Automatic Construction of Topic Maps for Navigation in Information Space

ChengXiang ("Cheng") Zhai

Department of Computer Science
University of Illinois at Urbana-Champaign
http://www.cs.uiuc.edu/homes/czhai

My Group: TIMAN@UIUC

Text Data

We develop general models, algorithms, systems for

Text Data Access

Pull:
- Retrieval models
- Personalized search
- **Topic map for browsing**

Push:
- Recommender Systems

Text Data Mining

Contextual topic mining
Opinion integration and summarization
Information trustworthiness

Applications in multiple domains

12 Ph.D. students
5 MS students
5 Undergraduates

Today's talk

http://timan.cs.uiuc.edu

Combatting Information Overload: Querying vs. Browsing

Information Seeking as Sightseeing

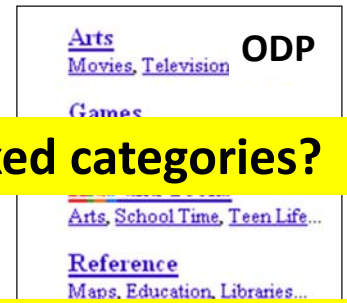
- **Know the address of an attraction site?**
  - Yes: take a taxi and go directly to the site
  - No: walk around or take a taxi to a nearby place then walk around
- **Know what exactly you want to find?**
  - Yes: use the right keywords as a query and find the information directly
  - No: browse the information space or start with a rough query and then browse

When query fails, browsing comes to rescue…

Current Support for Browsing is Limited

- **Hyperlinks**
  - Only page-to-page
  - Mostly manually constructe
  - Browsing step is very small
- **Web directories**
  - Manually constructed
  - Fixed categories
  - Only support *vertical*

Beyond hyperlinks?



Beyond fixed categories?

How to promote browsing as a "first-class citizen"?

Sightseeing Analogy Continues…

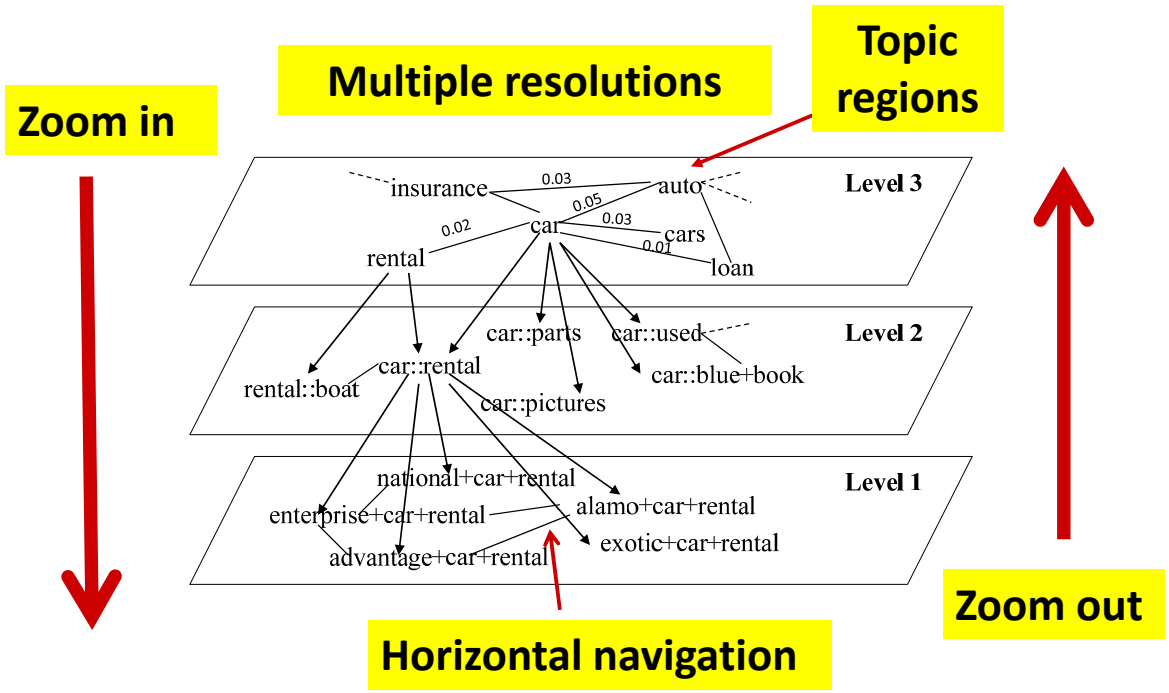
Region

Horizontal navigation

Zoom in

Zoom out

Topic Map for Touring Information Space



Topic-Map based Browsing

[Demo]

How can we construct such a multi-resolution topic map automatically?

Multiple possibilities…

Rest of the talk

- **Constructing a topic map based on user interests**
- **Constructing a topic map based on document content**
- **Summary & Future Directions**

Search Logs as Information Footprints

Footprints in information space

User 2722 searched for "*national car rental*" [!] at 2006-03-09 11:24:29
User 2722 searched for "*military car rental benefits*" [!] at 2006-03-10 09:33:37 (found http://www.valoans.com)
User 2722 searched for "*military car rental benefits*" [!] at 2006-03-10 09:33:37 (found http://benefits.military.com)
User 2722 searched for "*military car rental benefits*" [!] at 2006-03-10 09:33:37 (found http://www.avis.com)
User 2722 searched for "*enterprise rent a car*" [!] at 2006-04-05 23:37:42 (found http://www.enterprise.com)
User 2722 searched for "*meineke car care center*" [!] at 2006-05-02 09:12:49 (found http://www.meineke.com)
User 2722 searched for "*car rental*" [!] at 2006-05-25 15:54:36
User 2722 searched for "*autosave car rental*" [!] at 2006-05-25 23:26:54 (found http://eautosave.com)
User 2722 searched for "*budget car rental*" [!] at 2006-05-25 23:29:53
User 2722 searched for "*alamo car rental*" [!] at 2006-05-25 23:56:13
……

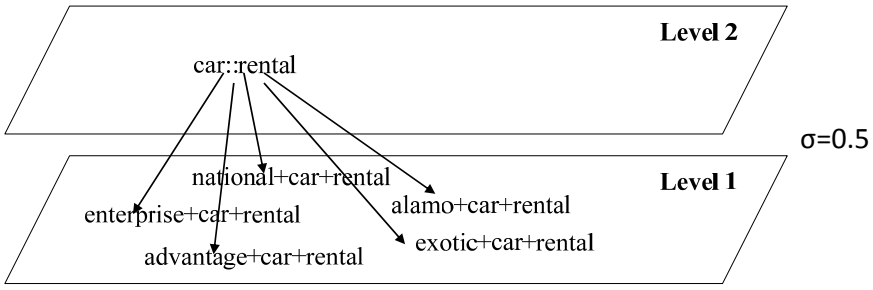
Information Footprints → Topic Map

- **Challenges**
  - How to define/construct a topic region
  - How to control granularities/resolutions of topic regions
  - How to connect topic regions to support effective browsing
- **Two approaches**
  - Multi-granularity clustering [Wang et al. CIKM 2009]
  - Query editing [Wang et al. CIKM 2008]

Xuanhui Wang, ChengXiang Zhai, **Mining term association patterns from search logs for effective query reformulation**, *Proceedings of the 17th ACM International Conference on Information and Knowledge Management* (**CIKM'08**), pages 479-488.

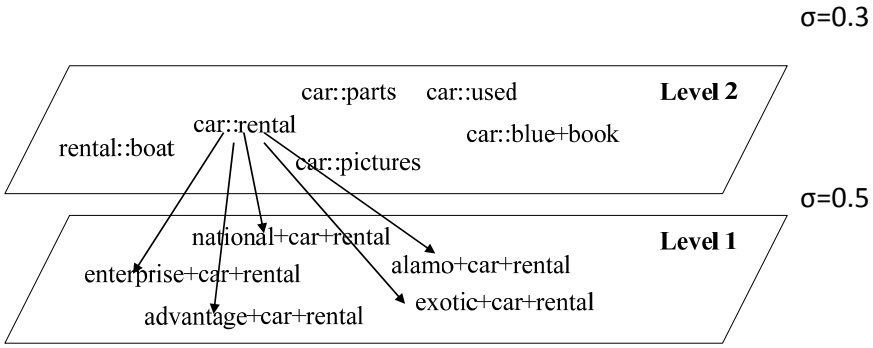
Xuanhui Wang, Bin Tan, Azadeh Shakery, ChengXiang Zhai, **Beyond Hyperlinks: Organizing Information Footprints in Search Logs to Support Effective Browsing**, *Proceedings of the 18th ACM International Conference on Information and Knowledge Management* (**CIKM'09**), pages 1237-1246, 2009.

Multi-Granularity Clustering



Star clustering

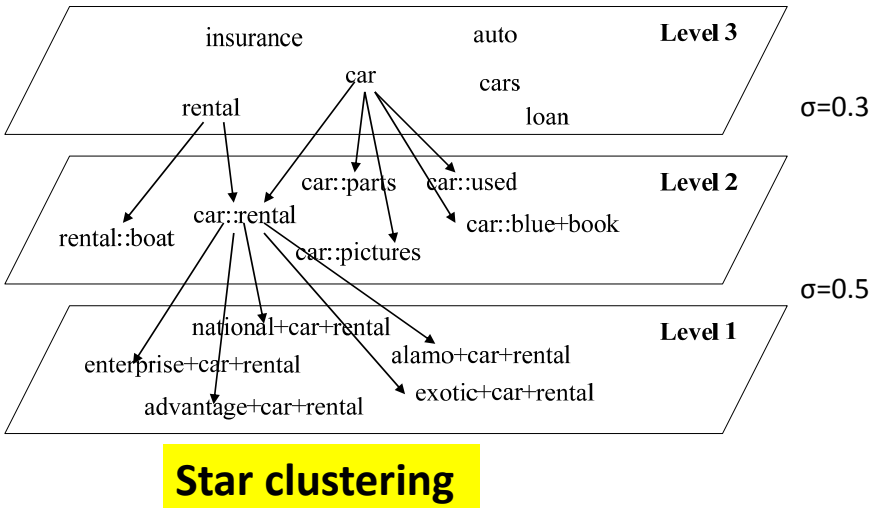
Multi-Granularity Clustering



Star clustering

Multi-Granularity Clustering

Control granularity

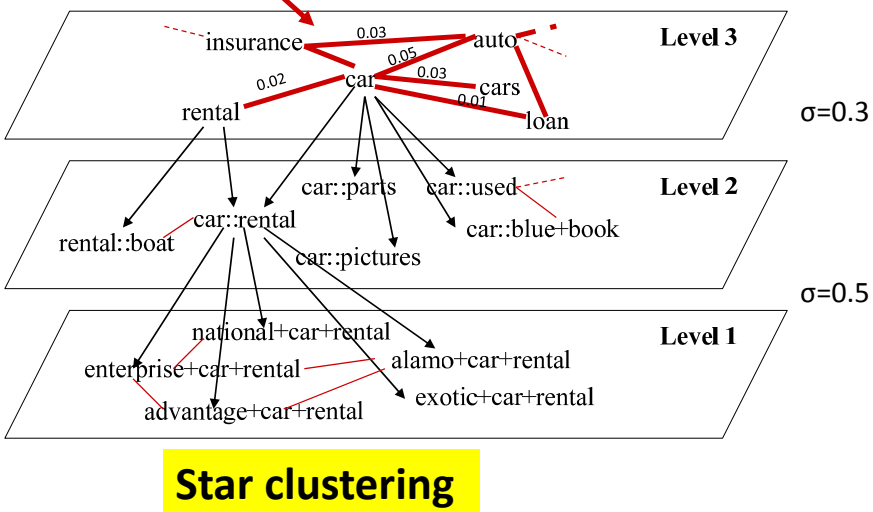


Star clustering

Multi-Granularity Clustering

Adding horizontal links

Control granularity

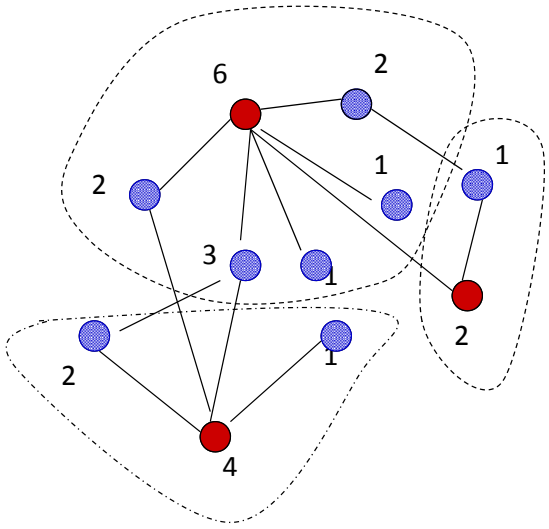


Star clustering

Star Clustering [Aslam et al. 04]

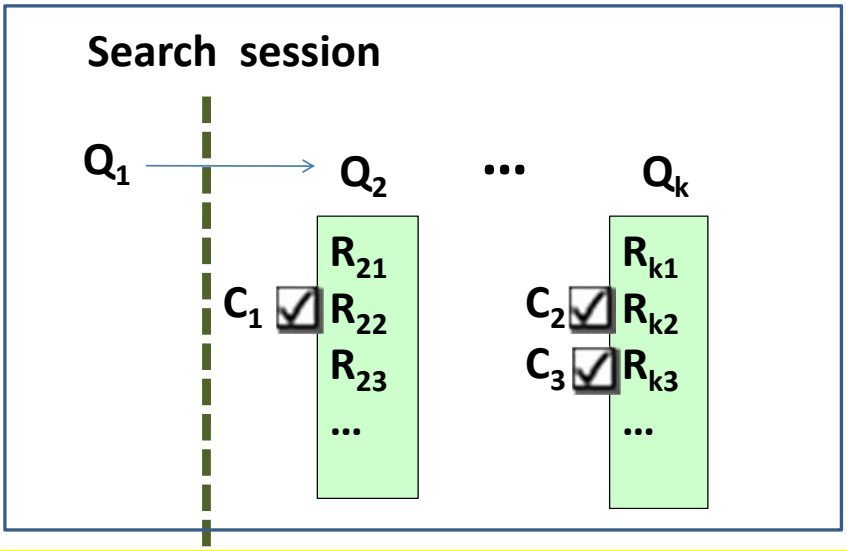
1. Form a similarity graph
   - -TF-IDF weight vectors
   - -Cosine similarity
   - -Thresholding

2. Iteratively identify a "star center" and its "satellites"



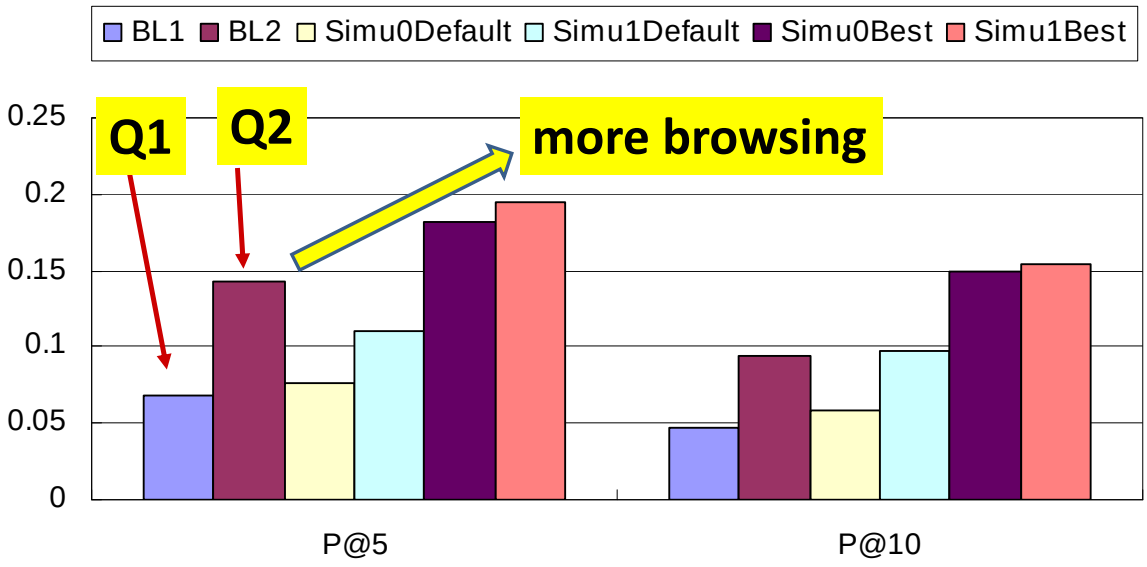
"Star center" query serves as a label for a cluster

Simulation Experiments

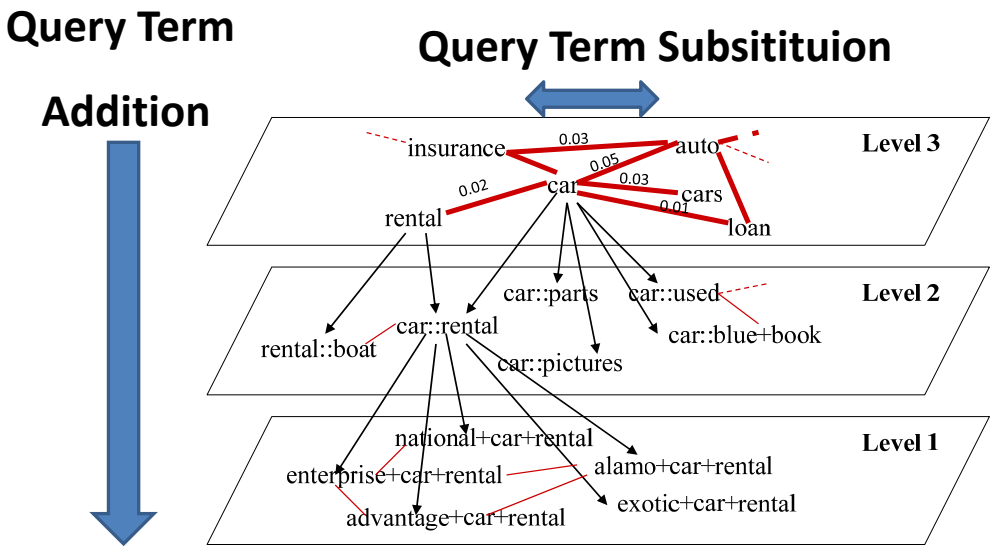


Could the user have browsed into C_1, C_2, and C_3 with a map without using Q2, …., Qk?

Browsing can be more effective than query reformulation



Topic Map as Systematic Query Editing



Map Construction
= Mining Query-Editing Patterns

- **Context-sensitive term substitution**

  auto → car | _ wash

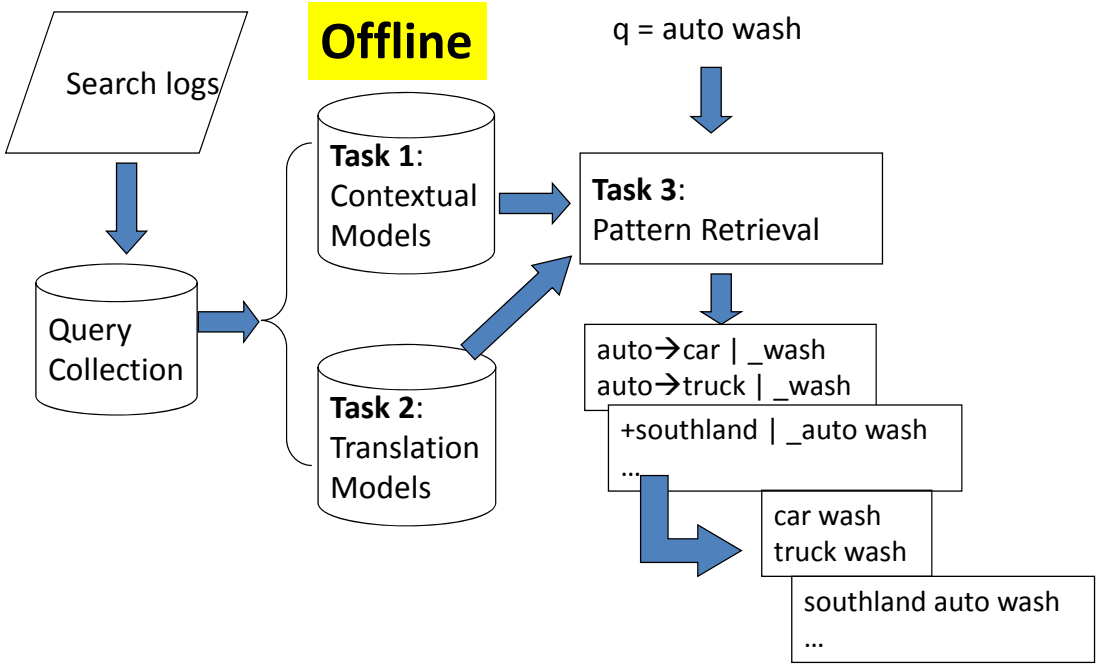
  yellowstone → glacier | _ park

- **Context-sensitive term addition**

  +sale | auto _ quotes

  +progressive | _ auto insurance

Dynamic Topic Map Construction



Examples of Contextual Models

w=car					
a	$P_G(a\|w)$	a	$P_{L_1}(a\|w)$	a	$P_{R_1}(a\|w)$
rental	0.152	rent	0.1187	rental	0.1937
rent	0.046	rental	0.0892	rentals	0.0517
rentals	0.0318	national	0.0656	seat	0.044
enterprise	0.0306	classic	0.0451	audio	0.0403
national	0.0301	enterprise	0.0396	dealers	0.0312
prices	0.0271	race	0.0257	insurance	0.031
audio	0.0246	budget	0.0237	wash	0.0275
budget	0.0197	alamo	0.0235	max	0.0251
insurance	0.0192	electric	0.0182	sales	0.0246
dealers	0.0191	hertz	0.0169	loan	0.0203

- Left and Right contexts are different
- General context mixed them together

Examples of Translation Models

w=computer		w=leg		w=chinese	
s	$t(s\|w)$	s	$t(s\|w)$	s	$t(s\|w)$
computer	0.00155	leg	0.00571	chinese	0.0027
computers	0.00014	abdominal	0.00024	japanese	0.00011
laptop	0.00011	stomach	0.00024	korean	0.0001
pc	0.00009	legs	0.00024	italian	0.00009
notebook	0.00009	muscle	0.00019	greek	0.00009

- Conceptually similar keywords have high translation probabilities
- Provide possibility for exploratory search in an interactive manner

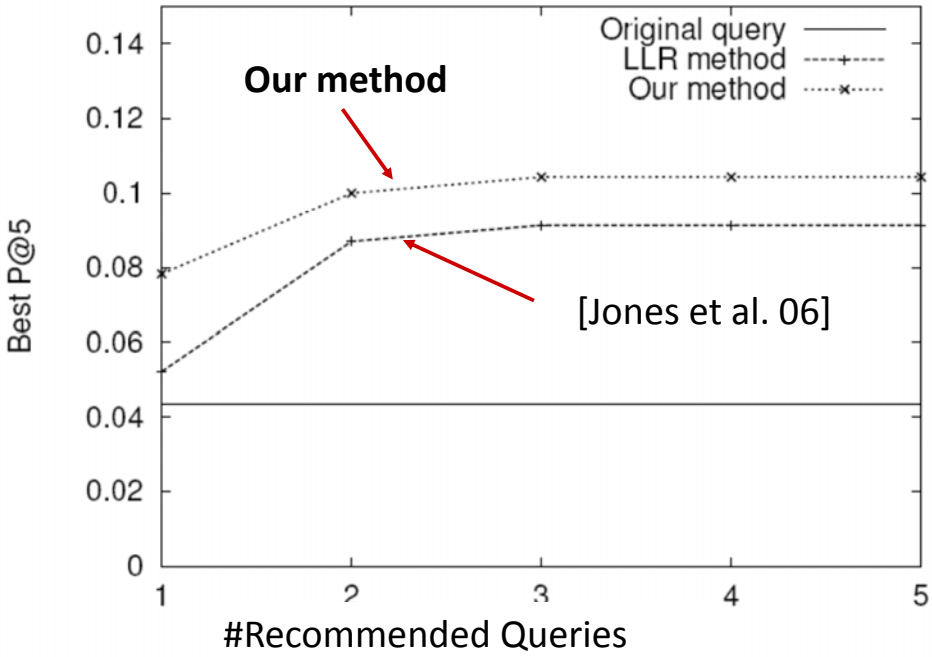
Sample Term Substitutions

Pattern	Reworded query
auto→car |_ wash	car wash
car→auto |_ trade	auto trade
children→kids |_ games	kids games
kids→children |_ clothing	children clothing
driving→maps | google_	google maps
military→army |_ acu	army acu
birthday→greeting |_ cards	greeting cards
lotto→lottery | florida _ results	florida lottery results
interpretation→meanings | _ of dreams	meanings of dreams
music→song |_ lyrics	song lyrics

Sample Term Addition Patterns

q ="wedding"	
pattern	refined query
+dresses|wedding_	wedding dresses
+cakes|wedding_	wedding cakes
+invitations|wedding_	wedding invitations
+songs|wedding_	wedding songs
+favors|wedding_	wedding favors
+flowers|wedding_	wedding flowers
+gowns|wedding_	wedding gowns
+rings|wedding_	wedding rings
+toasts|wedding_	wedding toasts
+vows|wedding_	wedding vows

Effectiveness of Query Suggestion



Rest of the talk

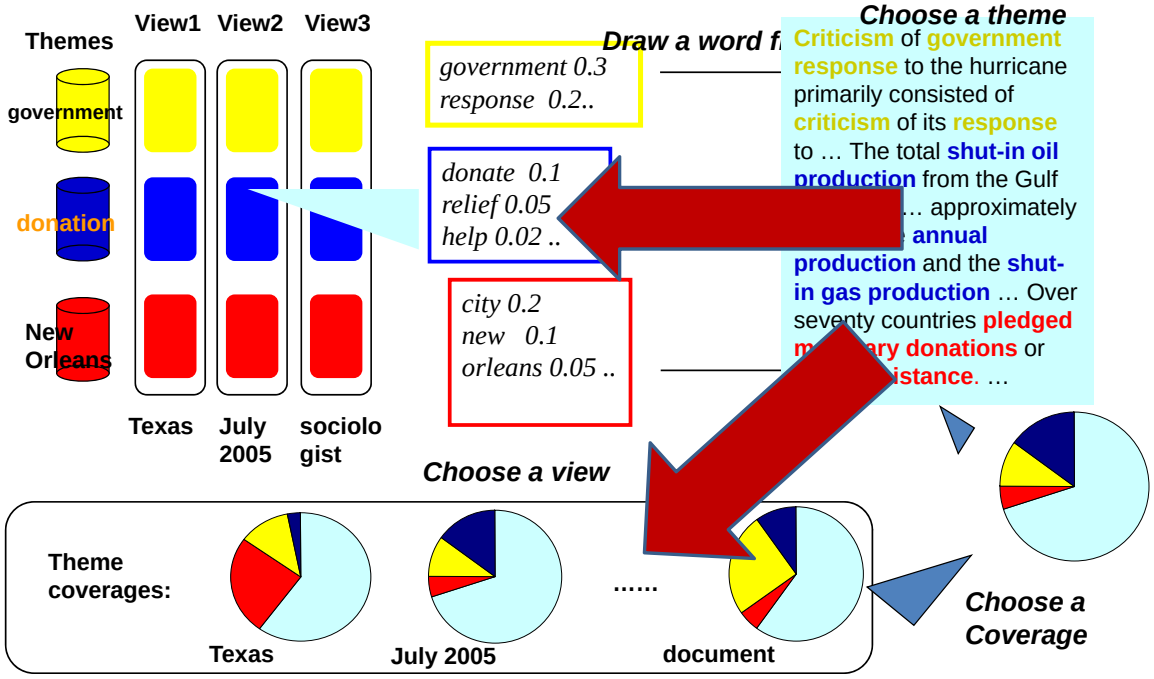
- **Constructing a topic map based on user interests**
- **Constructing a topic map based on document content**
- **Summary & Future Directions**

Document-Based Topic Map

- **Advantages over user-based map**
  - More complete coverage of topics in the information space
  - Can help satisfy long-tail information needs
- **Construction methods**
  - Traditional clustering approaches: hard to capture subtopics in text
  - Generative topic models: more promising and able to incorporate non-textual context variables
- **Two cases:**
  - Construct topic map with probabilistic latent topic analysis
  - Construct topic evolution map with probabilistic citation graph analysis

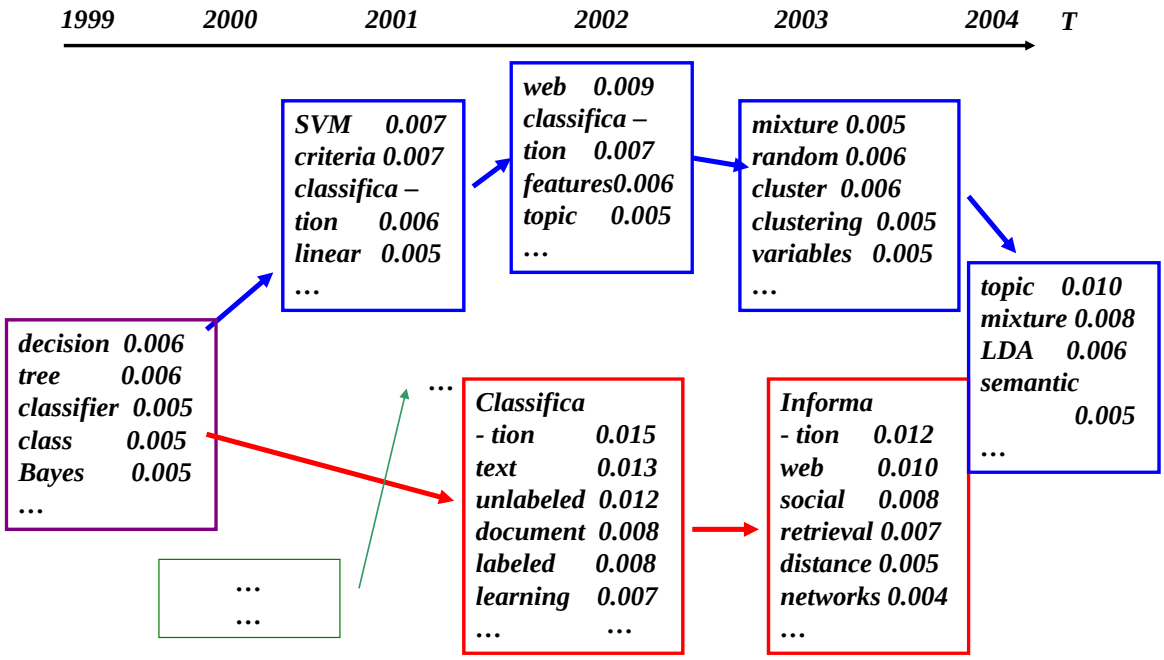
Contextual Probabilistic Latent Semantics Analysis

[Mei & Zhai KDD 2006]



Qiaozhu Mei, ChengXiang Zhai, **A Mixture Model for Contextual Text Mining**, *Proceedings of the 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , (**KDD'06**), pages 649-655.

Theme Evolution Graph: KDD [Mei & Zhai KDD 2005]



Qiaozhu Mei, ChengXiang Zhai, **Discovering Evolutionary Theme Patterns from Text -- An Exploration of Temporal Text Mining**, *Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , (**KDD'05**), pages 198-207, 2005

Joint Analysis of Text Collections and Associated Network Structures [Mei et al., WWW 2008]

Blog articles + friend network

Web page + hyperlink structure

News + geographic network

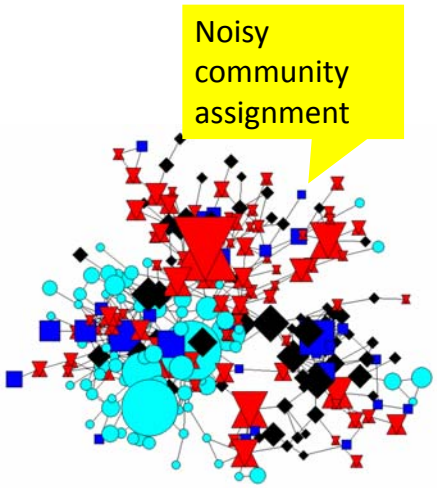
– Literature + coauthor/citation network
– Email + sender/receiver network
– …

Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. **Topic Modeling with Network Regularization**, *Proceedings of the World Wide Conference 2008* (**WWW'08**), pages 101-110

Topics from Pure Text Analysis

Topic 1		Topic 2		Topic 3		Topic 4	
term	0.02	peer	0.02	visual	0.02	interface	0.02
question	0.02	patterns	0.01	analog	0.02	towards	0.02
protein	0.01	mining	0.01	neurons	0.02	browsing	0.02
training	0.01	clusters	0.01	vlsi	0.01	xml	0.01
weighting	0.01	stream	0.01	motion	0.01	generation	0.01
multiple	0.01	frequent	0.01	chip	0.01	design	0.01
recognition	0.01	e	0.01	natural	0.01	engine	0.01
relations	0.01	page	0.01	cortex	0.01	service	0.01
library	0.01	gene	0.01	spike	0.01	social	0.01

? ? ?



Topical Communities Discovered from Joint Analysis

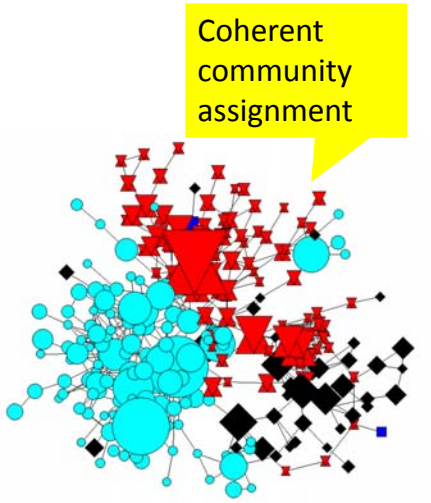
Topic 1		Topic 2		Topic 3		Topic 4	
retrieval	0.13	mining	0.11	neural	0.06	web	0.05
information	0.05	data	0.06	learning	0.02	services	0.03
document	0.03	discovery	0.03	networks	0.02	semantic	0.03
query	0.03	databases	0.02	recognition	0.02	services	0.03
text	0.03	rules	0.02	analog	0.01	peer	0.02
search	0.03	association	0.02	vlsi	0.01	ontologies	0.02
evaluation	0.02	patterns	0.02	neurons	0.01	rdf	0.02
user	0.02	frequent	0.01	gaussian	0.01	management	0.01
relevance	0.02	streams	0.01	network	0.01	ontology	0.01

Web

Information Retrieval

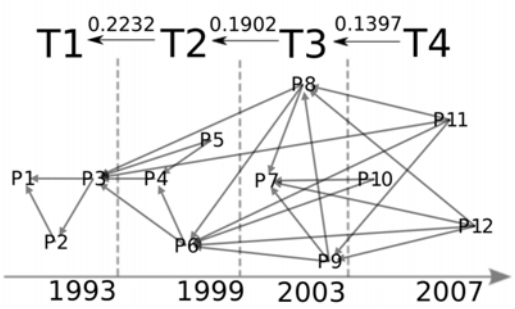
Data mining

Machine learning



Constructing Topic Evolution Map with Probabilistic Citation Analysis [Wang et al. under review]

- **Given research articles and citations in a research community**
- **Identify major research topics (themes) and their spans**
- **Construct a topic evolution map**



- **For each topic, identify milestone papers**

Probabilistic Modeling of Literature Citations

- **Modeling the generation of literature citations**
  - Document: bag of "citations"
  - Topic: distribution over documents
  - To generate a document:
    - draw topic sample $z \sim D_{doc\_topic}(z; d)$
    - draw document sample $c \sim D_{topic\_doc}(c; z)$
  - Any topic model can be used

Citation-LDA

- **Document-topic distribution:** $\theta_d \sim \mathrm{Dir}(\alpha)$
- **Topic-Document distribution:** $\varphi_z \sim \mathrm{Dir}(\beta)$

- **To generate citations in document** d_i
  - sample a topic $k \sim \mathrm{Multi}(\theta_i)$
  - sample a *document* to cite $c \sim \mathrm{Multi}(\varphi_k)$

Summarization of a Topic

- **Milestone papers: The topic-document** $\{\hat{\varphi}_{k,j}\}$ **distribution provides a natural ranking of papers**
- **Topic Key Words: weighted word counts in document titles**
- **Topic Life Span: Expected Topic Time:**

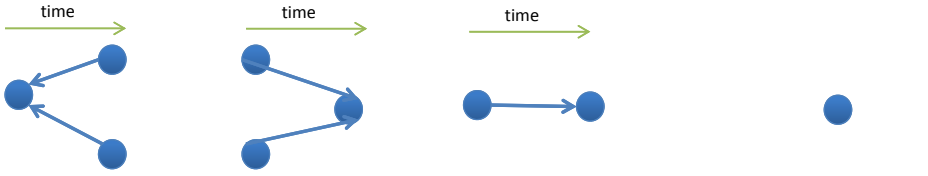
$$t(z=k) = \mathbf{E}_{z=k}[t(d)] \quad (5)$$
$$= \sum_i \Pr(d=i|z=k)t(d=i)$$

Citation Structure and Topic Evolution

- **Topic-level citation distribution:**

$$\begin{aligned} & \Pr(k^{(1)} \rightarrow k^{(2)} | k^{(1)}) \\ = & \sum_i \Pr(c = i | z = k^{(1)}) \Pr(z = k^{(2)} | d = i) \\ = & \sum_i \varphi_{k^{(1)},i} \theta_{i,k^{(2)}} \end{aligned} \quad (6)$$

- **Theme Evolution Patterns**



Sample Results: Major Topics in NLP Community

Topic	Weight	$\mathbf{E}(t)$	Top Word Phrases
94	0.02806	2005.16	dependency parsing, non-projective, shared tasks, multilingual
89	0.02761	2004.64	sentiment classification, opinion analysis, orientation, learning
8	0.02509	1991.37	word sense disambiguation, lexical semantics
92	0.02428	2004.98	machine translation, phrase-based models, alignment
96	0.02277	2005.45	machine translation, online, margin, discriminative learning
84	0.02093	2003.94	semantic role labeling, shared tasks
80	0.02069	2003.44	machine translation, reordering, alignment
73	0.01965	2002.76	discriminative parsing, sequential labeling, part-of-speech
50	0.01908	2000.87	machine translation, minimum error rate training, BLEU evaluation
72	0.01804	2002.74	coreference resolution, machine learning, anaphora, pronoun

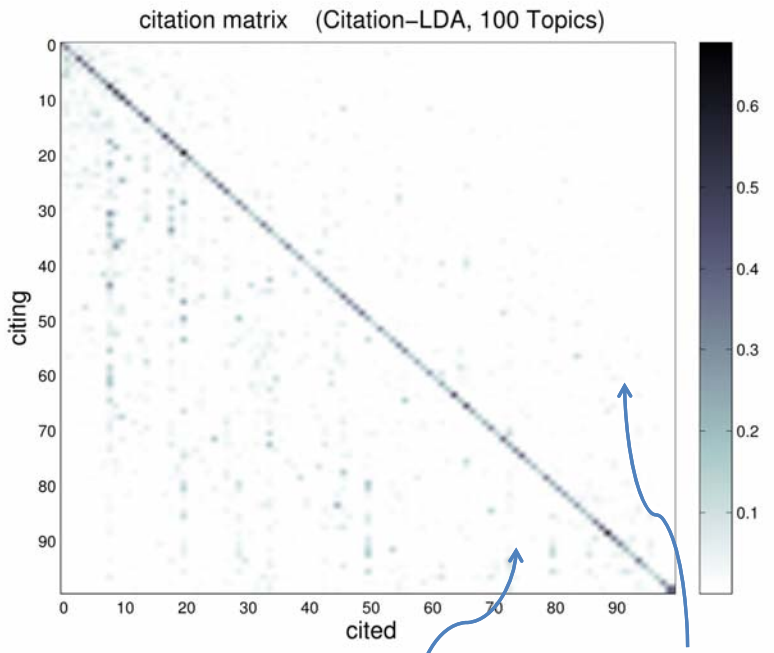
ACL Anthology Network (AAN)

Papers from NLP major conferences from 1965 - 2011

18,041 papers

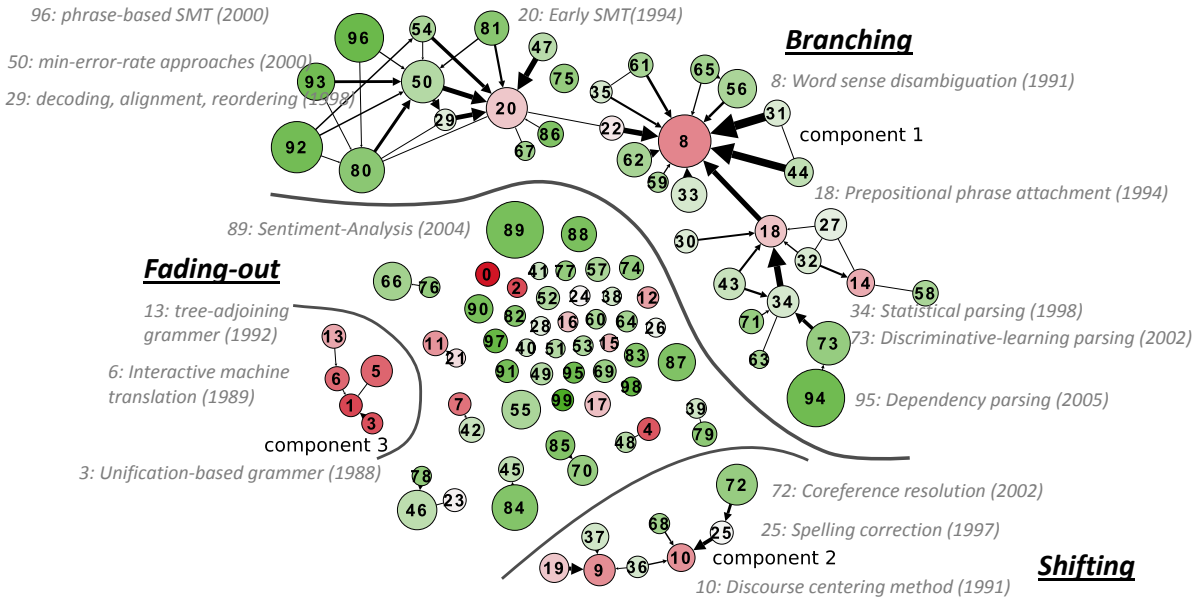
82,944 citations

Citation Structure



NLP-Community Topic Evolution

- **Topic Evolution: (green: newer, red: older)**



Detailed View of Topic "Statistical Machine Translation"

Topic	Year	ID	Paper Title
T_1	1990	P1	A Statistical Approach To Machine Translation
	1991	P2	A Program For Aligning Sentences In Bilingual Corpora
	1993	P3	The Mathematics Of Statistical Machine Translation: Parameter Estimation
T_2	1996	P4	HMM-Based Word Alignment In Statistical Translation
	1997	P5	Decoding Algorithm In Statistical Machine Translation
	1999	P6	Improved alignment models for statistical machine translation
T_3	2002	P7	Bleu: A Method For Automatic Evaluation Of Machine Translation
	2002	P8	Discriminative Training And Maximum Entropy Models For Statistical Machine Translation
	2003	P9	Minimum Error Rate Training In Statistical Machine Translation
T_4	2003	P10	Statistical Phrase-Based Translation
	2005	P11	A Hierarchical Phrase-Based Model For Statistical Machine Translation
	2007	P12	Hierarchical Phrase-Based Translation

Rest of the talk

- **Constructing a topic map based on user interests**
- **Constructing a topic map based on document content**
- **Summary & Future Directions**

Summary

- **Querying & Browsing are complementary ways of navigating in information space**
- **General support for browsing requires a topic map**
- **It's feasible to automatically construct topic maps**
  - Search logs ➔ multi-resolution topic map
  - Document content + context ➔ contextualized topic map
  - Citation graph ➔ topic evolution map
- **Topic maps naturally enable collaborative surfing**

Collaborative Surfing

New queries become new footprints

Navigation trace enriches map structures

Clickthroughs become new footprints

Browse logs offer more opportunities to understand user interests and intents

Future Research Questions

- **How do we evaluate a topic map?**
- **How do we visualize a topic map?**
- **How can we leverage ontology to construct a topic map?**
- **A navigation framework for unifying querying and browsing**
  - Formalization of a topic map
  - Algorithms for constructing a topic map
  - Topic maps with multiple views
- **A sequential decision model for optimal interactive information seeking**
  - Optimal topic/region/document ranking
  - Learn user interests and intents from browse logs + query logs
  - Intent clarification
- **Beyond information access to support knowledge service (information space→knowledge space)**

Future: Towards Multi-Mode Information Seeking & Analysis

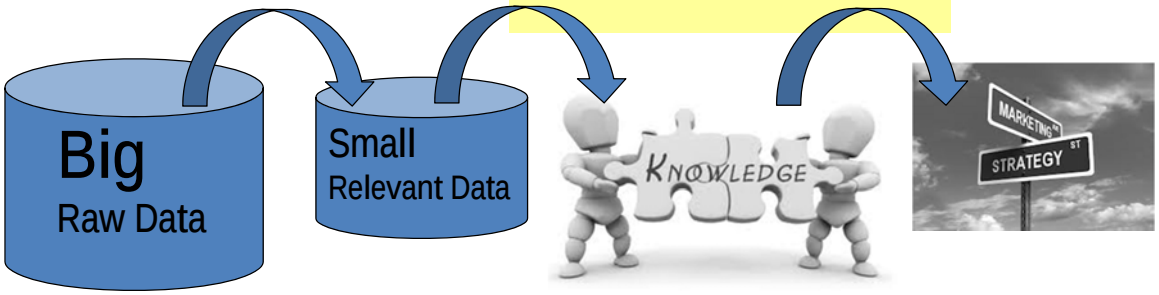
Multi-Mode Text Access

Pull: Querying + Browsing
Push: Recommendation

Multi-Mode Text Analysis

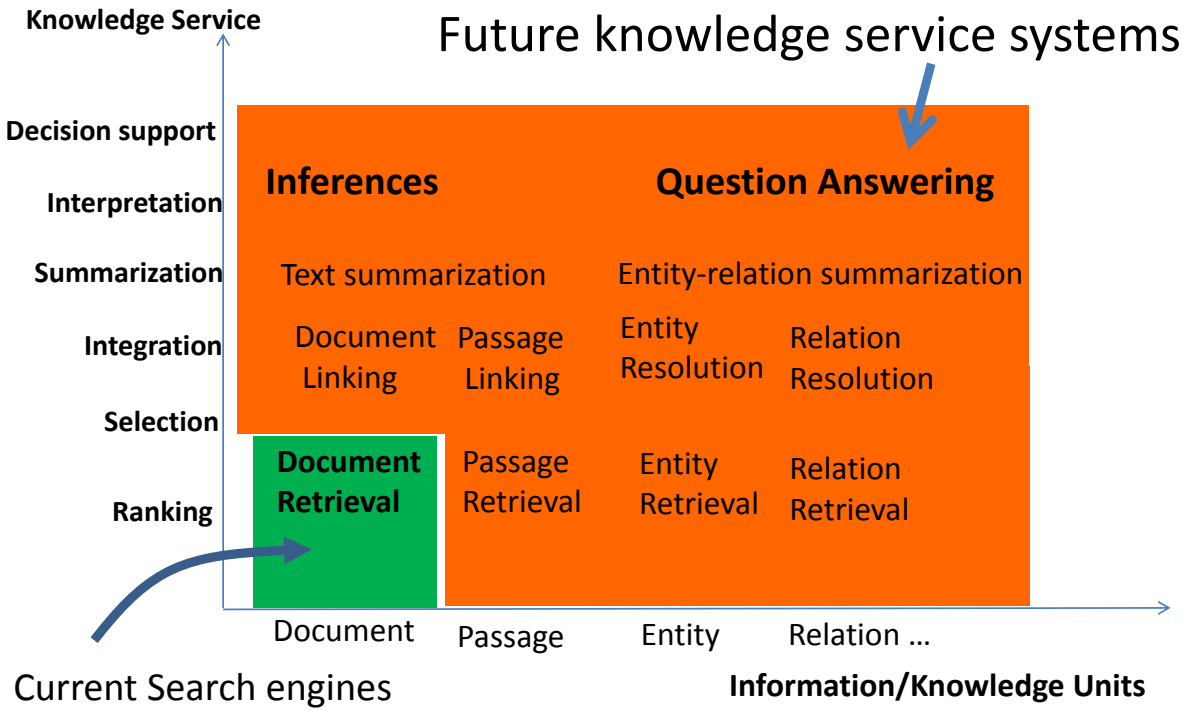
Topic extraction & analysis
Sentiment analysis
…

Interactive Decision Support



Need to develop a general framework to support all these

IKNOWX: Intelligent Knowledge Service (collaboration with Prof. Ying Ding)



Acknowledgments

- **Contributors: Xuanhui Wang, Xiaolong Wang, Qiaozhu Mei, Yanen Li, and many others**
- **Funding**

Thank You!

Questions/Comments?