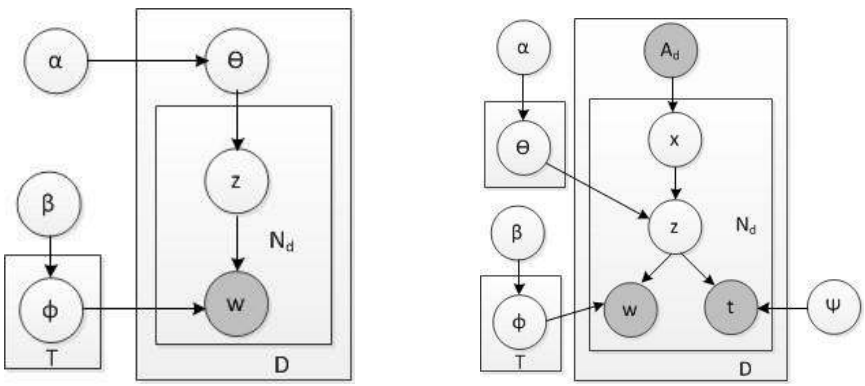
are semantically related to each other and a human subject is able to say that “these words are about X“, where X can be any domain like business, computer science, chemistry etc. There is no consensus in literature on what could be a formal definition of a topic model. So, we see a “Topic Model“ as a model of the generative process by which documents are created and captures the word co-occurrence patterns in a document corpus to produce semantically coherent topics. A variety of statistical models have been proposed for topic-based analysis and modeling of text documents. To name few of them are unigram model, mixture of unigram model [Nigam et al., 2000], latent semantic analysis (pLSA) (Hofmaan, 2001) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is a Bayesian multinomial mixture model which has become a state of the art and popular method in text analysis due to its ability to produce interpretable and semantically coherent topics. It uses the Dirichlet distribution to model the distribution of the topics for each document. In LDA each word is considered sampled from a multinomial distribution over words specific to this topic. LDA is a well-defined generative model and generalizes easily to new documents without overfitting. Since LDA is highly modular and hierarchical, therefore, it can easily be extended. Many extensions to basic LDA model have been proposed to incorporate document metadata. The simplest method of incorporating the metadata in generative topic models is to generate both the words and the metadata simultaneously given hidden topic variables. In this type of model, each topic has a distribution over words as in the standard model, as well as a distribution over metadata values. Examples of such model includes, Topics over Time model (Wang et al., 2006) of Wang and McCallum, Continuous Time Dynamic Topic Models. The Group-Topic model of Wang (Wang et al., 2008), Mohanty and McCallum (Wang et al., 2005), Author-Topic model (Rosen et al., 2004) of Rosen Zvi, Griffiths, Steyvers and Smyth, Linked Topic and Interest Model (Cheng et al., 2008) of Cheng and Li.

THE ATTention MODEL

We extend LDA by incorporating document metadata i.e. author and timestamp of the document. Figure 1 is a graphical representation of LDA and ATTention. In the model each author is modeled as having distribution overs topics and each topic is modeled as having distribution over words. The ATTention model has three sets of unknown parameters; the author distribution over topics θ, the topic distribution over words φ and the topic distribution over time ψ. Both θ and φ have multinomial distributions with symmetric Dirichlet priors having the hyperparameters α and β respectively. To avoid time discretization, we use a continuous per-topic parametric Beta distribution ψ over absolute time values in the generative process, this gives a natural distribution of topics over time. We normalize the time-stamps to values between 0 and 1 for parameter estimation.



(a) Topic-Word (LDA) (b) ATTention Model

Figure 1: Latent Dirichlet Allocation and ATTention Models for document content generation

The generative process of the ATTention model as given in Figure 1, which corresponds to the process used in Gibbs sampling for parameter estimation is described as follows.

- Draw θ $Dir(\alpha)$
- Draw φ $Dir(\beta)$

- For each document d, pick an author from the list of authors a_d and draw a multinomial θ_d from Dirichlet prior α; then for each of the N_d words, w_i,
  - Draw a topic z_{di} from multinomial θ_d
  - Draw a word w_{di} from multinomial φ_{zdi}
  - Draw a timestamp t_{di} from Beta ψ_{zdi}

In the ATTention model three parameters θ, φ, ψ are estimated. Exact inference of the parameters of LDA type models is intractable, therefore, we use Gibbs sampling to perform approximate inference. In the model there are three latent variables z, a and t. Each set (zi, ai, ti) of these latent variables is drawn as block conditioned on all other variables. We begin with the joint probability of dataset, and using the chain rule we obtain conditional probability for

$$p(z_i = j, x_i = k, t_i = l | w_i = m, z_{-i}, x_{-i}, t_{-i}, w_{-i}, a_d) \quad (1)$$

where zi, xi, ti represent topic, author and time assigned to wi whereas z_{-i}, x_{-i}, t_i are all other assignments of that topic, author and time excluding the current assignment. w_{-i} represents all other words in the document set and ad is the observed author of the document. Learning joint probabilities of these three latent variables enables us to query the model conditioned on any combination of these variables using Baye's rule. For example, given the author and time find the authors interest in that time period P (φ_d a, t) or given the topic and time find the top authors contributing to the topic in that time P (θ_d z, t). The presented approach can be used for variety of applications. For example, authors that have high probability for a topic when it starts emerging can be seen as "topic pioneers" who conduct innovative research in that topic. Moreover, active authors that frequently change their topics of interest can be considered as "trend setters" in the respective research community. On the other hand, authors that have high probability at the peak topic activity can be seen as "mainstream" researchers that follow general trends and interests of the community. Finally, authors that have time-independent profiles with stable topics of interest can be recognized as foundational researchers that act independently of fluctuating trends and popular issues. From the application perspective, this knowledge can be exploited in a variety of ways, e.g. for advanced impact ranking, similarity-based contact recommendation for future collaborations, or better summarization of recent research trends and prediction of their further evolution.

RESULTS AND DISCUSSION

We apply our approach to a subset of the CiteSeer dataset consisting of abstracts and titles of research papers published by authors having more than 150 publications from 2001 to 2009. The minimum limit of 150 publications is applied to have sufficient text for capturing author interest over time. Dataset is preprocessed to remove stop words and noise be removing highly frequent terms and terms occurring in less than 10 documents. We set the number of topics to K=100 and fix the hyperparameters α = 50/K and β = 0.01. The results shown are obtained by sampling from the 2000th iteration of Gibbs Sampler. Due to space constraints we are showing four topics with their most influential authors and beta PDF modeling the topic distribution over time. Table1 to Table 4 show the top 5 terms and the top 4 authors for each topic. The interesting observation from the results is that the activities in the Semantic Web and Database System topics are correlated. As one topic starts gaining, the activity in another topic starts decreasing. It also shows that as the topic of semantic web started to emerge, influential authors in the database systems topic shifted to semantic web topic.

CONCLUSION

In this paper, we have presented a probabilistic method of modeling text, authors and timestamps in a given set of text data enabling us to identify temporal activity of topics and finding out influential users for the identified topics. Joint modeling and learning posterior probabilities of text, author and time allows us to query model for any combination of these variables conditioned on each other for finding information about how author's interests change over time and how activity in topics changes with emergence of new topics. Results from the application of this model to the CiteSeer dataset show the applicability of the model to arbitrary document collections with author and temporal information for detecting topics trends, topic evolution and author's interests.

Table 1: Grid Computing topic captured by ATTention model with influential authors and beta PDF showing topic activity

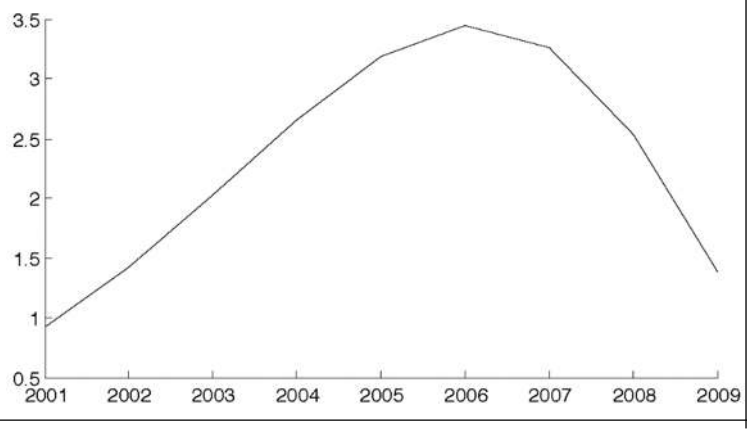
Grid Computing		
Word	**Prob.**	**Topic Activity**
grid	0.0297	
framework	0.019	
dynamic	0.0175	
resource	0.0175	
integration	0.0131	
Author	**Prob.**	
L. Tong	0.0746	
E. Gold	0.0207	
W. Zhao	0.0186	
H. Lin	0.0134	

Table 2: Image Analysis topic captured by ATTention model with influential authors and beta PDF showing topic activity

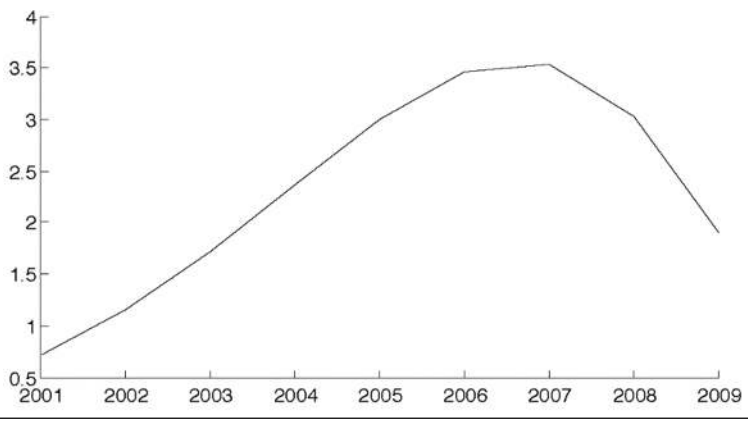
Image Analysis		
Word	**Prob.**	**Topic Activity**
image	0.0185	
spectral	0.015	
test	0.013	
cluster	0.012	
statistic	0.012	
Author	**Prob.**	
X. Gu	0.1201	
C. Chan	0.0047	
H. Lin	0.0025	
J. Gao	0.0013	

Table 3: Semantic Web topic captured by ATTention model with influential authors and beta PDF showing topic activity

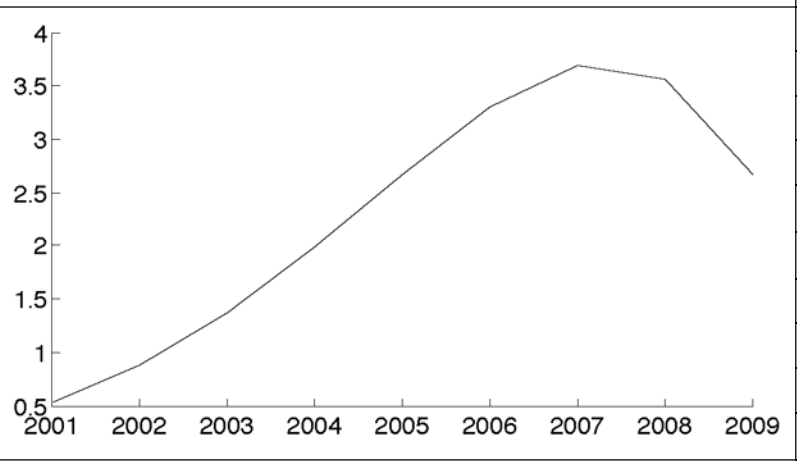
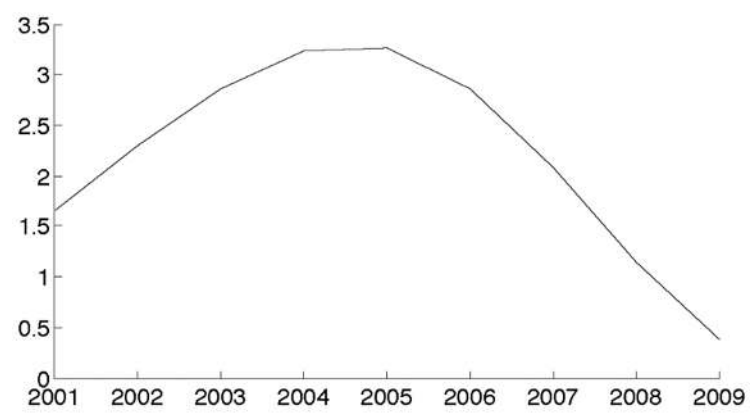
Semantic Web		
Word	**Prob.**	**Topic Activity**
resource	0.038	
web	0.0375	
metadata	0.0298	
rdf	0.0211	
semantic	0.0195	
Author	**Prob.**	
I. Horrocks	0.0892	
S. Staab	0.0686	
A. Lin	0.0042	
W. Nejdl	0.0011	

Table 4: Database Systems topic captured by ATTention model with influential authors and beta PDF showing topic activity

Database Systems		
Word	**Prob.**	**Topic Activity**
database	0.0223	
query	0.019	
sequence	0.0114	
control	0.01	
search	0.0095	
Author	**Prob.**	
S. Staab	0.1	
I. Horrocks	0.051	
A. Joshi	0.0478	
W. Nejdl	0.0328	

REFERENCES

- Blei DM, Ng AY, Jordan MI. (2003). Latent Dirichlet Allocation. J. Mach. Learn. Res. 3:993– 1022.
- Cheng V, Li CH. (2008). Linked topic and interest model for web forums. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 1: 279–284. IEEE Computer Society, Washington, DC, USA.

- Rosen-Zvi M, Griffiths T. Steyvers M, Smyth P. (2004). The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. UAI '04, AUAI Press, Arlington, Virginia, United States, pp. 487–494.

- Wang X, McCallum A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '06, ACM, New York, NY, USA, pp. 424–433.

- Wang X, Mohanty N, McCallum A. (2005). Group and topic discovery from relations and text. In: Proceedings of the 3rd international workshop on Link discovery. LinkKDD '05, ACM, New York, NY, USA, pp. 28–35.

- Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using em. Mach. Learn 39(2-3):103–134, 2000.

- Wang C, Blei DM, Heckerman D. (2008). Continuous time dynamic topic models. In UAI'08, 579–586

- Wang X, Mohanty N, McCallum A. (2005). Group and topic discovery from relations and text. In Proceedings of the 3rd international workshop on Link discovery, LinkKDD 05, New York, NY, USA, ACM, pp28–35,