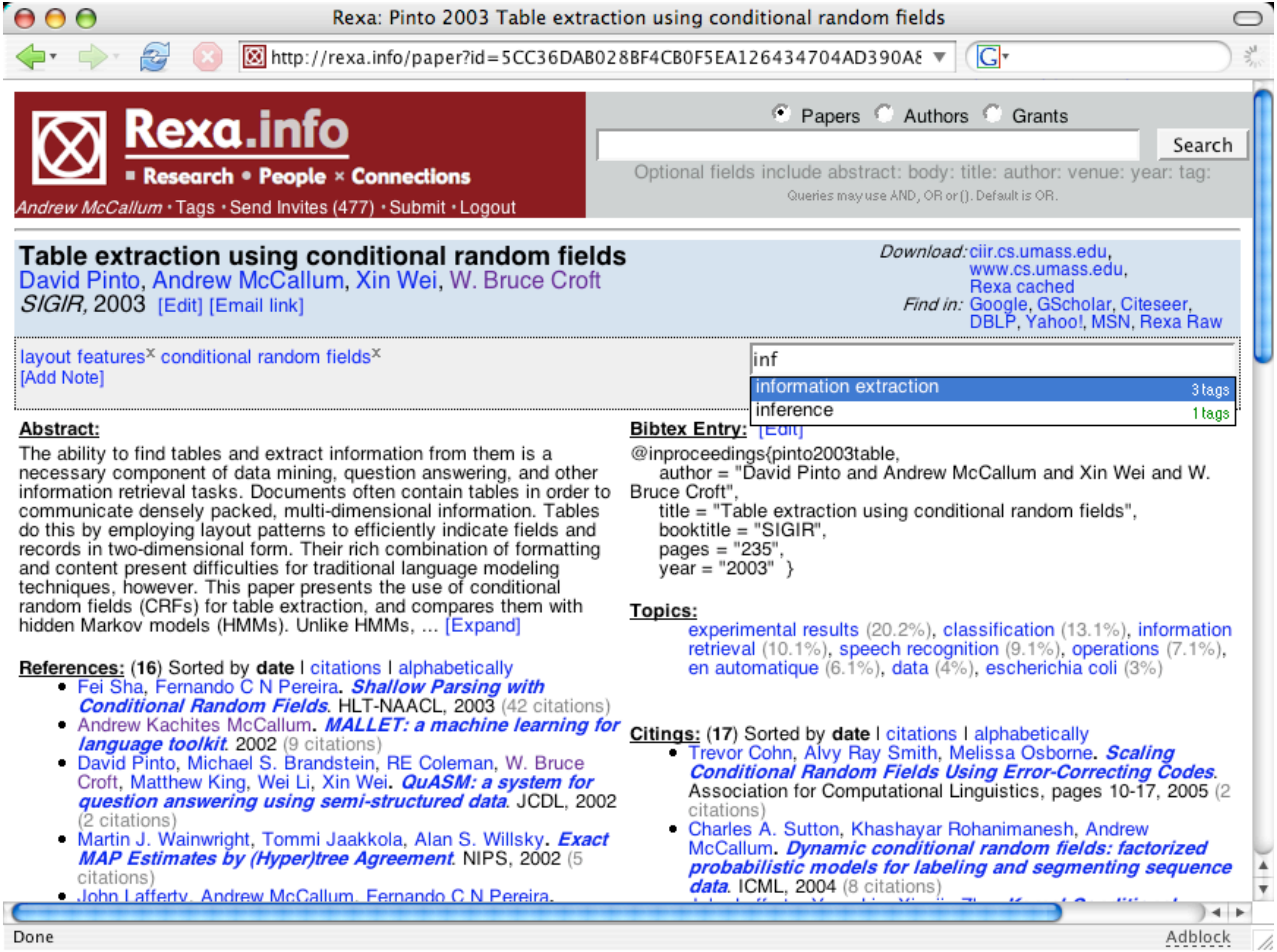

Rexa: Pinto 2003 Table extraction using conditional random fields
http://rexa.info/paper?id=5CC36DAB028BF4CB0F5EA126434704AD390A8
Rexa.info
■ Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.
Done
Adblock

Table extraction using conditional random fields

David Pinto, Andrew McCallum, Xin Wei, W. Bruce Croft

SIGIR, 2003 [Edit] [Email link]

Download: ciir.cs.umass.edu, www.cs.umass.edu, Rexa cached

Find in: Google, GScholar, Citeseer, DBLP, Yahoo!, MSN, Rexa Raw

layout features× conditional random fields×

[Add Note]

inf

information extraction	3 tags
inference	1 tags

Abstract:

The ability to find tables and extract information from them is a necessary component of data mining, question answering, and other information retrieval tasks. Documents often contain tables in order to communicate densely packed, multi-dimensional information. Tables do this by employing layout patterns to efficiently indicate fields and records in two-dimensional form. Their rich combination of formatting and content present difficulties for traditional language modeling techniques, however. This paper presents the use of conditional random fields (CRFs) for table extraction, and compares them with hidden Markov models (HMMs). Unlike HMMs, ... [Expand]

References: (**16**) Sorted by **date** | citations | alphabetically

- Fei Sha, Fernando C N Pereira. ***Shallow Parsing with Conditional Random Fields***. HLT-NAACL, 2003 (42 citations)
- Andrew Kachites McCallum. ***MALLET: a machine learning for language toolkit***. 2002 (9 citations)
- David Pinto, Michael S. Brandstein, RE Coleman, W. Bruce Croft, Matthew King, Wei Li, Xin Wei. ***QuASM: a system for question answering using semi-structured data***. JCDL, 2002 (2 citations)
- Martin J. Wainwright, Tommi Jaakkola, Alan S. Willsky. ***Exact MAP Estimates by (Hyper)tree Agreement***. NIPS, 2002 (5 citations)
- John Lafferty, Andrew McCallum, Fernando C N Pereira.

Bibtex Entry: [Edit]

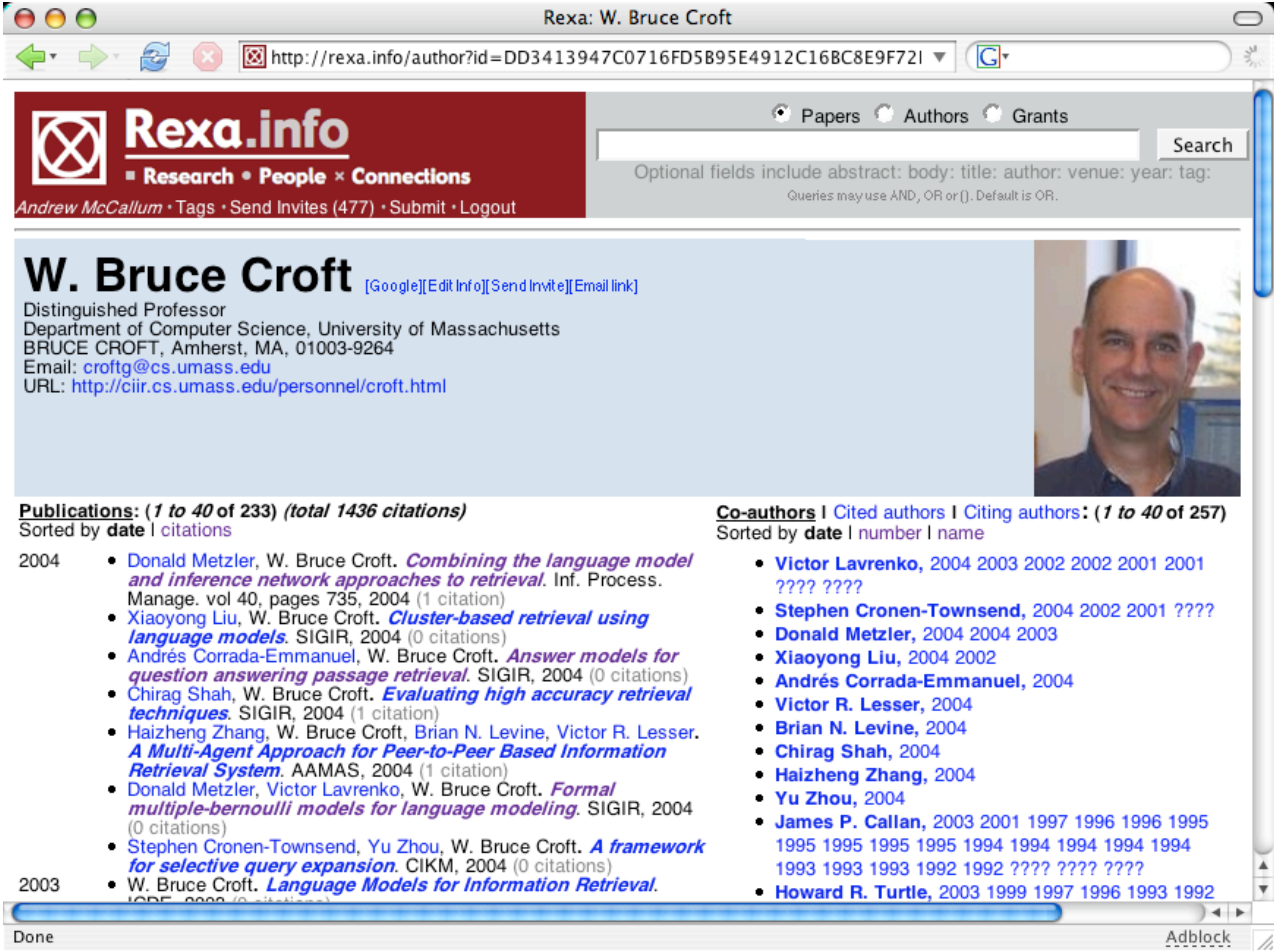
```
@inproceedings{pinto2003table,
    author = "David Pinto and Andrew McCallum and Xin Wei and W.
Bruce Croft",
    title = "Table extraction using conditional random fields",
    booktitle = "SIGIR",
    pages = "235",
    year = "2003"  }
```

Topics:

experimental results (20.2%), classification (13.1%), information retrieval (10.1%), speech recognition (9.1%), operations (7.1%), en automatique (6.1%), data (4%), escherichia coli (3%)

Citings: (**17**) Sorted by **date** | citations | alphabetically

- Trevor Cohn, Alvy Ray Smith, Melissa Osborne. ***Scaling Conditional Random Fields Using Error-Correcting Codes***. Association for Computational Linguistics, pages 10-17, 2005 (2 citations)
- Charles A. Sutton, Khashayar Rohanimanesh, Andrew McCallum. ***Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data***. ICML, 2004 (8 citations)


Rexa: W. Bruce Croft
http://rexa.info/author?id=DD3413947C0716FD5B95E4912C16BC8E9F72I
Rexa.info
▪ Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.
[Google][Edit Info][Send Invite][Email link]
Done
Adblock

W. Bruce Croft

Distinguished Professor
Department of Computer Science, University of Massachusetts
BRUCE CROFT, Amherst, MA, 01003-9264
Email: croftg@cs.umass.edu
URL: http://ciir.cs.umass.edu/personnel/croft.html

Publications: (*1 to 40* of 233) *(total 1436 citations)*
Sorted by **date** | citations

2004

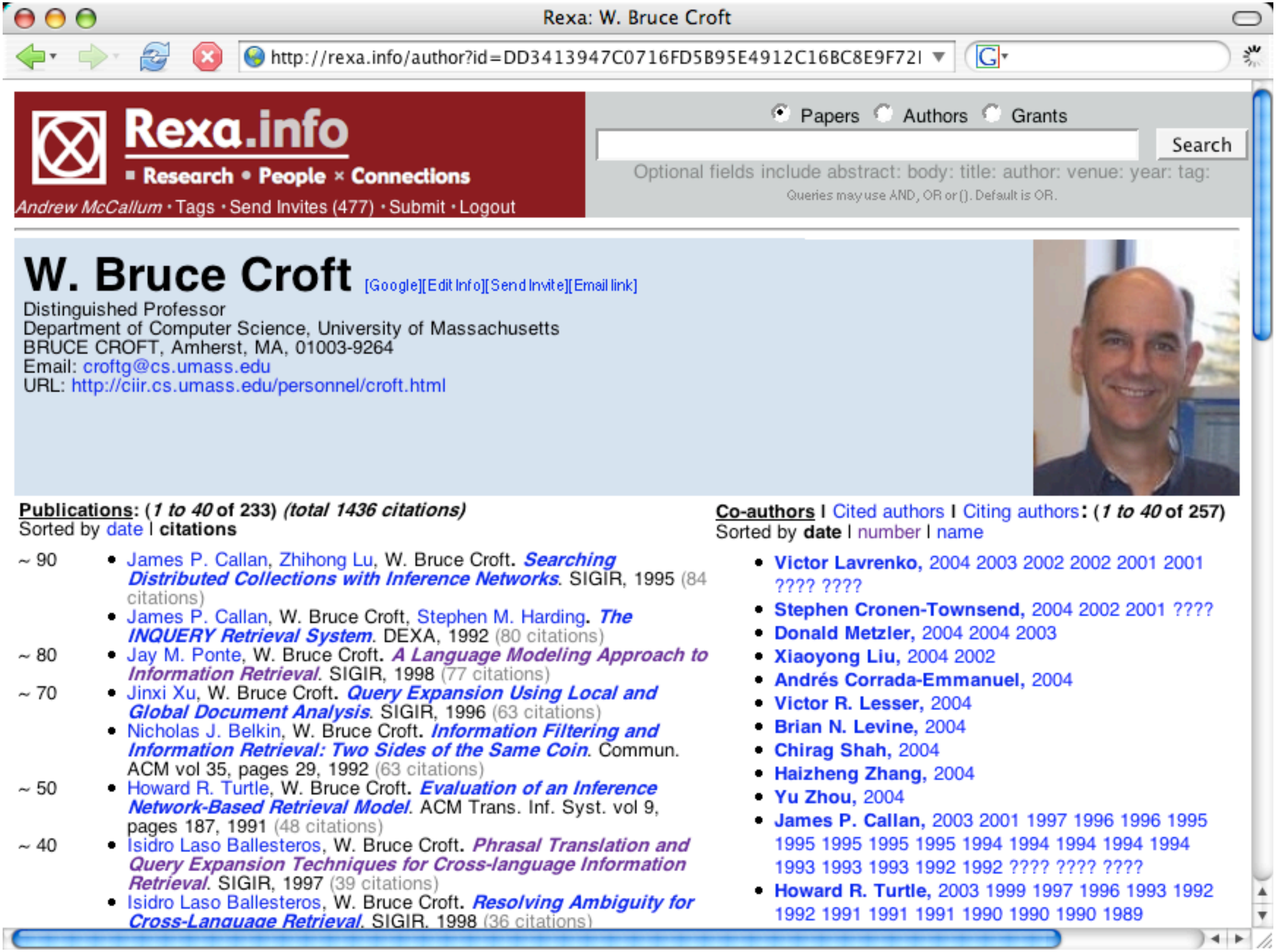
- Donald Metzler, W. Bruce Croft. ***Combining the language model and inference network approaches to retrieval***. Inf. Process. Manage. vol 40, pages 735, 2004 (1 citation)
- Xiaoyong Liu, W. Bruce Croft. ***Cluster-based retrieval using language models***. SIGIR, 2004 (0 citations)
- Andrés Corrada-Emmanuel, W. Bruce Croft. ***Answer models for question answering passage retrieval***. SIGIR, 2004 (0 citations)
- Chirag Shah, W. Bruce Croft. ***Evaluating high accuracy retrieval techniques***. SIGIR, 2004 (1 citation)
- Haizheng Zhang, W. Bruce Croft, Brian N. Levine, Victor R. Lesser. ***A Multi-Agent Approach for Peer-to-Peer Based Information Retrieval System***. AAMAS, 2004 (1 citation)
- Donald Metzler, Victor Lavrenko, W. Bruce Croft. ***Formal multiple-bernoulli models for language modeling***. SIGIR, 2004 (0 citations)
- Stephen Cronen-Townsend, Yu Zhou, W. Bruce Croft. ***A framework for selective query expansion***. CIKM, 2004 (0 citations)

2003

- W. Bruce Croft. ***Language Models for Information Retrieval***.

Co-authors | Cited authors | Citing authors: (*1 to 40* of 257)
Sorted by **date** | number | name

- **Victor Lavrenko,** 2004 2003 2002 2002 2001 2001 ???? ????
- **Stephen Cronen-Townsend,** 2004 2002 2001 ????
- **Donald Metzler,** 2004 2004 2003
- **Xiaoyong Liu,** 2004 2002
- **Andrés Corrada-Emmanuel,** 2004
- **Victor R. Lesser,** 2004
- **Brian N. Levine,** 2004
- **Chirag Shah,** 2004
- **Haizheng Zhang,** 2004
- **Yu Zhou,** 2004
- **James P. Callan,** 2003 2001 1997 1996 1996 1995 1995 1995 1995 1995 1994 1994 1994 1994 1994 1993 1993 1993 1992 1992 ???? ???? ????
- **Howard R. Turtle,** 2003 1999 1997 1996 1993 1992



Rexa: W. Bruce Croft
http://rexa.info/author?id=DD3413947C0716FD5B95E4912C16BC8E9F72I
Rexa.info
▪ Research • People × Connections

Andrew McCallum • Tags • Send Invites (477) • Submit • Logout

● Papers ○ Authors ○ Grants

Search

Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

W. Bruce Croft

[Google][Edit Info][Send Invite][Email link]

Distinguished Professor
Department of Computer Science, University of Massachusetts
BRUCE CROFT, Amherst, MA, 01003-9264
Email: croftg@cs.umass.edu
URL: http://ciir.cs.umass.edu/personnel/croft.html

Publications: (*1 to 40* of 233) *(total 1436 citations)*
Sorted by date | **citations**

~ 90
- James P. Callan, Zhihong Lu, W. Bruce Croft. ***Searching Distributed Collections with Inference Networks***. SIGIR, 1995 (84 citations)
- James P. Callan, W. Bruce Croft, Stephen M. Harding. ***The INQUERY Retrieval System***. DEXA, 1992 (80 citations)

~ 80
- Jay M. Ponte, W. Bruce Croft. ***A Language Modeling Approach to Information Retrieval***. SIGIR, 1998 (77 citations)

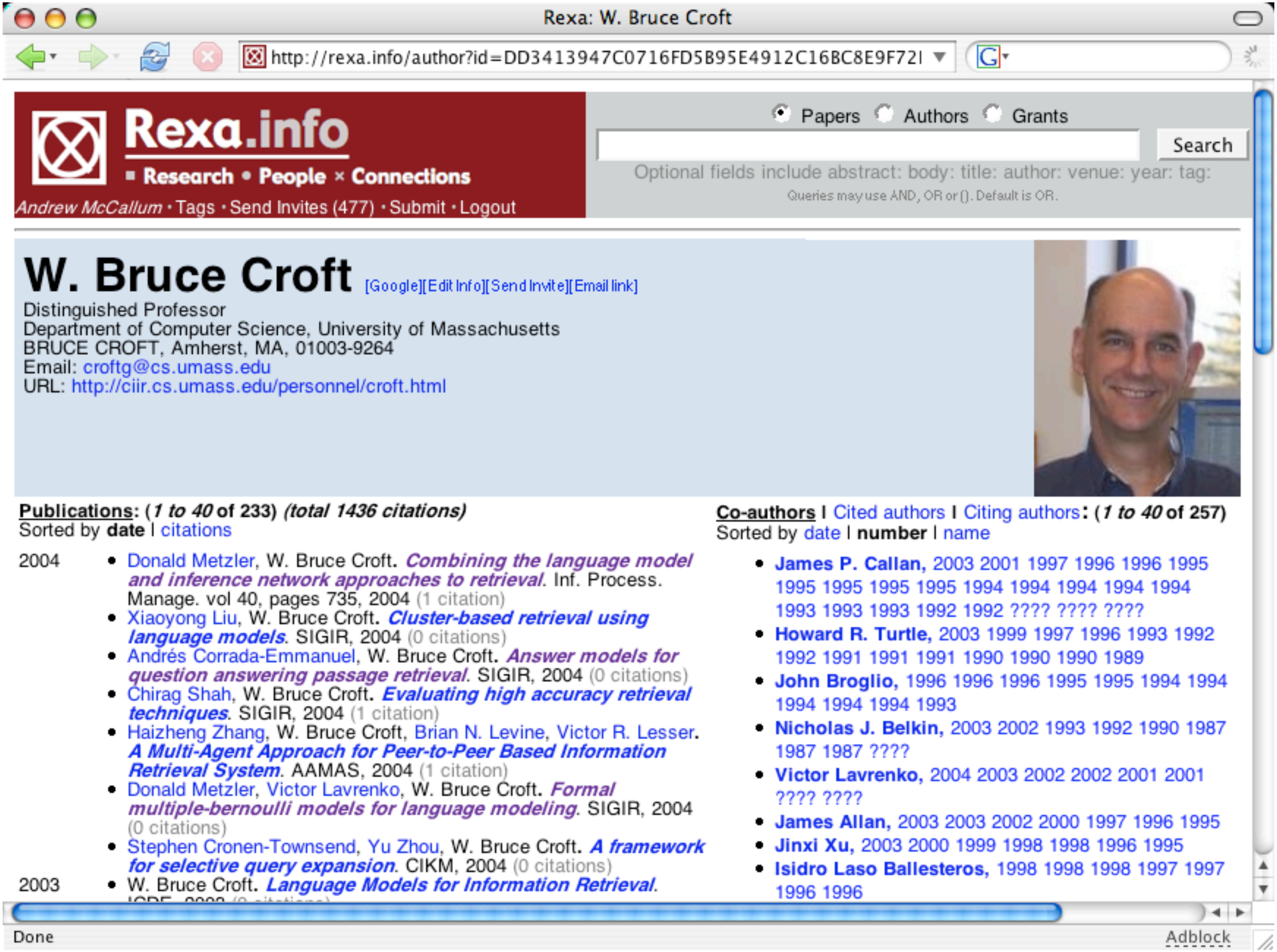
~ 70
- Jinxi Xu, W. Bruce Croft. ***Query Expansion Using Local and Global Document Analysis***. SIGIR, 1996 (63 citations)
- Nicholas J. Belkin, W. Bruce Croft. ***Information Filtering and Information Retrieval: Two Sides of the Same Coin***. Commun. ACM vol 35, pages 29, 1992 (63 citations)

~ 50
- Howard R. Turtle, W. Bruce Croft. ***Evaluation of an Inference Network-Based Retrieval Model***. ACM Trans. Inf. Syst. vol 9, pages 187, 1991 (48 citations)

~ 40
- Isidro Laso Ballesteros, W. Bruce Croft. ***Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval***. SIGIR, 1997 (39 citations)
- Isidro Laso Ballesteros, W. Bruce Croft. ***Resolving Ambiguity for Cross-Language Retrieval***. SIGIR. 1998 (36 citations)

Co-authors | Cited authors | Citing authors: (*1 to 40* of 257)
Sorted by **date** | number | name

- **Victor Lavrenko,** 2004 2003 2002 2002 2001 2001 ???? ????
- **Stephen Cronen-Townsend,** 2004 2002 2001 ????
- **Donald Metzler,** 2004 2004 2003
- **Xiaoyong Liu,** 2004 2002
- **Andrés Corrada-Emmanuel,** 2004
- **Victor R. Lesser,** 2004
- **Brian N. Levine,** 2004
- **Chirag Shah,** 2004
- **Haizheng Zhang,** 2004
- **Yu Zhou,** 2004
- **James P. Callan,** 2003 2001 1997 1996 1996 1995 1995 1995 1995 1995 1994 1994 1994 1994 1994 1993 1993 1993 1992 1992 ???? ???? ????
- **Howard R. Turtle,** 2003 1999 1997 1996 1993 1992 1992 1991 1991 1991 1990 1990 1990 1989



Rexa: W. Bruce Croft
http://rexa.info/author?id=DD3413947C0716FD5B95E4912C16BC8E9F72I
Rexa.info
■ Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.
Done
Adblock

W. Bruce Croft [Google][Edit Info][Send Invite][Email link]

Distinguished Professor
Department of Computer Science, University of Massachusetts
BRUCE CROFT, Amherst, MA, 01003-9264
Email: croftg@cs.umass.edu
URL: http://ciir.cs.umass.edu/personnel/croft.html

Publications: (*1 to 40* of 233) *(total 1436 citations)*
Sorted by **date** | citations

2004

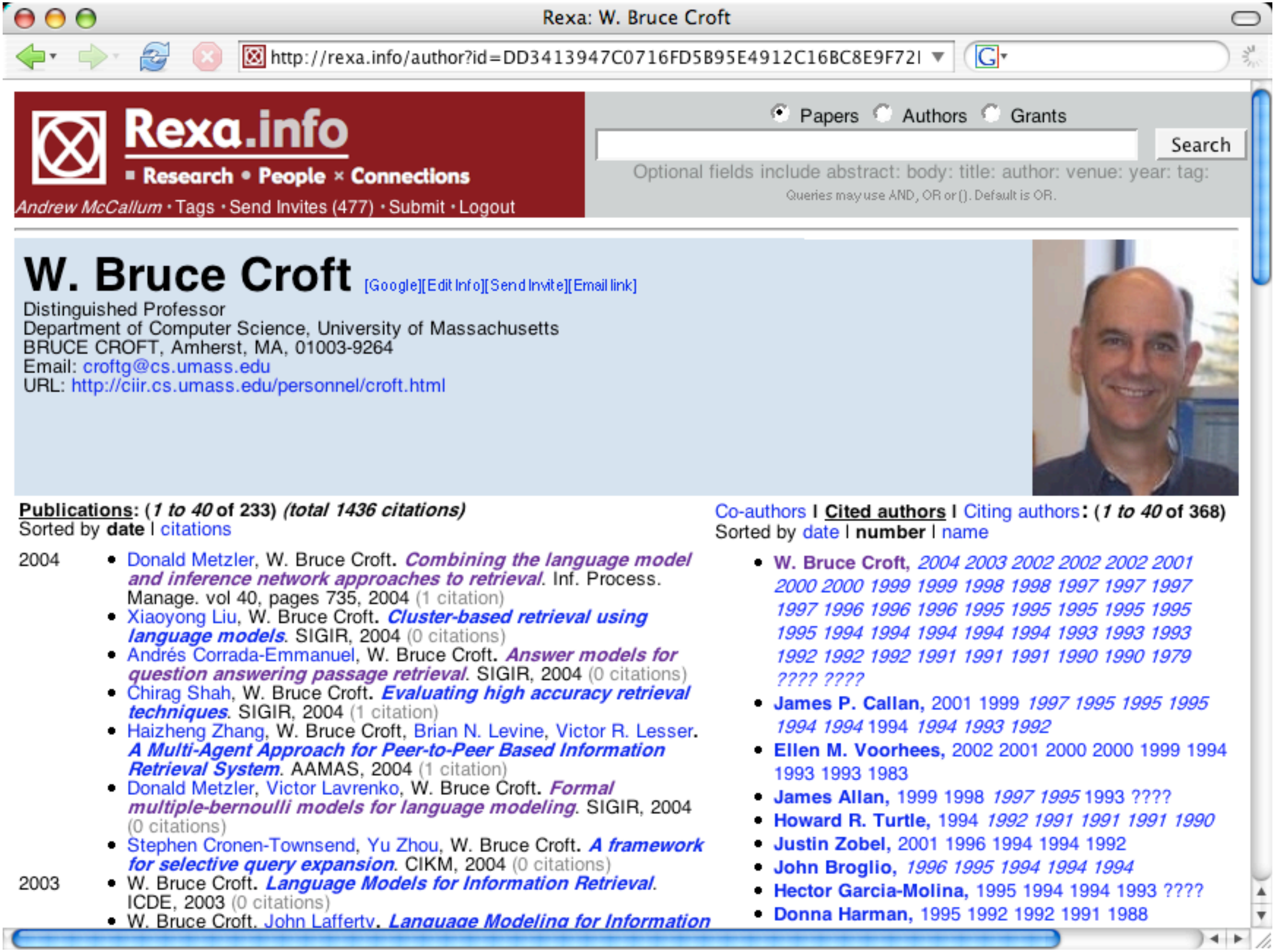
- Donald Metzler, W. Bruce Croft. ***Combining the language model and inference network approaches to retrieval***. Inf. Process. Manage. vol 40, pages 735, 2004 (1 citation)
- Xiaoyong Liu, W. Bruce Croft. ***Cluster-based retrieval using language models***. SIGIR, 2004 (0 citations)
- Andrés Corrada-Emmanuel, W. Bruce Croft. ***Answer models for question answering passage retrieval***. SIGIR, 2004 (0 citations)
- Chirag Shah, W. Bruce Croft. ***Evaluating high accuracy retrieval techniques***. SIGIR, 2004 (1 citation)
- Haizheng Zhang, W. Bruce Croft, Brian N. Levine, Victor R. Lesser. ***A Multi-Agent Approach for Peer-to-Peer Based Information Retrieval System***. AAMAS, 2004 (1 citation)
- Donald Metzler, Victor Lavrenko, W. Bruce Croft. ***Formal multiple-bernoulli models for language modeling***. SIGIR, 2004 (0 citations)
- Stephen Cronen-Townsend, Yu Zhou, W. Bruce Croft. ***A framework for selective query expansion***. CIKM, 2004 (0 citations)

2003

- W. Bruce Croft. ***Language Models for Information Retrieval***. ICDE, 2003 (0 citations)

Co-authors | Cited authors | Citing authors: (*1 to 40* of 257)
Sorted by date | **number** | name

- **James P. Callan,** 2003 2001 1997 1996 1996 1995 1995 1995 1995 1995 1994 1994 1994 1994 1994 1993 1993 1993 1992 1992 ???? ???? ????
- **Howard R. Turtle,** 2003 1999 1997 1996 1993 1992 1992 1991 1991 1991 1990 1990 1990 1989
- **John Broglio,** 1996 1996 1996 1995 1995 1994 1994 1994 1994 1994 1993
- **Nicholas J. Belkin,** 2003 2002 1993 1992 1990 1987 1987 1987 ????
- **Victor Lavrenko,** 2004 2003 2002 2002 2001 2001 ???? ????
- **James Allan,** 2003 2003 2002 2000 1997 1996 1995
- **Jinxi Xu,** 2003 2000 1999 1998 1998 1996 1995
- **Isidro Laso Ballesteros,** 1998 1998 1998 1997 1997 1996 1996



Rexa: W. Bruce Croft
http://rexa.info/author?id=DD3413947C0716FD5B95E4912C16BC8E9F72I
Rexa.info
▪ Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.

W. Bruce Croft [Google][Edit Info][Send Invite][Email link]

Distinguished Professor
Department of Computer Science, University of Massachusetts
BRUCE CROFT, Amherst, MA, 01003-9264
Email: croftg@cs.umass.edu
URL: http://ciir.cs.umass.edu/personnel/croft.html

Publications: (*1 to 40* of 233) *(total 1436 citations)*
Sorted by **date** | citations

2004

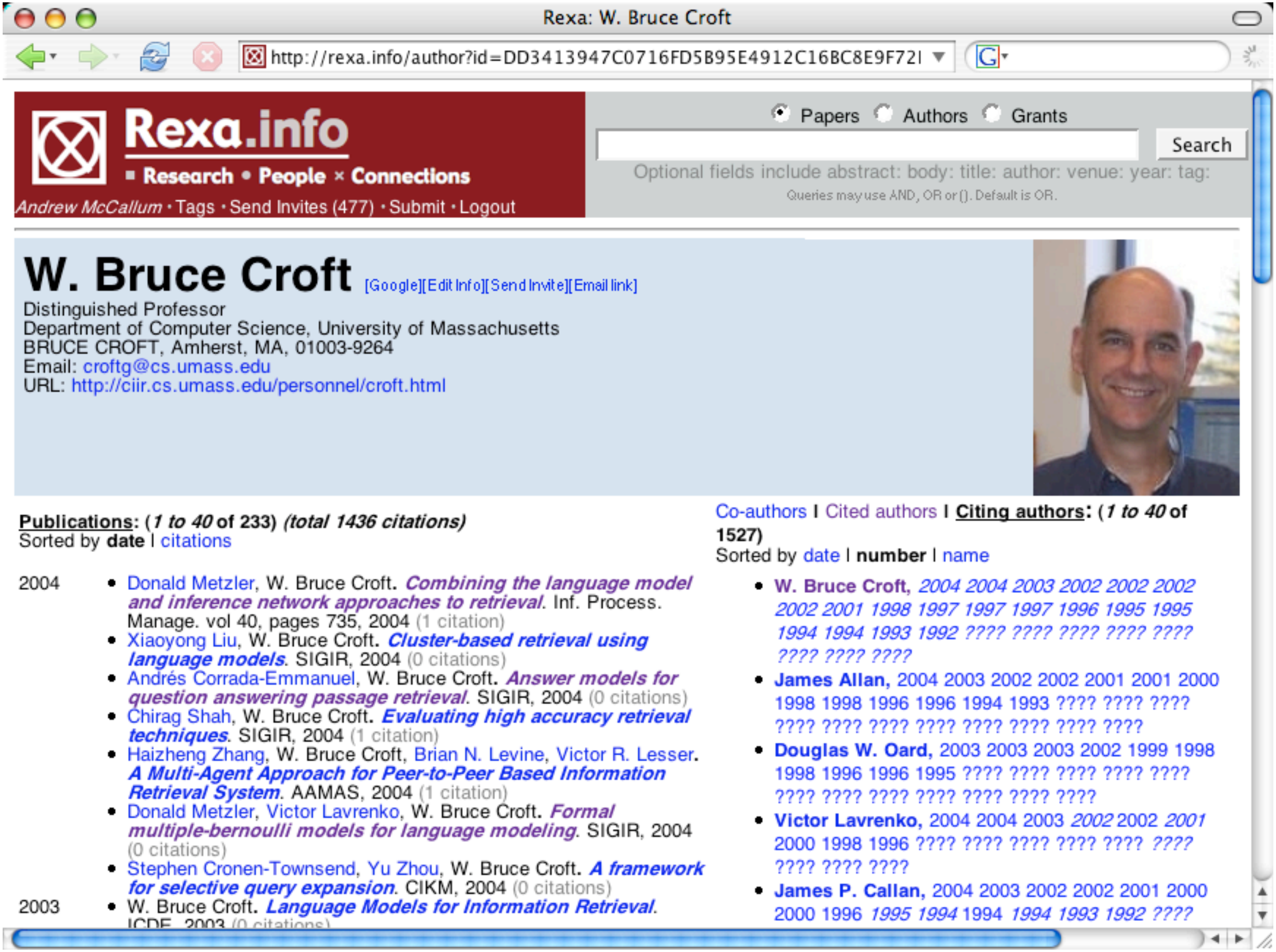
- Donald Metzler, W. Bruce Croft. ***Combining the language model and inference network approaches to retrieval***. Inf. Process. Manage. vol 40, pages 735, 2004 (1 citation)
- Xiaoyong Liu, W. Bruce Croft. ***Cluster-based retrieval using language models***. SIGIR, 2004 (0 citations)
- Andrés Corrada-Emmanuel, W. Bruce Croft. ***Answer models for question answering passage retrieval***. SIGIR, 2004 (0 citations)
- Chirag Shah, W. Bruce Croft. ***Evaluating high accuracy retrieval techniques***. SIGIR, 2004 (1 citation)
- Haizheng Zhang, W. Bruce Croft, Brian N. Levine, Victor R. Lesser. ***A Multi-Agent Approach for Peer-to-Peer Based Information Retrieval System***. AAMAS, 2004 (1 citation)
- Donald Metzler, Victor Lavrenko, W. Bruce Croft. ***Formal multiple-bernoulli models for language modeling***. SIGIR, 2004 (0 citations)
- Stephen Cronen-Townsend, Yu Zhou, W. Bruce Croft. ***A framework for selective query expansion***. CIKM, 2004 (0 citations)

2003

- W. Bruce Croft. ***Language Models for Information Retrieval***. ICDE, 2003 (0 citations)
- W. Bruce Croft, John Lafferty. ***Language Modeling for Information***

Co-authors | **Cited authors** | Citing authors: (*1 to 40* of 368)
Sorted by date | **number** | name

- **W. Bruce Croft,** *2004 2003 2002 2002 2002 2001 2000 2000 1999 1999 1998 1998 1997 1997 1997 1997 1996 1996 1996 1995 1995 1995 1995 1995 1995 1994 1994 1994 1994 1994 1993 1993 1993 1992 1992 1992 1991 1991 1991 1990 1990 1979 ???? ????*
- **James P. Callan,** 2001 1999 *1997 1995 1995 1995 1994 1994* 1994 *1994 1993 1992*
- **Ellen M. Voorhees,** 2002 2001 2000 2000 1999 1994 1993 1993 1983
- **James Allan,** 1999 1998 *1997 1995* 1993 ????
- **Howard R. Turtle,** 1994 *1992 1991 1991 1991 1990*
- **Justin Zobel,** 2001 1996 1994 1994 1992
- **John Broglio,** *1996 1995 1994 1994 1994*
- **Hector Garcia-Molina,** 1995 1994 1994 1993 ????
- **Donna Harman,** 1995 1992 1992 1991 1988


Rexa: W. Bruce Croft
http://rexa.info/author?id=DD3413947C0716FD5B95E4912C16BC8E9F72I
Rexa.info
▪ Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.

W. Bruce Croft [Google][Edit Info][Send Invite][Email link]

Distinguished Professor
Department of Computer Science, University of Massachusetts
BRUCE CROFT, Amherst, MA, 01003-9264
Email: croftg@cs.umass.edu
URL: http://ciir.cs.umass.edu/personnel/croft.html

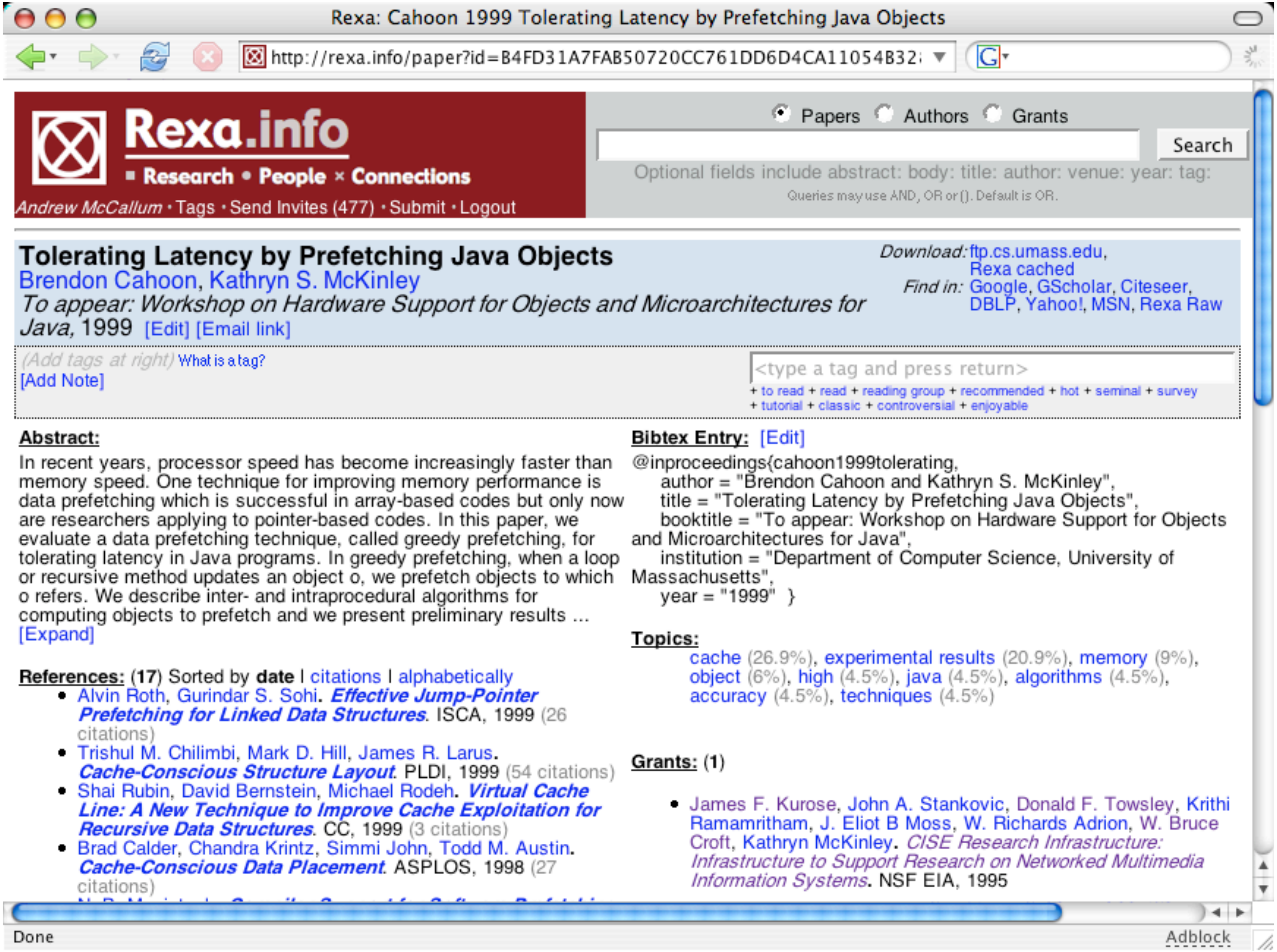
Publications: (*1 to 40* of 233) *(total 1436 citations)*
Sorted by **date** | citations

2004
- Donald Metzler, W. Bruce Croft. ***Combining the language model and inference network approaches to retrieval***. Inf. Process. Manage. vol 40, pages 735, 2004 (1 citation)
- Xiaoyong Liu, W. Bruce Croft. ***Cluster-based retrieval using language models***. SIGIR, 2004 (0 citations)
- Andrés Corrada-Emmanuel, W. Bruce Croft. ***Answer models for question answering passage retrieval***. SIGIR, 2004 (0 citations)
- Chirag Shah, W. Bruce Croft. ***Evaluating high accuracy retrieval techniques***. SIGIR, 2004 (1 citation)
- Haizheng Zhang, W. Bruce Croft, Brian N. Levine, Victor R. Lesser. ***A Multi-Agent Approach for Peer-to-Peer Based Information Retrieval System***. AAMAS, 2004 (1 citation)
- Donald Metzler, Victor Lavrenko, W. Bruce Croft. ***Formal multiple-bernoulli models for language modeling***. SIGIR, 2004 (0 citations)
- Stephen Cronen-Townsend, Yu Zhou, W. Bruce Croft. ***A framework for selective query expansion***. CIKM, 2004 (0 citations)

2003
- W. Bruce Croft. ***Language Models for Information Retrieval***. ICDE, 2003 (0 citations)

Co-authors | Cited authors | **Citing authors**: (*1 to 40* of 1527)
Sorted by date | **number** | name

- **W. Bruce Croft,** *2004 2004 2003 2002 2002 2002 2002 2001 1998 1997 1997 1997 1996 1995 1995 1994 1994 1993 1992 ???? ???? ???? ???? ???? ???? ???? ????*
- **James Allan,** 2004 2003 2002 2002 2001 2001 2000 1998 1998 1996 1996 1994 1993 ???? ???? ???? ???? ???? ???? ???? ???? ???? ???? ????
- **Douglas W. Oard,** 2003 2003 2003 2002 1999 1998 1998 1996 1996 1995 ???? ???? ???? ???? ???? ???? ???? ???? ???? ???? ???? ????
- **Victor Lavrenko,** 2004 2004 2003 *2002* 2002 *2001* 2000 1998 1996 ???? ???? ???? ???? ???? *????* ???? ???? ????
- **James P. Callan,** 2004 2003 2002 2002 2001 2000 2000 1996 *1995 1994* 1994 *1994 1993 1992 ????*


Rexa: Cahoon 1999 Tolerating Latency by Prefetching Java Objects
http://rexa.info/paper?id=B4FD31A7FAB50720CC761DD6D4CA11054B32
Rexa.info
▪ Research • People × Connections
Andrew McCallum • Tags • Send Invites (477) • Submit • Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.
Done
Adblock

Tolerating Latency by Prefetching Java Objects

Brendon Cahoon, Kathryn S. McKinley

To appear: Workshop on Hardware Support for Objects and Microarchitectures for Java, 1999 [Edit] [Email link]

Download: ftp.cs.umass.edu, Rexa cached

Find in: Google, GScholar, Citeseer, DBLP, Yahoo!, MSN, Rexa Raw

(Add tags at right) What is a tag?

[Add Note]

<type a tag and press return>

+ to read + read + reading group + recommended + hot + seminal + survey + tutorial + classic + controversial + enjoyable

Abstract:

In recent years, processor speed has become increasingly faster than memory speed. One technique for improving memory performance is data prefetching which is successful in array-based codes but only now are researchers applying to pointer-based codes. In this paper, we evaluate a data prefetching technique, called greedy prefetching, for tolerating latency in Java programs. In greedy prefetching, when a loop or recursive method updates an object o, we prefetch objects to which o refers. We describe inter- and intraprocedural algorithms for computing objects to prefetch and we present preliminary results ... [Expand]

References: (17) Sorted by **date** | citations | alphabetically

- Alvin Roth, Gurindar S. Sohi. ***Effective Jump-Pointer Prefetching for Linked Data Structures***. ISCA, 1999 (26 citations)
- Trishul M. Chilimbi, Mark D. Hill, James R. Larus. ***Cache-Conscious Structure Layout***. PLDI, 1999 (54 citations)
- Shai Rubin, David Bernstein, Michael Rodeh. ***Virtual Cache Line: A New Technique to Improve Cache Exploitation for Recursive Data Structures***. CC, 1999 (3 citations)
- Brad Calder, Chandra Krintz, Simmi John, Todd M. Austin. ***Cache-Conscious Data Placement***. ASPLOS, 1998 (27 citations)

Bibtex Entry: [Edit]

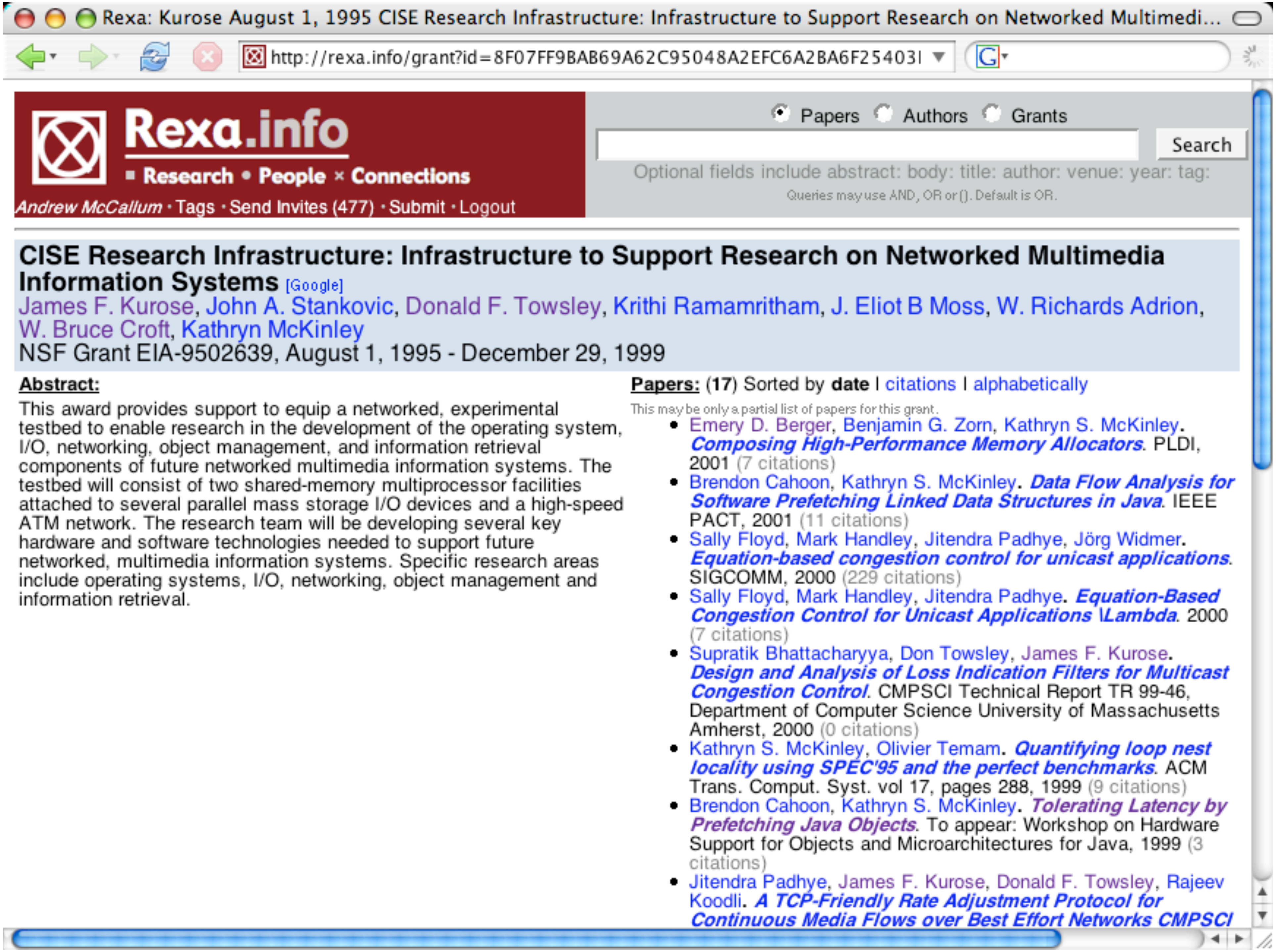
```
@inproceedings{cahoon1999tolerating,
    author = "Brendon Cahoon and Kathryn S. McKinley",
    title = "Tolerating Latency by Prefetching Java Objects",
    booktitle = "To appear: Workshop on Hardware Support for Objects
and Microarchitectures for Java",
    institution = "Department of Computer Science, University of
Massachusetts",
    year = "1999"  }
```

Topics:

cache (26.9%), experimental results (20.9%), memory (9%), object (6%), high (4.5%), java (4.5%), algorithms (4.5%), accuracy (4.5%), techniques (4.5%)

Grants: (1)

- James F. Kurose, John A. Stankovic, Donald F. Towsley, Krithi Ramamritham, J. Eliot B Moss, W. Richards Adrion, W. Bruce Croft, Kathryn McKinley. *CISE Research Infrastructure: Infrastructure to Support Research on Networked Multimedia Information Systems*. NSF EIA, 1995


Rexa: Kurose August 1, 1995 CISE Research Infrastructure: Infrastructure to Support Research on Networked Multimedi...
http://rexa.info/grant?id=8F07FF9BAB69A62C95048A2EFC6A2BA6F25403I
Rexa.info
■ Research • People × Connections
Andrew McCallum · Tags · Send Invites (477) · Submit · Logout
● Papers ○ Authors ○ Grants
Search
Optional fields include abstract: body: title: author: venue: year: tag:
Queries may use AND, OR or (). Default is OR.

CISE Research Infrastructure: Infrastructure to Support Research on Networked Multimedia Information Systems [Google]

James F. Kurose, John A. Stankovic, Donald F. Towsley, Krithi Ramamritham, J. Eliot B Moss, W. Richards Adrion, W. Bruce Croft, Kathryn McKinley

NSF Grant EIA-9502639, August 1, 1995 - December 29, 1999

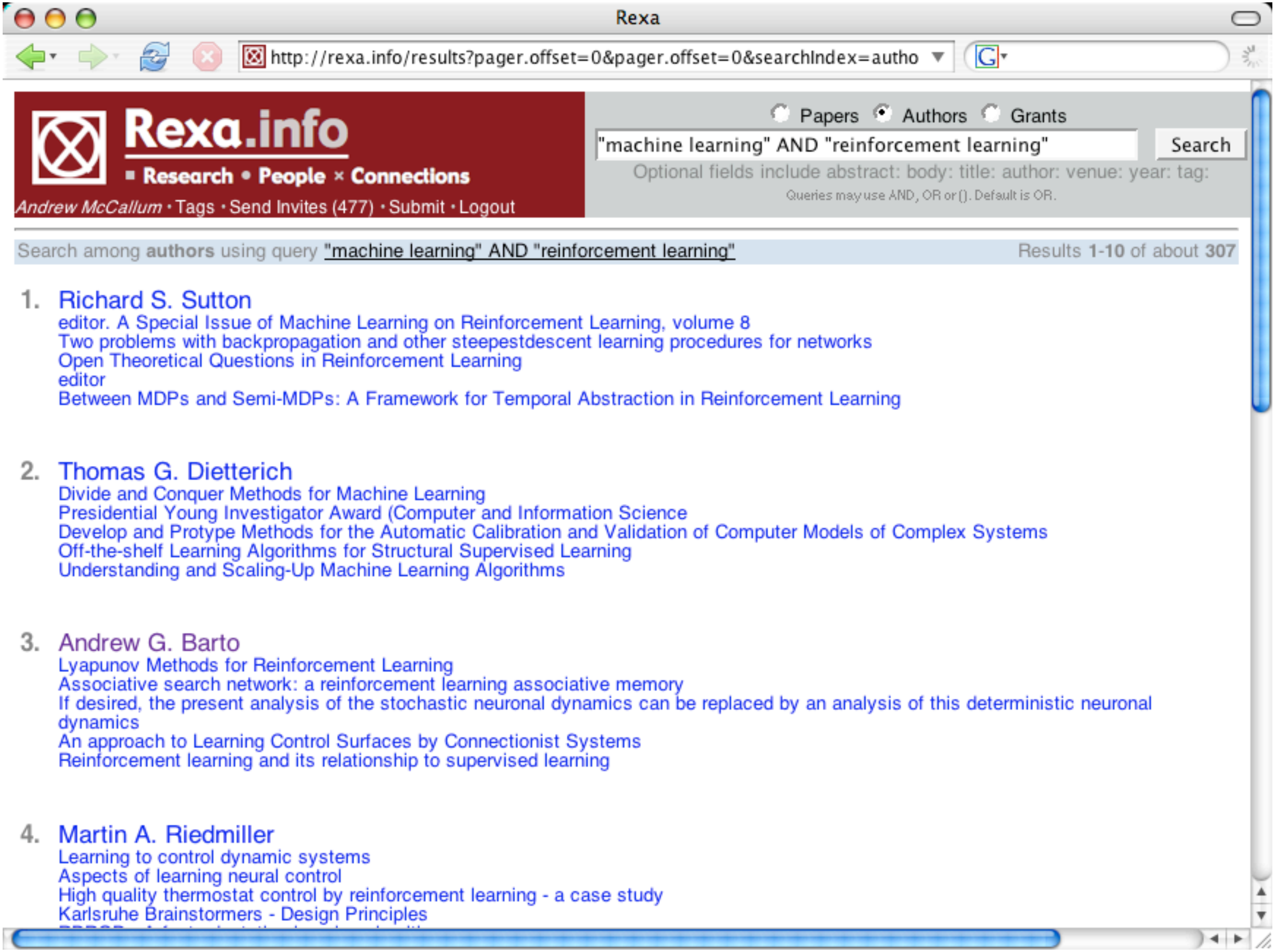
Abstract:

This award provides support to equip a networked, experimental testbed to enable research in the development of the operating system, I/O, networking, object management, and information retrieval components of future networked multimedia information systems. The testbed will consist of two shared-memory multiprocessor facilities attached to several parallel mass storage I/O devices and a high-speed ATM network. The research team will be developing several key hardware and software technologies needed to support future networked, multimedia information systems. Specific research areas include operating systems, I/O, networking, object management and information retrieval.

Papers: (17) Sorted by **date** | citations | alphabetically

This may be only a partial list of papers for this grant.

- Emery D. Berger, Benjamin G. Zorn, Kathryn S. McKinley. ***Composing High-Performance Memory Allocators***. PLDI, 2001 (7 citations)
- Brendon Cahoon, Kathryn S. McKinley. ***Data Flow Analysis for Software Prefetching Linked Data Structures in Java***. IEEE PACT, 2001 (11 citations)
- Sally Floyd, Mark Handley, Jitendra Padhye, Jörg Widmer. ***Equation-based congestion control for unicast applications***. SIGCOMM, 2000 (229 citations)
- Sally Floyd, Mark Handley, Jitendra Padhye. ***Equation-Based Congestion Control for Unicast Applications \Lambda***. 2000 (7 citations)
- Supratik Bhattacharyya, Don Towsley, James F. Kurose. ***Design and Analysis of Loss Indication Filters for Multicast Congestion Control***. CMPSCI Technical Report TR 99-46, Department of Computer Science University of Massachusetts Amherst, 2000 (0 citations)
- Kathryn S. McKinley, Olivier Temam. ***Quantifying loop nest locality using SPEC'95 and the perfect benchmarks***. ACM Trans. Comput. Syst. vol 17, pages 288, 1999 (9 citations)
- Brendon Cahoon, Kathryn S. McKinley. ***Tolerating Latency by Prefetching Java Objects***. To appear: Workshop on Hardware Support for Objects and Microarchitectures for Java, 1999 (3 citations)
- Jitendra Padhye, James F. Kurose, Donald F. Towsley, Rajeev Koodli. ***A TCP-Friendly Rate Adjustment Protocol for Continuous Media Flows over Best Effort Networks CMPSCI***


Rexa
http://rexa.info/results?pager.offset=0&pager.offset=0&searchIndex=autho
Rexa.info
▪ Research • People × Connections
Search

Andrew McCallum • Tags • Send Invites (477) • Submit • Logout

○ Papers ● Authors ○ Grants

"machine learning" AND "reinforcement learning"

Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

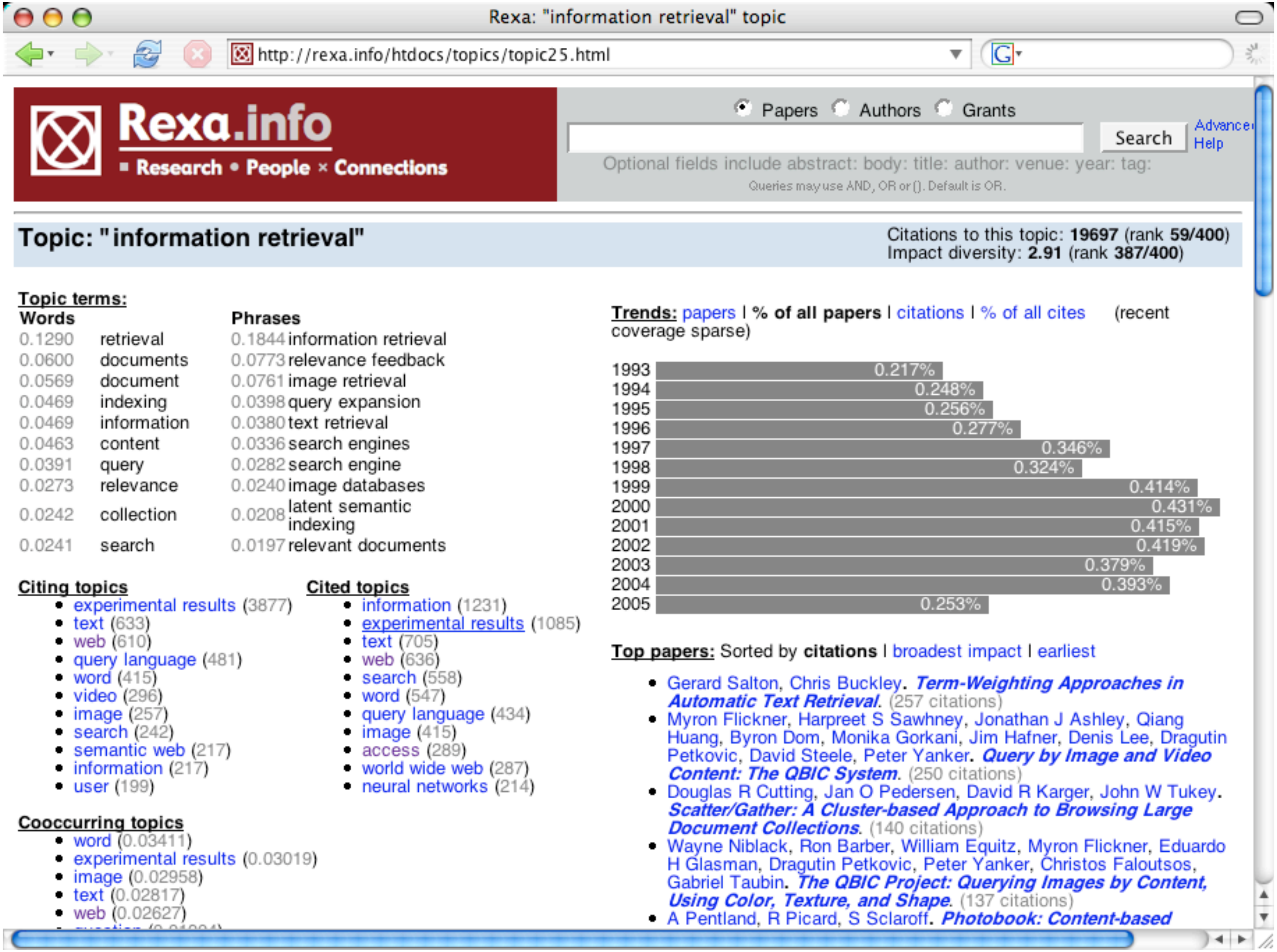
Search among **authors** using query "machine learning" AND "reinforcement learning" Results **1-10** of about **307**

1. Richard S. Sutton
   editor. A Special Issue of Machine Learning on Reinforcement Learning, volume 8
   Two problems with backpropagation and other steepestdescent learning procedures for networks
   Open Theoretical Questions in Reinforcement Learning
   editor
   Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning

2. Thomas G. Dietterich
   Divide and Conquer Methods for Machine Learning
   Presidential Young Investigator Award (Computer and Information Science
   Develop and Protype Methods for the Automatic Calibration and Validation of Computer Models of Complex Systems
   Off-the-shelf Learning Algorithms for Structural Supervised Learning
   Understanding and Scaling-Up Machine Learning Algorithms

3. Andrew G. Barto
   Lyapunov Methods for Reinforcement Learning
   Associative search network: a reinforcement learning associative memory
   If desired, the present analysis of the stochastic neuronal dynamics can be replaced by an analysis of this deterministic neuronal dynamics
   An approach to Learning Control Surfaces by Connectionist Systems
   Reinforcement learning and its relationship to supervised learning

4. Martin A. Riedmiller
   Learning to control dynamic systems
   Aspects of learning neural control
   High quality thermostat control by reinforcement learning - a case study
   Karlsruhe Brainstormers - Design Principles



Rexa: "information retrieval" topic
http://rexa.info/htdocs/topics/topic25.html
Rexa.info
Research • People × Connections
● Papers ○ Authors ○ Grants
Search
Advanced
Help

Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

Topic: "information retrieval"

Citations to this topic: **19697** (rank **59/400**)
Impact diversity: **2.91** (rank **387/400**)

Topic terms:

Words		**Phrases**	
0.1290	retrieval	0.1844	information retrieval
0.0600	documents	0.0773	relevance feedback
0.0569	document	0.0761	image retrieval
0.0469	indexing	0.0398	query expansion
0.0469	information	0.0380	text retrieval
0.0463	content	0.0336	search engines
0.0391	query	0.0282	search engine
0.0273	relevance	0.0240	image databases
0.0242	collection	0.0208	latent semantic indexing
0.0241	search	0.0197	relevant documents

Citing topics

- experimental results (3877)
- text (633)
- web (610)
- query language (481)
- word (415)
- video (296)
- image (257)
- search (242)
- semantic web (217)
- information (217)
- user (199)

Cited topics

- information (1231)
- experimental results (1085)
- text (705)
- web (636)
- search (558)
- word (547)
- query language (434)
- image (415)
- access (289)
- world wide web (287)
- neural networks (214)

Cooccurring topics

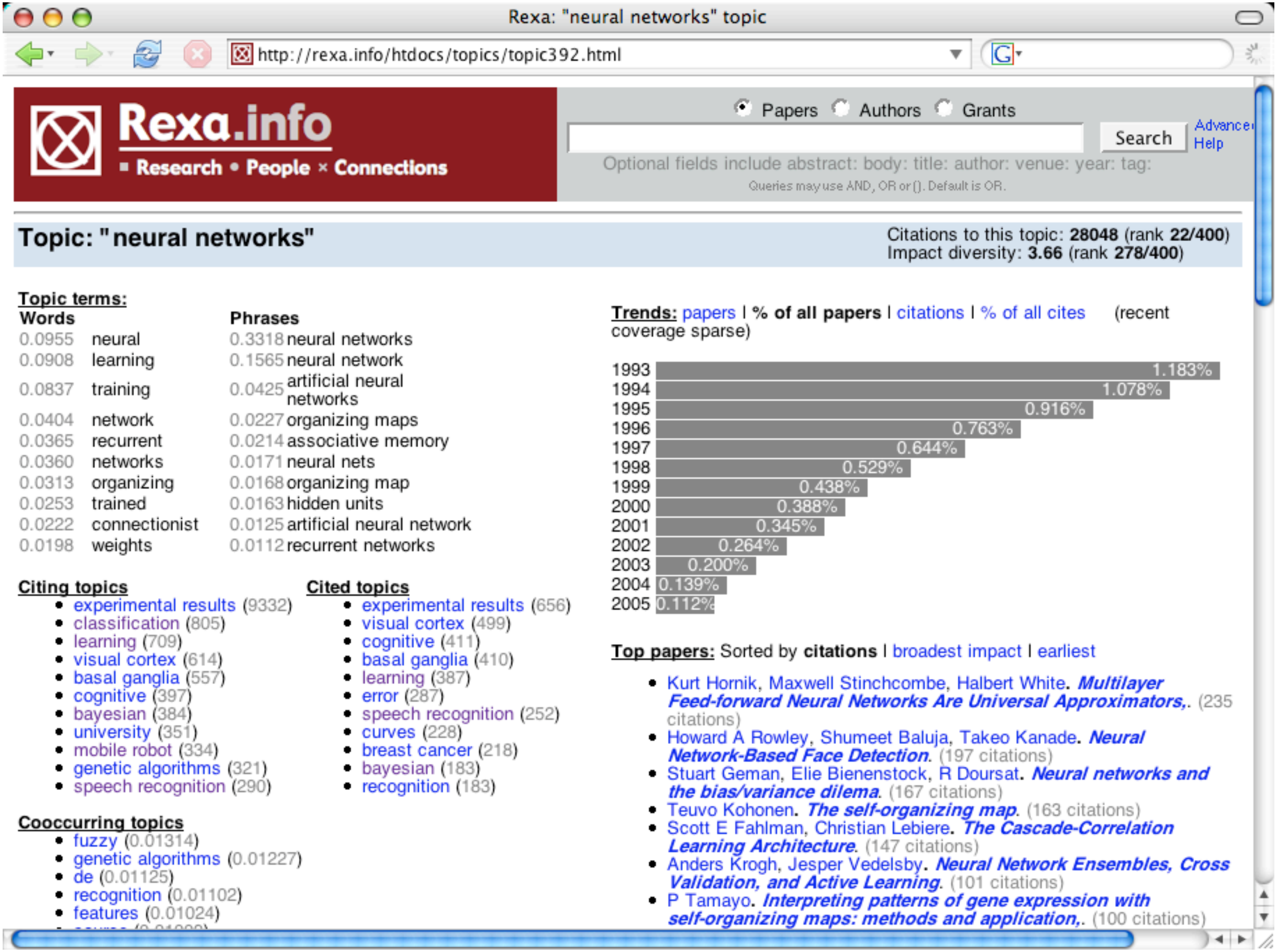
- word (0.03411)
- experimental results (0.03019)
- image (0.02958)
- text (0.02817)
- web (0.02627)

Trends: papers | **% of all papers** | citations | % of all cites (recent coverage sparse)

1993	0.217%
1994	0.248%
1995	0.256%
1996	0.277%
1997	0.346%
1998	0.324%
1999	0.414%
2000	0.431%
2001	0.415%
2002	0.419%
2003	0.379%
2004	0.393%
2005	0.253%

Top papers: Sorted by **citations** | broadest impact | earliest

- Gerard Salton, Chris Buckley. ***Term-Weighting Approaches in Automatic Text Retrieval.*** (257 citations)
- Myron Flickner, Harpreet S Sawhney, Jonathan J Ashley, Qiang Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, Peter Yanker. ***Query by Image and Video Content: The QBIC System.*** (250 citations)
- Douglas R Cutting, Jan O Pedersen, David R Karger, John W Tukey. ***Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections.*** (140 citations)
- Wayne Niblack, Ron Barber, William Equitz, Myron Flickner, Eduardo H Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, Gabriel Taubin. ***The QBIC Project: Querying Images by Content, Using Color, Texture, and Shape.*** (137 citations)
- A Pentland, R Picard, S Sclaroff. ***Photobook: Content-based***



Rexa: "neural networks" topic
http://rexa.info/htdocs/topics/topic392.html
Rexa.info
▪ Research • People × Connections
● Papers ○ Authors ○ Grants
Search
Advance
Help

Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

Topic: "neural networks"

Citations to this topic: **28048** (rank **22/400**)
Impact diversity: **3.66** (rank **278/400**)

Topic terms:

Words		**Phrases**	
0.0955	neural	0.3318	neural networks
0.0908	learning	0.1565	neural network
0.0837	training	0.0425	artificial neural networks
0.0404	network	0.0227	organizing maps
0.0365	recurrent	0.0214	associative memory
0.0360	networks	0.0171	neural nets
0.0313	organizing	0.0168	organizing map
0.0253	trained	0.0163	hidden units
0.0222	connectionist	0.0125	artificial neural network
0.0198	weights	0.0112	recurrent networks

Citing topics

- experimental results (9332)
- classification (805)
- learning (709)
- visual cortex (614)
- basal ganglia (557)
- cognitive (397)
- bayesian (384)
- university (351)
- mobile robot (334)
- genetic algorithms (321)
- speech recognition (290)

Cited topics

- experimental results (656)
- visual cortex (499)
- cognitive (411)
- basal ganglia (410)
- learning (387)
- error (287)
- speech recognition (252)
- curves (228)
- breast cancer (218)
- bayesian (183)
- recognition (183)

Cooccurring topics

- fuzzy (0.01314)
- genetic algorithms (0.01227)
- de (0.01125)
- recognition (0.01102)
- features (0.01024)

Trends: papers | **% of all papers** | citations | % of all cites (recent coverage sparse)

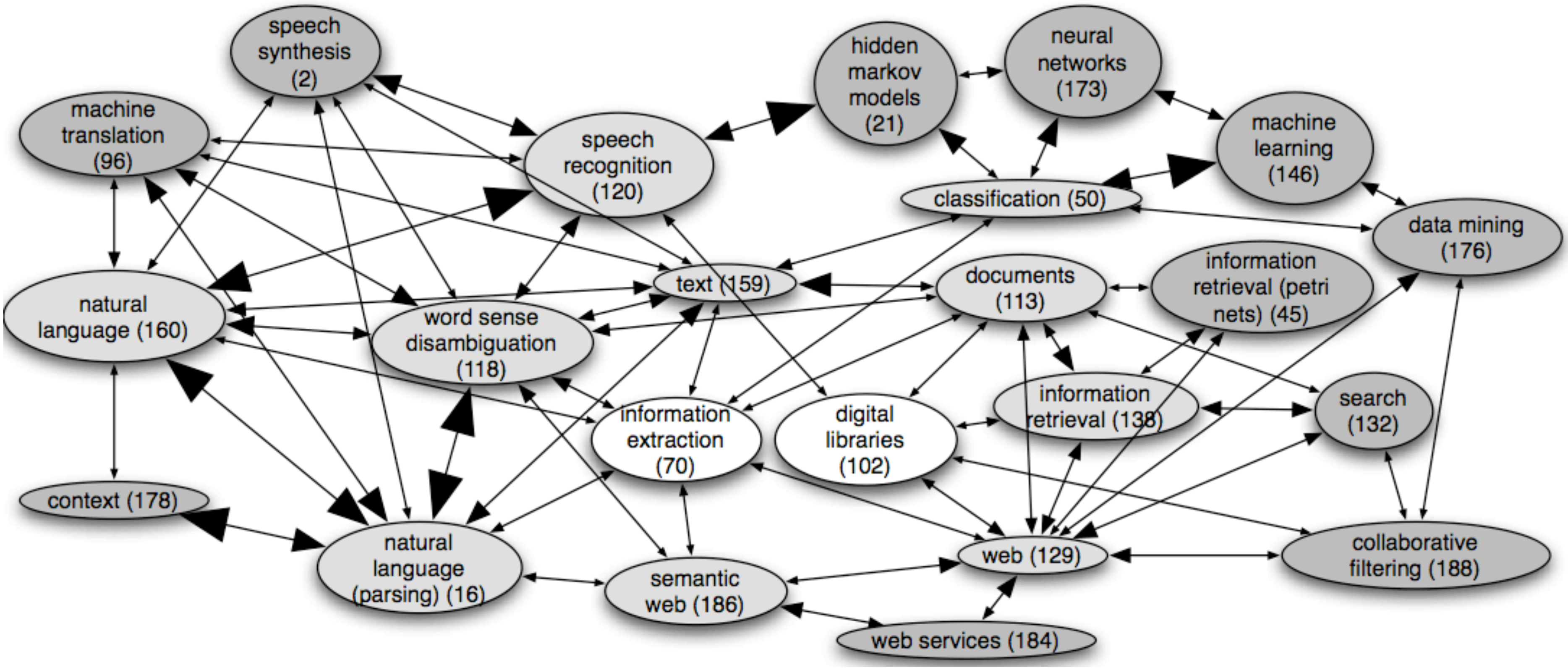
1993	1.183%
1994	1.078%
1995	0.916%
1996	0.763%
1997	0.644%
1998	0.529%
1999	0.438%
2000	0.388%
2001	0.345%
2002	0.264%
2003	0.200%
2004	0.139%
2005	0.112%

Top papers: Sorted by **citations** | broadest impact | earliest

- Kurt Hornik, Maxwell Stinchcombe, Halbert White. ***Multilayer Feed-forward Neural Networks Are Universal Approximators,***. (235 citations)
- Howard A Rowley, Shumeet Baluja, Takeo Kanade. ***Neural Network-Based Face Detection***. (197 citations)
- Stuart Geman, Elie Bienenstock, R Doursat. ***Neural networks and the bias/variance dilema***. (167 citations)
- Teuvo Kohonen. ***The self-organizing map***. (163 citations)
- Scott E Fahlman, Christian Lebiere. ***The Cascade-Correlation Learning Architecture***. (147 citations)
- Anders Krogh, Jesper Vedelsby. ***Neural Network Ensembles, Cross Validation, and Active Learning***. (101 citations)
- P Tamayo. ***Interpreting patterns of gene expression with self-organizing maps: methods and application,***. (100 citations)

Topical Transfer

Citation counts from one topic to another.
Map "producers and consumers"



Topical Bibliometric Impact Measures

[Mann, Mimno, McCallum, 2006]

- Topical Citation Counts
- Topical Impact Factors
- Topical Longevity
- Topical Precedence
- Topical Diversity
- Topical Transfer

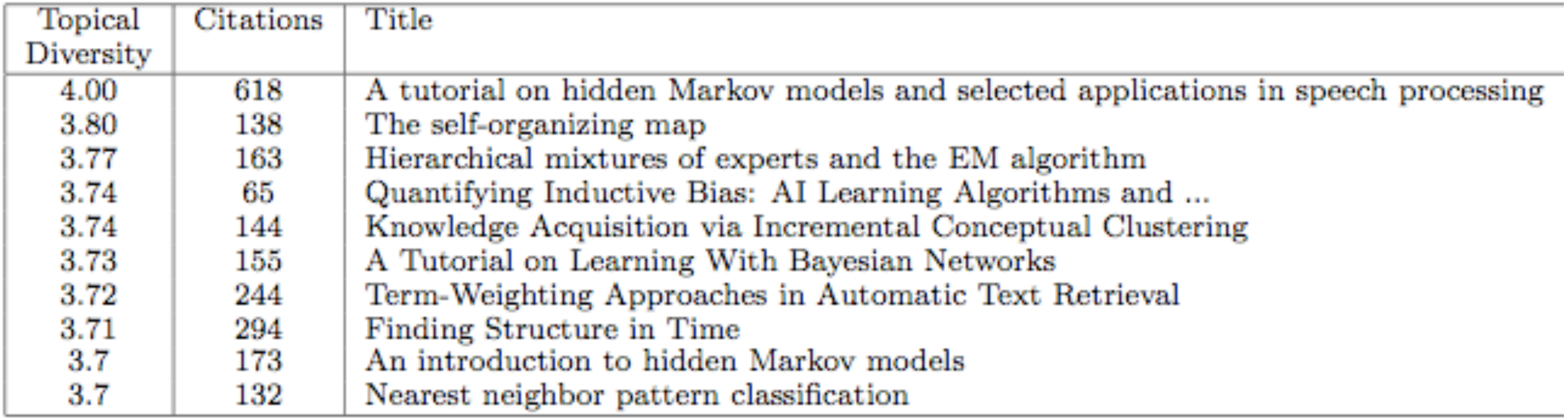
Topical Transfer

Transfer from Digital Libraries to other topics

Other topic	Cit's	Paper Title
Web Pages	31	*Trawling the Web for Emerging Cyber-Communities,* Kumar, Raghavan,... 1999.
Computer Vision	14	*On being 'Undigital' with digital cameras: extending the dynamic...*
Video	12	*Lessons learned from the creation and deployment of a terabyte digital video libr..*
Graphs	12	*Trawling the Web for Emerging Cyber-Communities*
Web Pages	11	*WebBase: a repository of Web pages*

Topical Diversity

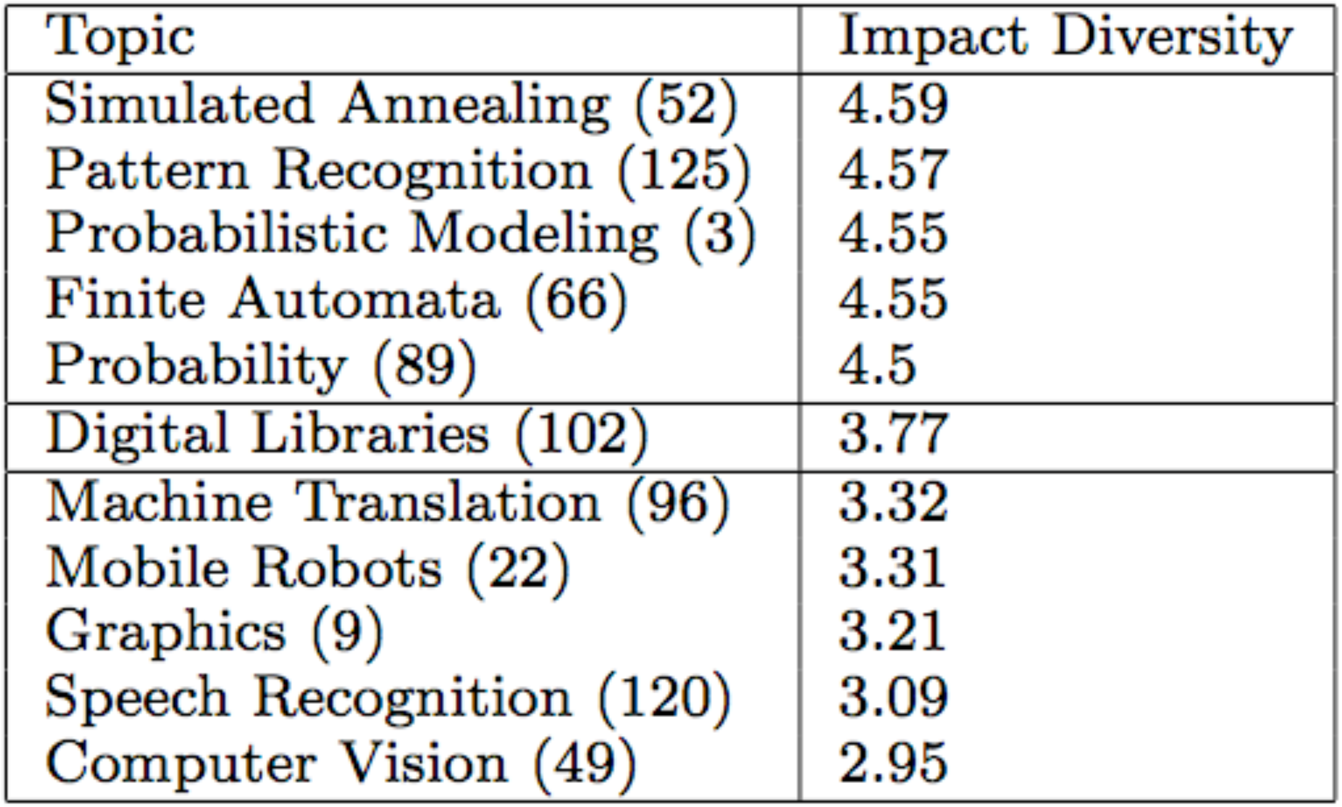
Papers that had the most influence across many other fields...



Topical Diversity	Citations	Title
4.00	618	A tutorial on hidden Markov models and selected applications in speech processing
3.80	138	The self-organizing map
3.77	163	Hierarchical mixtures of experts and the EM algorithm
3.74	65	Quantifying Inductive Bias: AI Learning Algorithms and ...
3.74	144	Knowledge Acquisition via Incremental Conceptual Clustering
3.73	155	A Tutorial on Learning With Bayesian Networks
3.72	244	Term-Weighting Approaches in Automatic Text Retrieval
3.71	294	Finding Structure in Time
3.7	173	An introduction to hidden Markov models
3.7	132	Nearest neighbor pattern classification

Topical Diversity

Entropy of the topic distribution among papers that cite this paper (this topic).



Topic	Impact Diversity
Simulated Annealing (52)	4.59
Pattern Recognition (125)	4.57
Probabilistic Modeling (3)	4.55
Finite Automata (66)	4.55
Probability (89)	4.5
Digital Libraries (102)	3.77
Machine Translation (96)	3.32
Mobile Robots (22)	3.31
Graphics (9)	3.21
Speech Recognition (120)	3.09
Computer Vision (49)	2.95

High Diversity

Low Diversity

Topical Bibliometric Impact Measures

[Mann, Mimno, McCallum, 2006]

- Topical Citation Counts
- Topical Impact Factors
- Topical Longevity
- Topical Precedence
- Topical Diversity
- Topical Transfer

Topical Precedence

“Early-ness”

Within a topic, what are the earliest papers that received more than *n* citations?

Speech Recognition:

Some experiments on the recognition of speech, with one and two ears,
E. Colin Cherry (1953)

Spectrographic study of vowel reduction,
B. Lindblom (1963)

Automatic Lipreading to enhance speech recognition,
Eric D. Petajan (1965)

Effectiveness of linear prediction characteristics of the speech wave for...,
B. Atal (1974)

Automatic Recognition of Speakers from Their Voices,
B. Atal (1976)

Topical Precedence

“Early-ness”

Within a topic, what are the earliest papers that received more than *n* citations?

Information Retrieval:

On Relevance, Probabilistic Indexing and Information Retrieval,
Kuhns and Maron (1960)

Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems,
Cooper (1968)

Relevance feedback in information retrieval,
Rocchio (1971)

Relevance feedback and the optimization of retrieval effectiveness,
Salton (1971)

New experiments in relevance feedback,
Ide (1971)

Automatic Indexing of a Sound Database Using Self-organizing Neural Nets,
Feiten and Gunzel (1982)

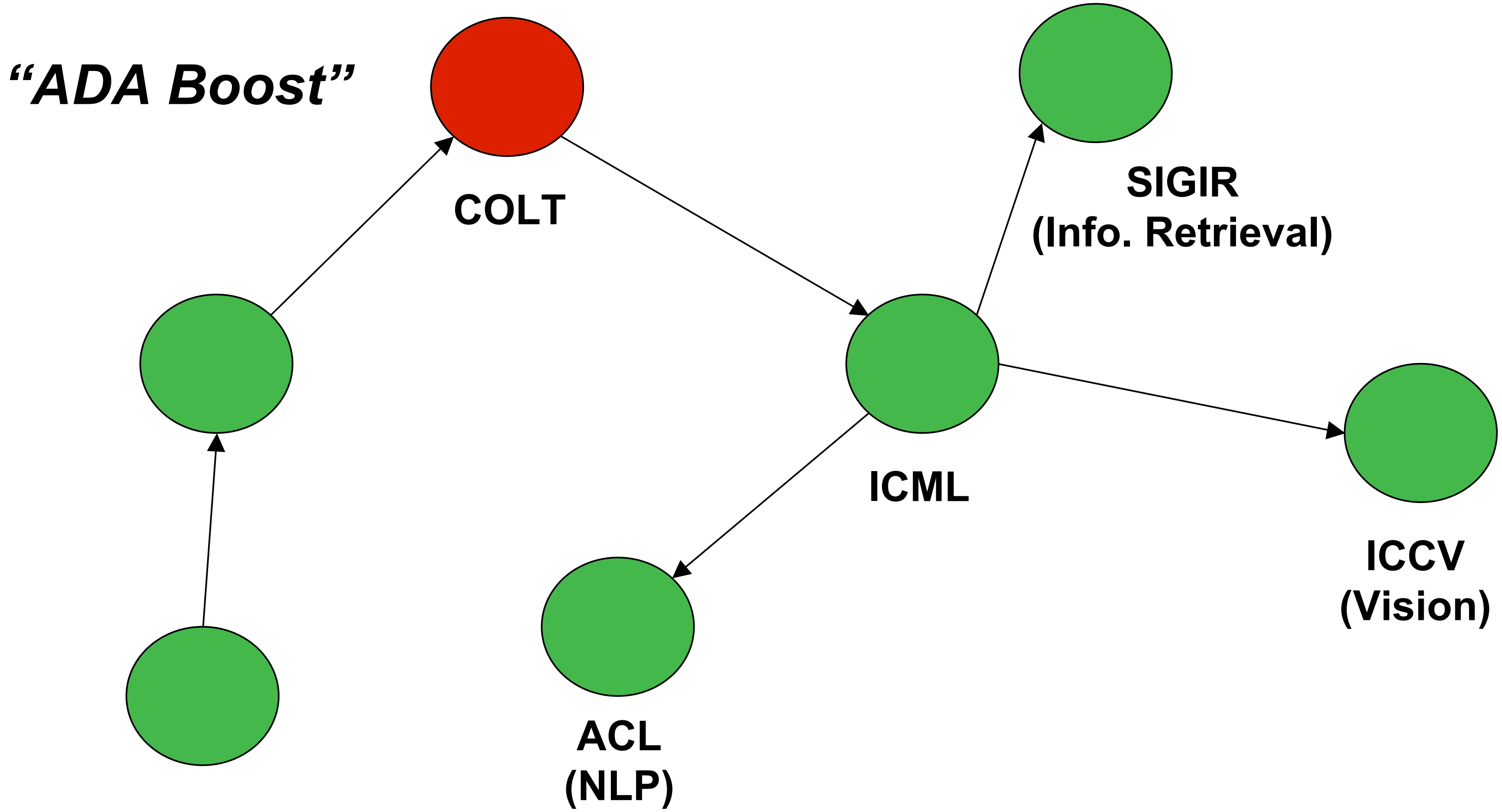
Topical Transfer Through Time

- Can we predict which research topics will be “hot” at ICML *next year*?

- ...based on
  - the hot topics in “neighboring” venues *last year*
  - learned “neighborhood” distances for venue pairs

How do Ideas Progress Through Social Networks?

Hypothetical Example:



"ADA Boost"

COLT

SIGIR
(Info. Retrieval)

ICML

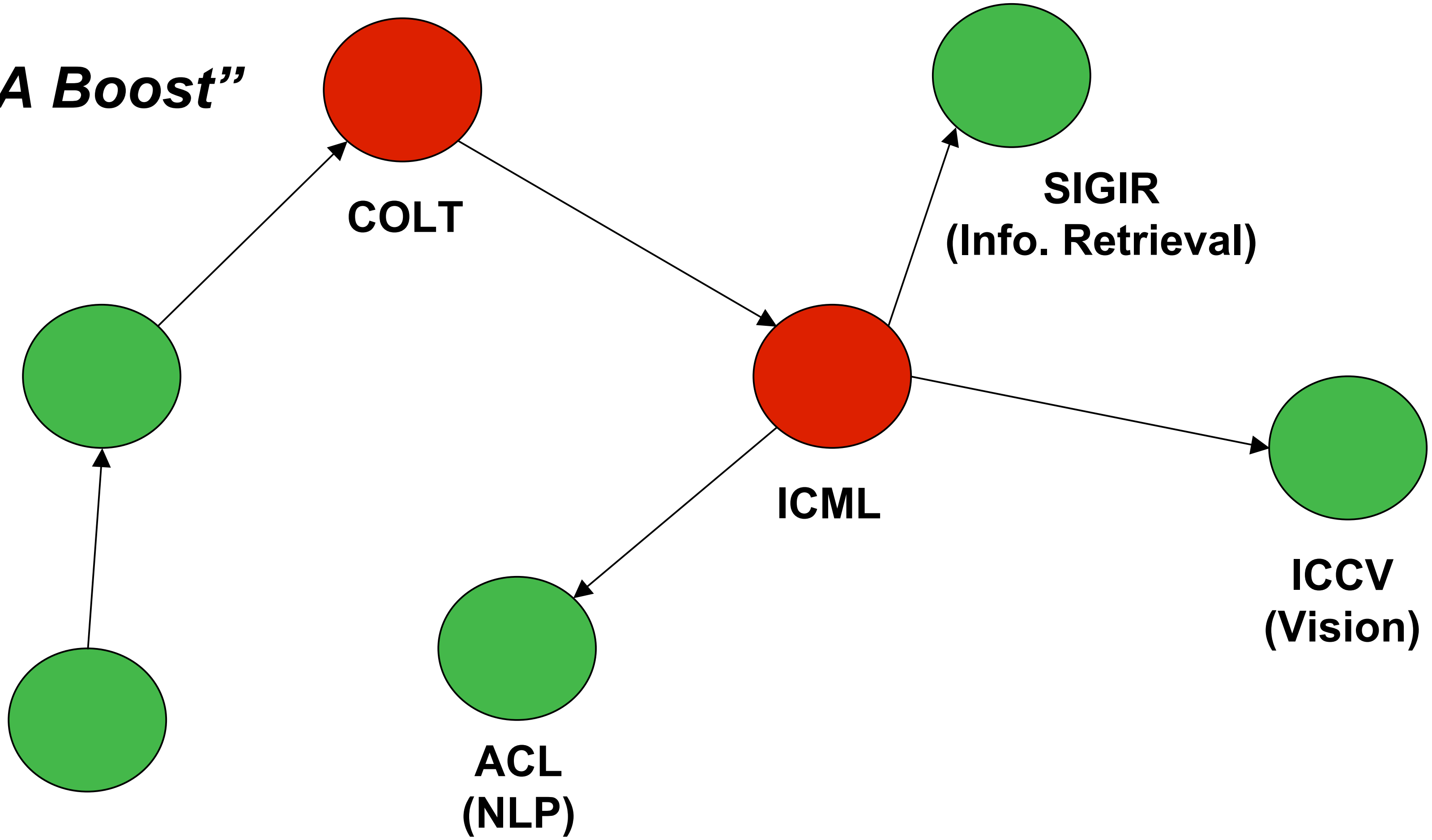
ICCV
(Vision)

ACL
(NLP)

How do Ideas Progress Through Social Networks?

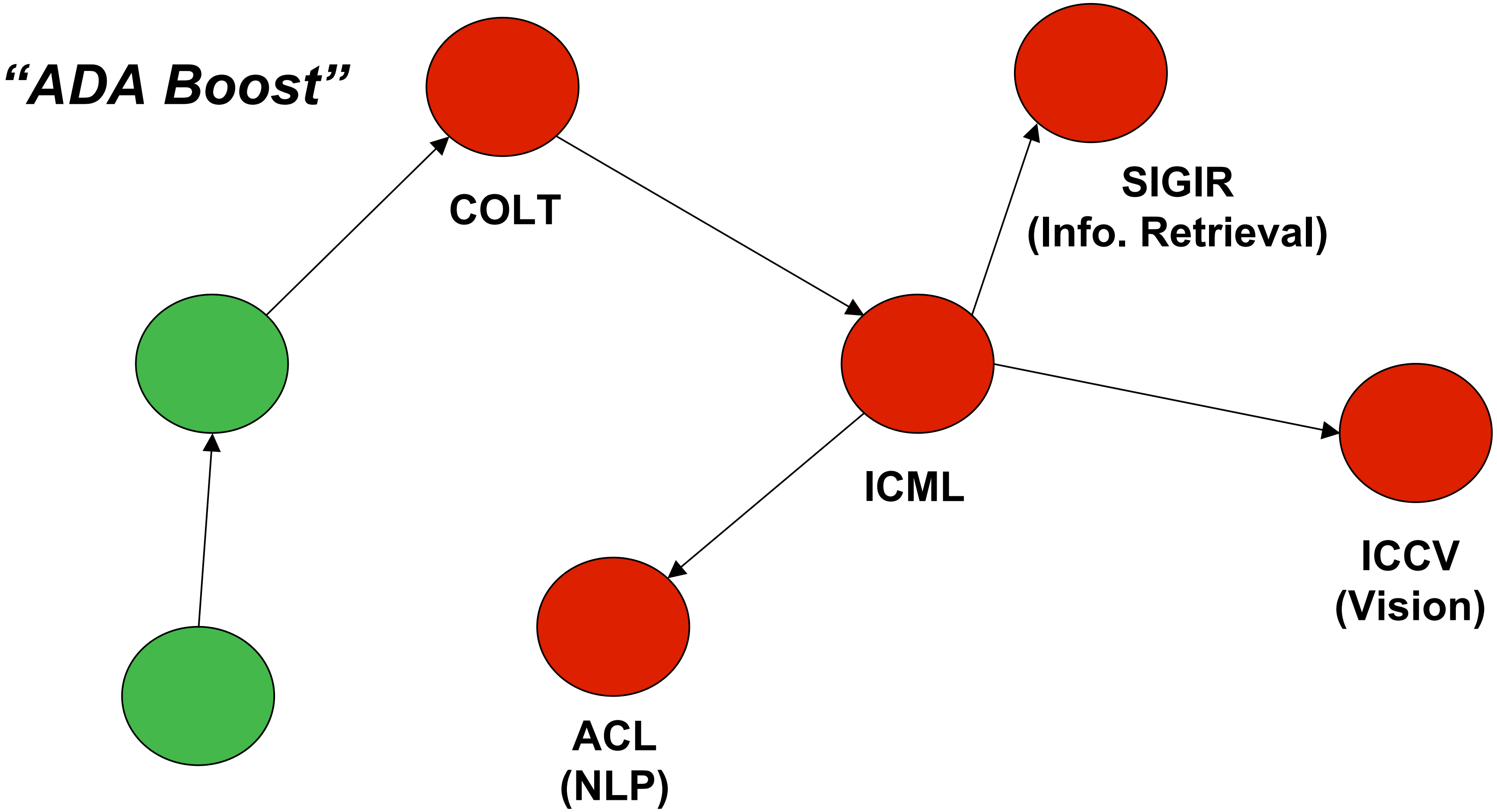
Hypothetical Example:

"ADA Boost"



How do Ideas Progress Through Social Networks?

Hypothetical Example:



Topic Prediction Models

Static Model $$Z_v^i = \lambda_v$$

Transfer Model $$Z_v^i = \lambda_v + \sum_{v'} \theta_v^{v'} Z_{v'}^{i-1}$$

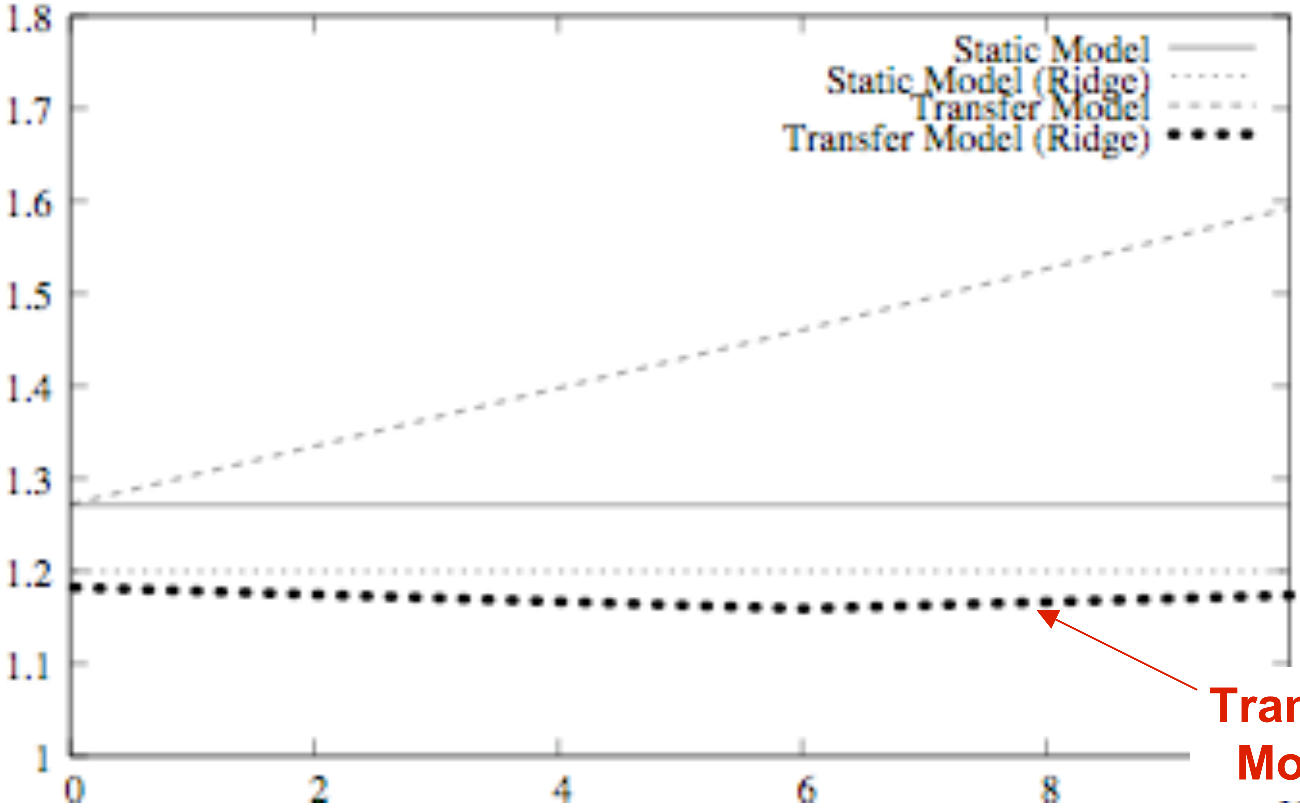
- Z_v^i : proportion of topic Z in venue v in year i
- λ_v : static topic coefficient
- $\theta_v^{v'}$: topic transfer coefficient

Linear Regression and Ridge Regression
Used for Coefficient Training.

Preliminary Results

Mean Squared Prediction Error

(Smaller Is better)



Transfer Model

# Venues used for prediction

Transfer Model with Ridge Regression is a good Predictor

Topic Model Musings

- 3 years ago Latent Dirichlet Allocation appeared as a complex innovation
  ...but now these methods & mechanics are well-understood.

- Innovation now is to understand
  data and modeling needs,
  how to structure a new model to capture these.