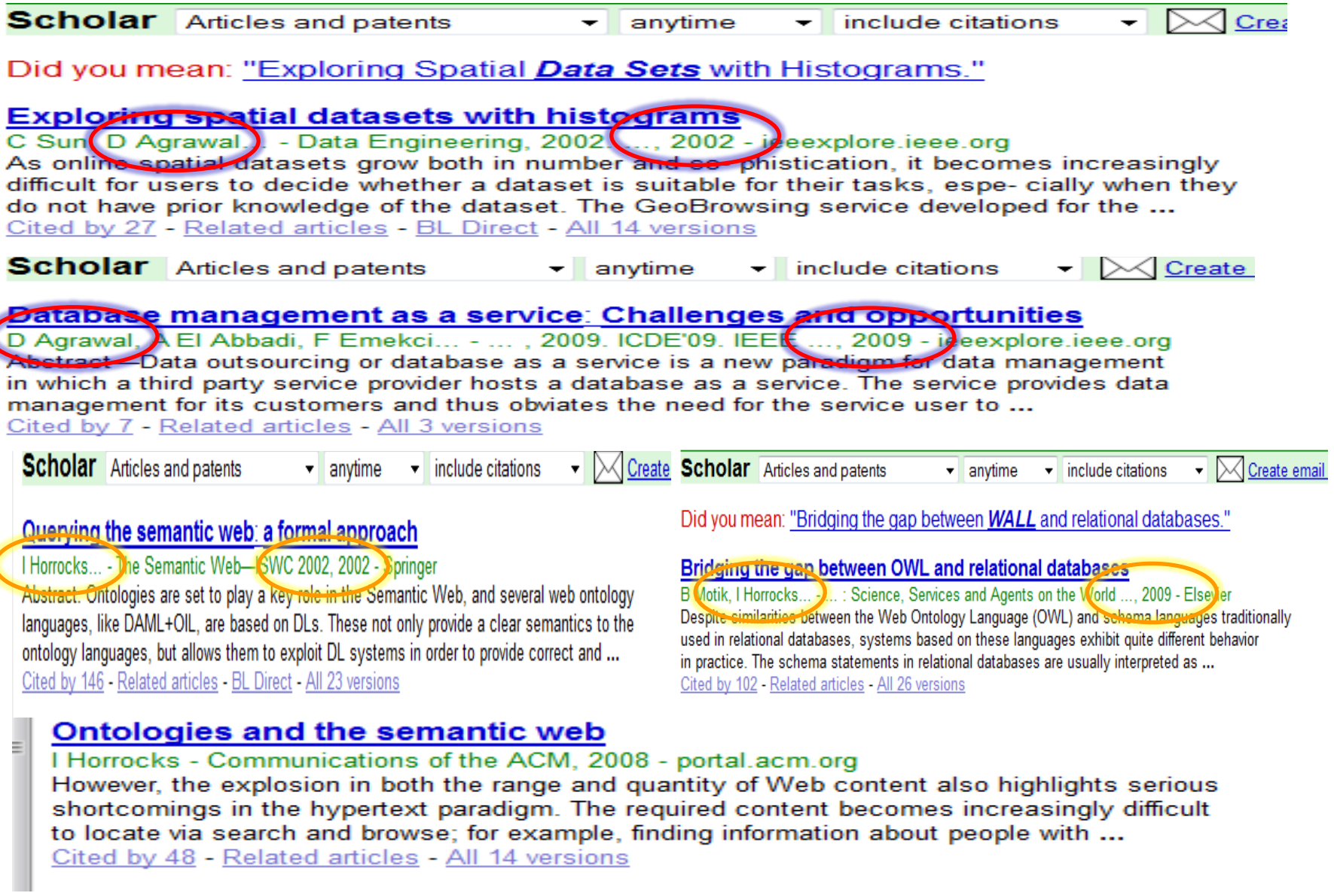
ATT: Analyzing Temporal Dynamics of Topics and Authors in Social Media

Nasir Naveed, Sergej Sizov and Steffen Staab
naveed,sizov,staab@uni-koblenz.de

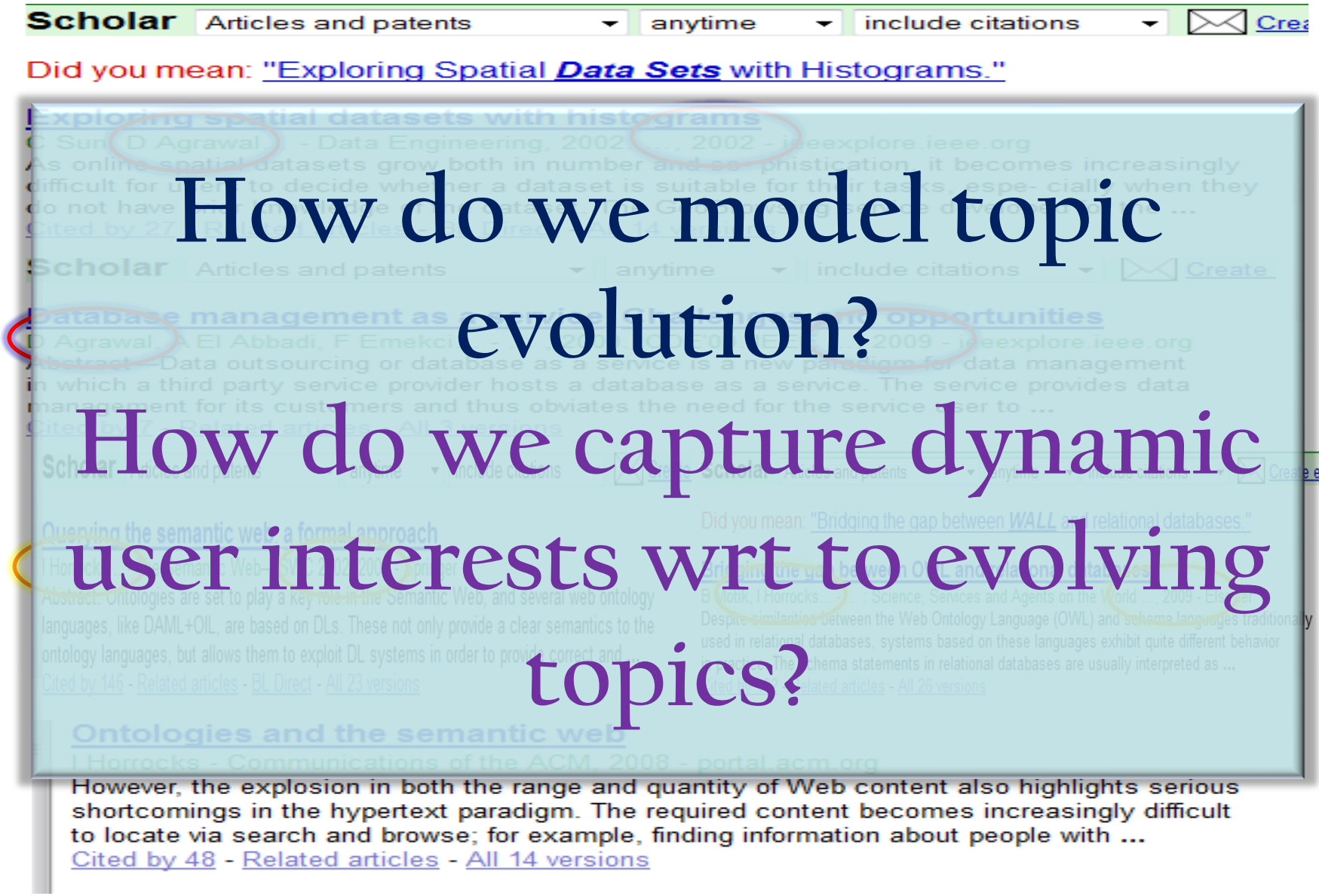
WeST
People and Knowledge Networks

UNIVERSITÄT KOBLENZ · LANDAU

Motivation


Scholar
Articles and patents
anytime
include citations
Did you mean: "Exploring Spatial Data Sets with Histograms."
Exploring spatial datasets with histograms
C Sun, D Agrawal, ... - Data Engineering, 2002..., 2002 - ieeexplore.ieee.org
As online spatial datasets grow both in number and sophistication, it becomes increasingly difficult for users to decide whether a dataset is suitable for their tasks, espe- cially when they do not have prior knowledge of the dataset. The GeoBrowsing service developed for the ...
Cited by 27 - Related articles - BL Direct - All 14 versions
Scholar
Articles and patents
anytime
include citations
Create
Database management as a service: Challenges and opportunities
D Agrawal, A El Abbadi, F Emekci... - ... , 2009. ICDE'09. IEEE ..., 2009 - ieeexplore.ieee.org
Abstract—Data outsourcing or database as a service is a new paradigm for data management in which a third party service provider hosts a database as a service. The service provides data management for its customers and thus obviates the need for the service user to ...
Cited by 7 - Related articles - All 3 versions
Scholar
Articles and patents
anytime
include citations
Create
Querying the semantic web: a formal approach
I Horrocks... - The Semantic Web—ISWC 2002, 2002 - Springer
Abstract. Ontologies are set to play a key role in the Semantic Web, and several web ontology languages, like DAML+OIL, are based on DLs. These not only provide a clear semantics to the ontology languages, but allows them to exploit DL systems in order to provide correct and ...
Cited by 146 - Related articles - BL Direct - All 23 versions
Scholar
Articles and patents
anytime
include citations
Create email
Did you mean: "Bridging the gap between WALL and relational databases."
Bridging the gap between OWL and relational databases
B Motik, I Horrocks... - ... : Science, Services and Agents on the World ..., 2009 - Elsevier
Despite similarities between the Web Ontology Language (OWL) and schema languages traditionally used in relational databases, systems based on these languages exhibit quite different behavior in practice. The schema statements in relational databases are usually interpreted as ...
Cited by 102 - Related articles - All 26 versions
Ontologies and the semantic web
I Horrocks - Communications of the ACM, 2008 - portal.acm.org
However, the explosion in both the range and quantity of Web content also highlights serious shortcomings in the hypertext paradigm. The required content becomes increasingly difficult to locate via search and browse; for example, finding information about people with ...
Cited by 48 - Related articles - All 14 versions

Motivation


Scholar
Articles and patents
anytime
include citations
Did you mean: "Exploring Spatial Data Sets with Histograms."
How do we model topic evolution?
How do we capture dynamic user interests wrt to evolving topics?
However, the explosion in both the range and quantity of Web content also highlights serious shortcomings in the hypertext paradigm. The required content becomes increasingly difficult to locate via search and browse; for example, finding information about people with ...
Cited by 48 - Related articles - All 14 versions

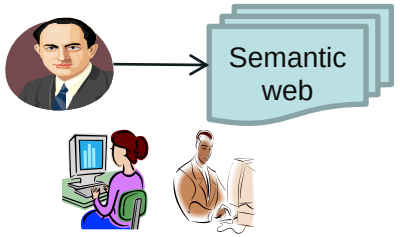
Why?

We model topics to:

- Observe topic trends
- Find related documents

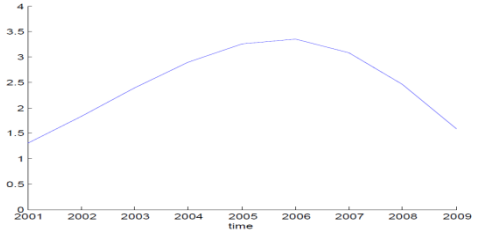
We model users interests to:

- Identify what users write “about”
- Identify authors with similar interests
- Find unusual work by an author



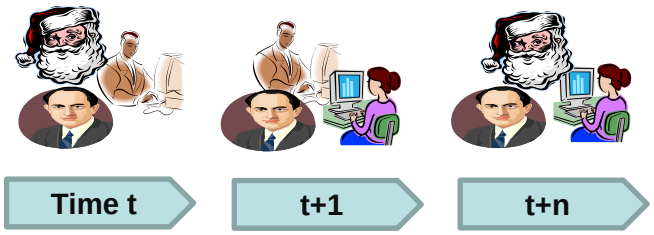
We model time to:

- Identify topic evolution
- Capture users dynamic interests

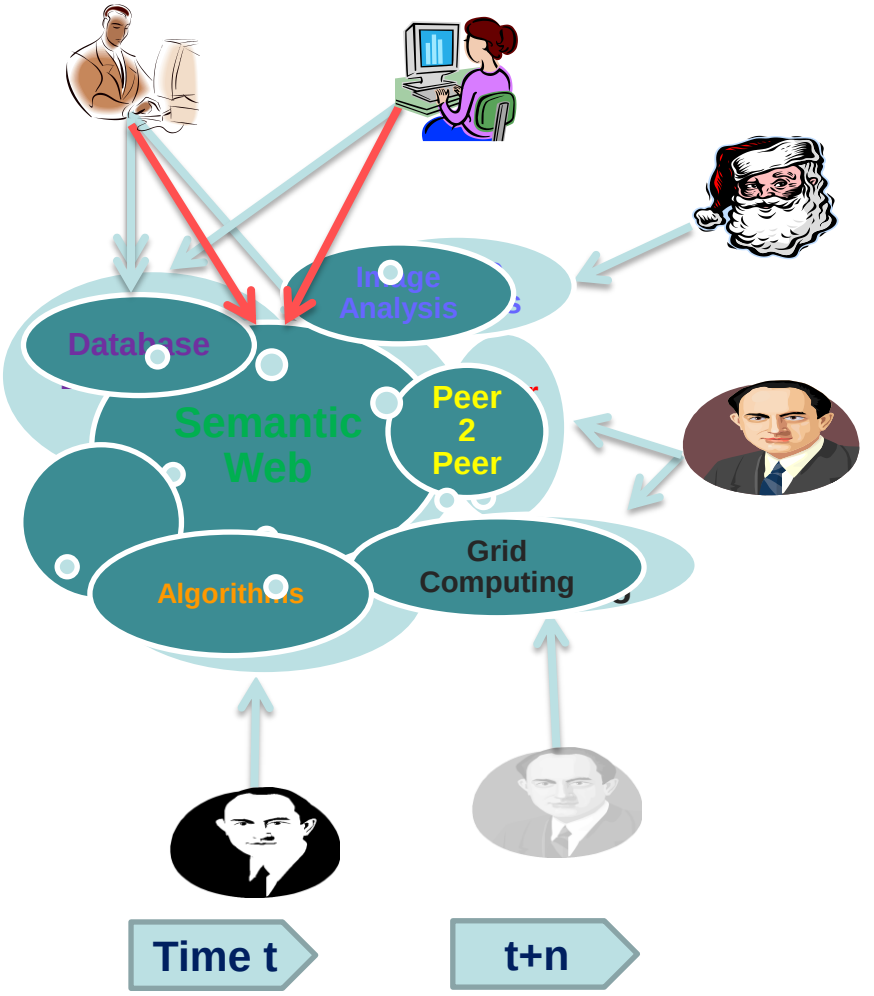


Problem Understanding

Available Input



Ideal Output



What is a Topic?

Example Topics

Top topic words

Database Management		Semantic Web	
Word	**Probability**	**Word**	**Probability**
databas	0.0223	web	0.04307
queri	0.0190	ontology	0.03784
sequenc	0.0114	semant	0.02898
control	0.0100	languag	0.02254
search	0.0095	rdf	0.01852
law	0.0095	knowledg	0.0153
molecular	0.0095	schema	0.0149
schema	0.0090	servic	0.01248

Text snippets

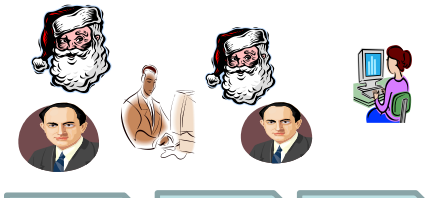
Words defining topic

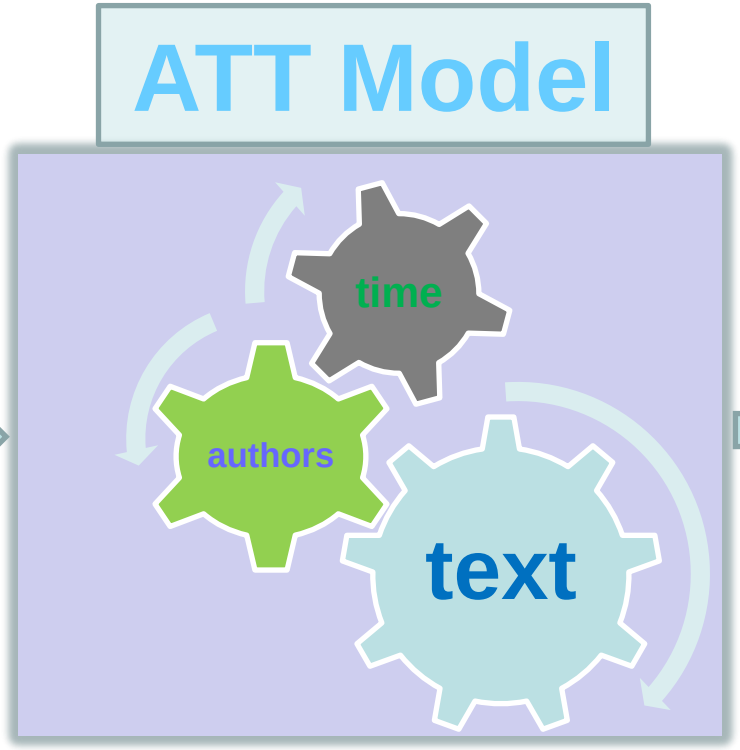
Bridging the gap between OWL and relational databases
Despite similarities between the Web Ontology Language (OWL) and schema languages traditionally used in relational databases, systems based on languages exhibit quite different behavior in pra The schema statements in relational databases usually interpreted as .....

hat Is Relational Database Schema?
What Is Relational Database Schema?. A relational database schema is the tables, columns and relationships that make up a relational database.

What Is Relational Database Schema?. A relational database schema is the tables, columns and relationships that make up a relational database.

Author-Topic-Time Model (ATT)





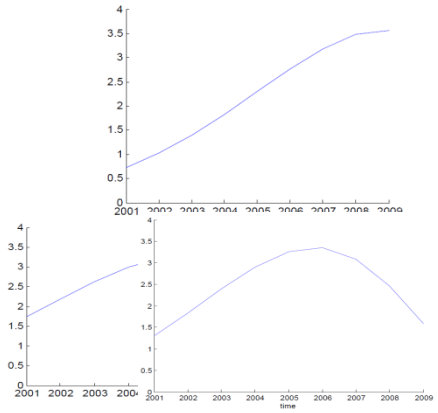


Image Analysis	
Word	**Prob.**
image	0.0185
spectral	0.0150
test	0.0130
	0.0120
	0.0120
	Prob.
	0.1201
	0.0047
	0.0025
	0.0013

Semantic Web	
Word	**Prob.**
ontolog	0.0380
web	0.0375
metadata	0.0298
rdf	0.0211
semantic	0.0195
Author	**Prob.**
I. Horrocks	0.0892
S. Staab	0.0686
A. Lin	0.0042
W. Nejdl	0.0011

Applications of Author-Topic-Time Model

Help answer questions like

- Which are the important authors for a given topic?
- Who are the pioneers of a topic?
- Given the time slice find the top topics with associated authors

Potentially useful for

- Monitoring topic evolution
- Monitoring author's profiles
- Predicting author profiles for a given interest
- ….

Experimental Dataset

CiteSeer Publications from Comptue Science Domain

- Total Docs: *5000*
- Time Period: *2001 – 2009*
- Total Authors: *18*
- No. of Publications per Author: *180 – 300*
- Data cleaing: stop word removal and stemming
- Used JAGS for implementing the ATT model

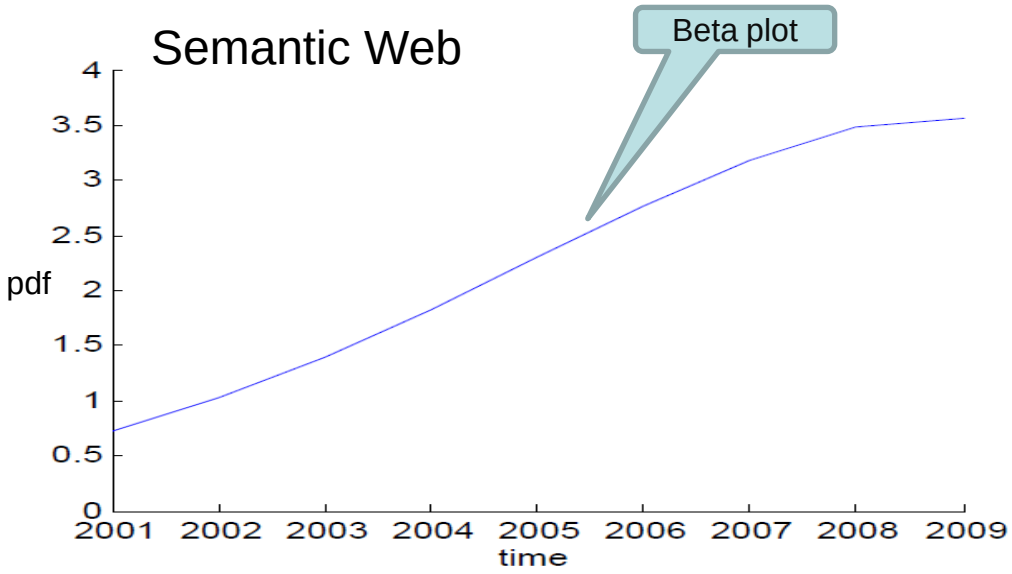
Why Limited Dataset?

- Publication limit to include some specific authors
- For manual confirmation of authors profiles
- Data cleaning to reduce noice and sparsity
- JAGS[1] is not scalable to larger dataset

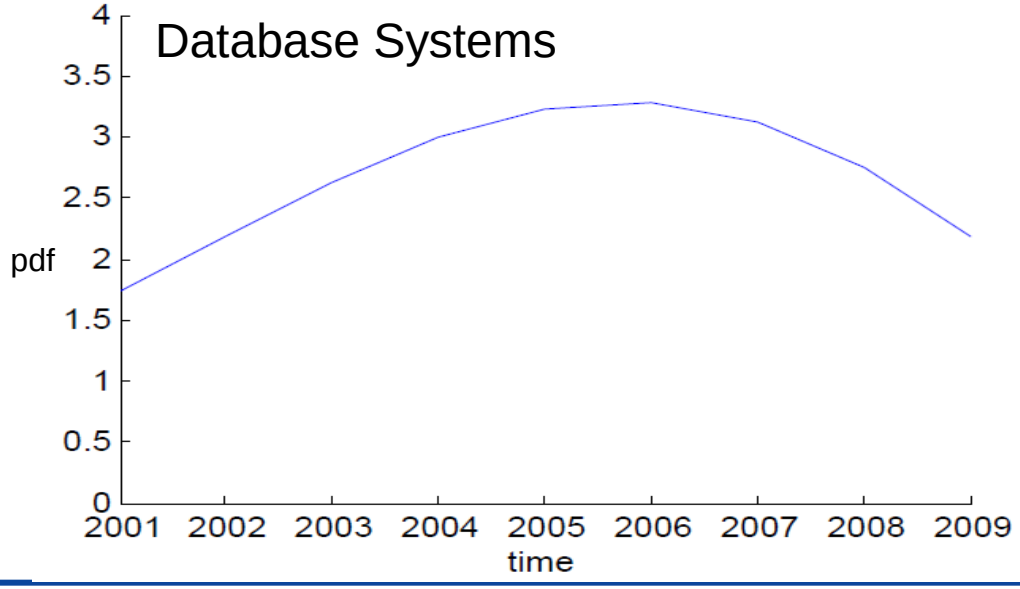
1. http://www-ice.iarc.fr/~martyn/software/jags/

Topics Life Cycle and Associated Authors (CiteSeer dataset)

Semantic Web	
Word	**Prob.**
ontolog	0.0380
web	0.0375
metadata	0.0298
rdf	0.0211
semantic	0.0195
Author	**Prob.**
I. Horrocks	0.0892
S. Staab	0.0686
A. Lin	0.0042
W. Nejdl	0.0011



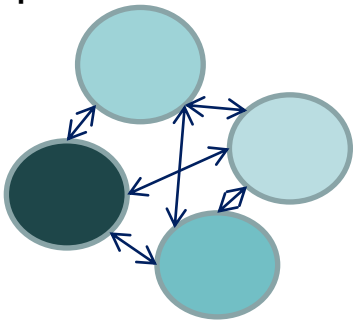
Database Systems	
Word	**Prob.**
queri	0.0226
databas	0.0184
knowledg	0.0179
cube	0.0160
logic	0.0146
Author	**Prob.**
D. Agrawal	0.0927
I. Horrocks	0.0610
A. Joshi	0.0378
W. Nejdl	0.0328

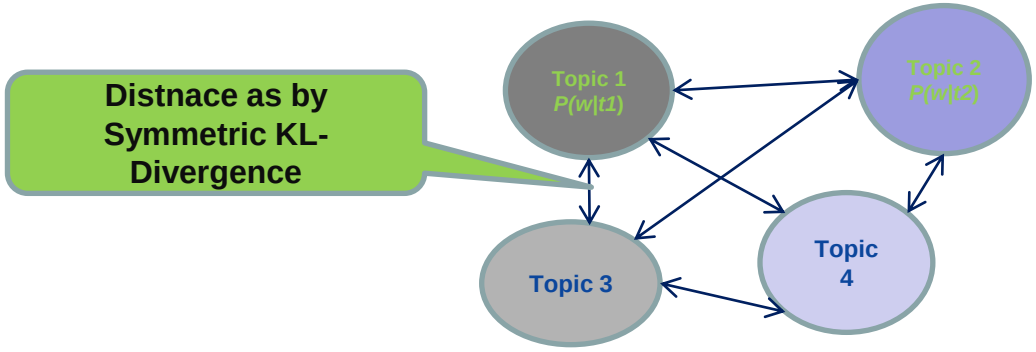


Evaluation

- Quality of topics:
  - Subjectively
  - Measure the distance between word distributions $P(w/t)$ for a pair of topics.

Model	Average Symm. KL Divergence
ATT	14.59341783
LDA	8.452413348





Topics detected by LDA

Topic 4		Topic 9		Topic 11	
Word	Prob	Word	Prob	Word	Prob
peer	0.01996	ontolog	0.02465	ontolog	0.02163
queri	0.01568	logic	0.02178	learn	0.01378
databas	0.01426	role	0.01756	logic	0.01064
metadata	0.01141	knowledg	0.01705	delay	0.00977
resourc	0.00894	languag	0.01587	tree	0.00959
view	0.00894	descript	0.01503	reason	0.00942
grid	0.00843	web	0.01503	function	0.00924
search	0.00804	reason	0.01435	power	0.0089

Topics detected by ATT

Topic 9		Topic 26		Topic 79	
Word	Prob.	Word	Prob.	Word	Prob.
storag	0.028	ontolog	0.04307	file	0.0285
disk	0.0235	web	0.03784	metadata	0.02296
failur	0.019	semant	0.02898	secur	0.01979
reliabl	0.0175	languag	0.02254	analysi	0.019
select	0.016	rdf	0.01852	safeti	0.01504
server	0.015	knowledg	0.0153	share	0.01425
fault	0.015	schema	0.0149	storag	0.01267
cach	0.013	servic	0.01248	express	0.01267

Author Similarity (Symmetric KL-Divergence)

Inter-Topic Authors	
Author Pair	**Symm. KL Divergence**
D. Agrawal[4] - L. Tong[1]	9.07
I. Horrocks[4] - X. Gu[2]	8.69
S. Staab[3] - X.Gu[2]	8.64
L. Tong[1] - X. Gu[2]	7.64

Intra-Topic Authors	
Author Pair	**Symm. KL Divergence**
D. Agrawal[4] - I. Horrocks[4]	2.12
S. Staab[3] - I. Horrocks[3]	1.88
C. Chan[2] - H. Lin[2]	2.19
H. Lin[2] - J. Gao[2]	2.06

Key	Topic
1	Grid Computing
2	Image Analysis
3	Semanic Web
4	Database Systems

Summary and Future Work

Summary :

- We described the problem of identifying author's interests over time wrt evolving topics
- Proposed the ATT model for capturing author, topic and time dependencies
- Results from the application of ATT to CiteSeer dataset

Future Work:

- Implementing Gibbs Sampler specific to ATT for scaling to larger dataset
- Evaluation of the ATT in Recommender Scenario
- Make better use of the time information for querying the model

Thank You