A Continuous-Time Model of Topic Co-occurrence Trends

Wei Li, Xuerui Wang and Andrew McCallum
Department of Computer Science
University of Massachusetts
140 Governors Drive
Amherst, MA 01003-9264

Abstract

Recent work in statistical topic models has investigated richer structures to capture either temporal or inter-topic correlations. This paper introduces a topic model that combines the advantages of two recently proposed models: (1) The Pachinko Allocation model (PAM), which captures arbitrary topic correlations with a directed acyclic graph (DAG), and (2) the Topics over Time model (TOT), which captures time-localized shifts in topic prevalence with a continuous distribution over timestamps. Our model can thus capture not only temporal patterns in individual topics, but also the temporal patterns in their co-occurrences. We present results on a research paper corpus, showing interesting correlations among topics and their changes over time.

Introduction

The increasing amount of data on the Web has heightened demand for methods that extract structured information from unstructured text. A recent study shows that there were over 11.5 billion pages on the world wide web as of January 2005 (Gulli & Signorini 2005), mostly in unstructured data. Lately, studies in information extraction and synthesis have led to more advanced techniques to automatically organize raw data into structured forms. For example, the Automatic Content Extraction (ACE) program[1] has added a new Event Detection and Characterization (EDC) task. Compared to entity and relation extraction, EDC extracts more complex structures that involve many named entities and multi-relations.

While ACE-style techniques provide detailed structures, they are also brittle. State-of-the-art named entity recognition and relation extraction techniques are both errorful, and when combined, yield even lower accuracies for event extraction. Furthermore, they require significant amount of labeled training data. In this paper, we explore an alternative approach that models topical events. Instead of extracting individual entities and relations for specific events, we capture the rise and fall of themes, ideas and their co-occurrence patterns over time. Similar to Topic Detection and Tracking (TDT) (Allan *et al.* 1998), which extracts related contents from a stream of newswire stories, we discover topics from documents and study their changes over time. While this approach is less detailed and structured, it is more robust and does not require supervised training. In addition to providing a coarse-grained view of events in time, these models could also be used as a filter, or as an input to later, more detailed traditional processing methods.

Statistical topic models such as Latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan 2003) have been shown to be effective tools in topic extraction and analysis. Although these models do not capture the dynamic properties of topics, many of the large data sets to which they are applied are often collected over time. Topics rise and fall in prominence; they split apart; they merge to form new topics; words change their correlations. More importantly, topic co-occurrences also change significantly over time, and time-sensitive patterns can be learned from them as well.

Some previous work has performed post-hoc analysis to study the dynamical behaviors of topics—discovering topics without the use of timestamps and then projecting their occurrence counts into discretized time (Griffiths & Steyvers 2004)— but this misses the opportunity for time to improve topic discovery. A more systematic approach is the state transition based methods (Blei & Lafferty 2006) using the Markov assumption. Recently, we proposed a simple new topics over time (TOT) model to use temporal information in topic models (Wang & McCallum 2006) which represents timestamps of documents as observed continuous variables. A significant difference between TOT and previous work with similar goals is that TOT does not discretize time and does not make Markov assumptions over state transitions in time. Each topic is associated with a continuous distribution over time, and topics are responsible for generating both observed timestamps as well as words.

To generate the words in a document, TOT follows the same procedure as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan 2003). Each document is represented as a mixture of topics and each topic is a multinomial distribution over a word vocabulary. The mixture components in the documents are sampled from a single Dirichlet distribution. Therefore, TOT focuses on modeling individual topics and their changes over time.

However, in real-world text data, topics are often correlated. That is, some topics are more likely to co-occur than

[1] http://www.itl.nist.gov/iad/894.01/tests/ace/

others. In order to model this correlation, we have recently inrtoduced the Pachinko Allocation model (PAM), an extension to LDA that assumes a directed acyclic graph (DAG) to capture arbitrary topic correlations. Each leaf in the DAG is associated with a word in the vocabulary, and each interior node represents a correlation among its children, which are either leaves or other interior nodes. Therefore, PAM captures not only correlations among words like LDA, and also correlations among topic themselves. For each topic, PAM parameterizes a Dirichlet distribution over its children.

Although PAM captures topic correlations, and TOT captures the distribution of topics over time, neither captures the phenomena that the correlation among topics evolves over time. In this paper, we combine PAM and TOT to model the temporal aspects and dynamic correlations of extracted topics from large text collections. To generate a document, we first draw a multinomial distribution from each Dirichlet associated with the topics. Then for each word in the document, we sample a topic path in the DAG based on these multinomials. Each topic on the path generates a timestamp for this word based on a per-topic Beta distribution over time.

With this combined approach, we can discover not only how topics are correlated, but also when such correlations occur or disappear. In contrast to other work that models trajectories of individual topics over time, PAMTOT topics and their meanings are modeled as constant. PAMTOT captures the changes in topic co-occurrences, not the changes in the word distribution of each topic.

The Model

In this section, we first present the Pachinko allocation model (PAM) which captures individual topics and their correlations from a static point of view. Then we introduce a combined approach of PAM and topics over time (TOT) for modeling the evolution of topics. We are especially interested in the timelines of topic correlations. While PAM allows arbitrary DAGs, we will focus on a special four-level hierarchical structure and describe the corresponding inference algorithm and parameter estimation method.

Pachinko Allocation Model

The notation for the Pachinko allocation model is summarized below.

$V = \{x_1, x_2, ..., x_v\}$: a word vocabulary.

$S = \{y_1, y_2, ..., y_s\}$: a set of topic nodes. Each of them captures some correlation among words or topics. Note that there is a special node called the root r. It has no incoming links and every topic path starts from it.

D: a DAG that consists of nodes in V and S. The topic nodes occupy the interior levels and the leaves are words.

$G = \{g_1(\alpha_1), g_2(\alpha_2), ..., g_s(\alpha_s)\}$: g_i, parameterized by α_i, is a Dirichlet distribution associated with topic y_i. α_i is a vector with the same dimension as the number of children in y_i, specifying the correlation among them.

To generate a document d, we follow a two-step process:

1. Sample $\theta^{(d)}_{y_1}$, $\theta^{(d)}_{y_2}$, ..., $\theta^{(d)}_{y_s}$ from $g_1(\alpha_1)$, $g_2(\alpha_2)$, ..., $g_s(\alpha_s)$, where $\theta^{(d)}_{y_i}$ is a multinomial distribution of topic y_i over its children.
2. For each word w in the document,
   - Sample a topic path $\mathbf{z}_w$ of L_w topics : $< z_{w1}, z_{w2}, ..., z_{wL_w} >$. z_{w1} is always the root and z_{w2} through z_{wL_w} are topic nodes in S. z_{wi} is a child of $z_{w(i-1)}$ and it is sampled from the multinomial distribution $\theta^{(d)}_{z_{w(i-1)}}$.
   - Sample word w from $\theta^{(d)}_{z_{wL_w}}$.

Following this process, the joint probability of generating a document d, the topic assignments $\mathbf{z}^{(d)}$ and the multinomial distributions $\theta^{(d)}$ is

$$P(d, \mathbf{z}^{(d)}, \theta^{(d)}|\alpha) = \prod_{i=1}^{s} P(\theta^{(d)}_{y_i}|\alpha_i) \times \prod_{w} (\prod_{i=2}^{L_w} P(z_{wi}|\theta^{(d)}_{z_{w(i-1)}}) P(w|\theta^{(d)}_{z_{wL_w}}))$$

Combining PAM and TOT

Now we introduce a combined approach of PAM and topics over time (TOT), which captures the evolution of topics and their correlations. Each document is associated with one timestamp, but for the convenience of inference (Wang & McCallum 2006), we consider it to be shared by all the words in the document. In order to generate both words and their timestamps, we modify the generative process in PAM as follows. For each word w in a document d, we still sample a topic path $\mathbf{z}_w$ based on multinomial distributions $\theta^{(d)}_{y_1}$, $\theta^{(d)}_{y_2}$, ..., $\theta^{(d)}_{y_s}$. Simultaneously, we also sample a timestamp t_{wi} from each topic z_{wi} on the path based on the corresponding Beta distribution Beta($\psi_{z_{wi}}$), where ψ_z is the parameters of the Beta distribution for topic z.

Now the joint probability of generating a document d, the topic assignments $\mathbf{z}^{(d)}$, the timestamps $\mathbf{t}^{(d)}$ and the multinomial distributions $\theta^{(d)}$ is

$$P(d, \mathbf{z}^{(d)}, \mathbf{t}^{(d)}, \theta^{(d)}|\alpha, \Psi) = \prod_{i=1}^{s} P(\theta^{(d)}_{y_i}|\alpha_i) \times \prod_{w} (\prod_{i=1}^{L_w} P(t_{wi}|\psi_{z_{wi}}) \prod_{i=2}^{L_w} P(z_{wi}|\theta^{(d)}_{z_{w(i-1)}}) P(w|\theta^{(d)}_{z_{wL_w}}))$$

Integrating out $\theta^{(d)}$ and summing over $\mathbf{z}^{(d)}$, we get the marginal probability of a document and its timestamps as:

$$P(d, \mathbf{t}^{(d)}|\alpha, \Psi) = \int \prod_{i=1}^{s} P(\theta^{(d)}_{y_i}|\alpha_i) \times \prod_{w} (\sum_{\mathbf{z}_w} \prod_{i=1}^{L_w} P(t_{wi}|\psi_{z_{wi}}) \prod_{i=2}^{L_w} P(z_{wi}|\theta^{(d)}_{z_{w(i-1)}}) P(w|\theta^{(d)}_{z_{wL_w}})) \mathrm{d}\theta^{(d)}$$

Finally, the probability of generating a corpus with timestamps is the product of the probability for every document:

$$P(\mathbf{D}, \mathbf{T}|\alpha, \Psi) = \prod_{d} P(d, \mathbf{t}^{(d)}|\alpha, \Psi)$$

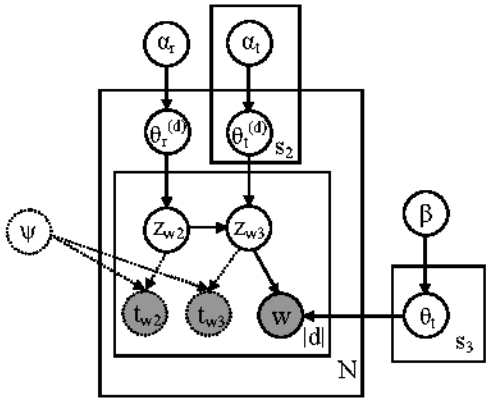

Figure 1: Graphical model for four-level PAMTOT. The dotted lines describe how to sample timestamps from topics, while the solid lines correspond to PAM without time.

Based on the above generative process, each word is associated with multiple timestamps sampled from different topics. When fitting our model, each training document's timestamp is shared by all the words in the document. But after fitting, if it actually runs as a generative model, this process would generate different timestamps for every word.

Four-Level PAMTOT

While PAM allows arbitrary DAGs to model the topic correlations, in this paper, we focus on one special structure in our experiments. It is a four-level hierarchy consisting of one root topic, s_2 topics at the second level, s_3 topics at the third level and words at the bottom. We call the topics at the second level super-topics and the ones at the third level sub-topics. The root is connected to all super-topics, super-topics are fully connected to sub-topics and sub-topics are fully connected to words. We also make a simplification similar to LDA; i.e., the multinomial distributions for sub-topics are sampled once for the whole corpus, from a single Dirichlet distribution $g(\beta)$. The multinomials for the root and super-topics are still sampled individually for each document. In comparison to LDA, this special setting of PAM has one additional layer of super-topics modeled with Dirichlet distributions, which is the key component to capturing topic correlations in this model. We present the graphical model in Figure 1. Since the root is fixed, we only show the variables for super-topics and sub-topics.

Inference and Parameter Estimation

The hidden variables in PAMTOT include the sampled multinomial distributions θ and topic assignments $\mathbf{z}$. In addition, we need to learn the parameters in the Dirichlet distributions $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_s\}$ and the Beta distributions $\Psi = \{\psi_1, \psi_2, ..., \psi_s\}$. We could employ the Expectation-Maximization (EM) algorithm for inference, which is often used to estimate parameters for models with hidden variables. However, EM has been shown to perform poorly for topic models, as they have many local maxima.

Instead, we apply Gibbs Sampling to perform approximate inference and parameter estimation. For an arbitrary DAG, we need to sample a topic path for each word given other variable assignments enumerating all possible paths and calculating their conditional probabilities. In our special four-level PAMTOT structure, each topic path contains the root, a super-topic and a sub-topic. Since the root is fixed, we only need to jointly sample the super-topic and sub-topic assignments for each word, based on their conditional probabilities given observations and other assignments, and having integrated out the multinomial distributions θ. The following equation shows the joint probability of a super-topic and a sub-topic. For word w in document d, we have:

$$P(z_{w2} = y_k, z_{w3} = y_p | \mathbf{D}, \mathbf{T}, \mathbf{z}_{-w}, \alpha, \beta, \Psi) \propto$$
$$\frac{n_{1k}^{(d)} + \alpha_{1k}}{n_1^{(d)} + \sum_{k'} \alpha_{1k'}} \times \frac{n_{kp}^{(d)} + \alpha_{kp}}{n_k^{(d)} + \sum_{p'} \alpha_{kp'}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_m \beta_m}$$
$$\times \frac{(1 - t_{w2})^{\psi_{k1}-1} t_{w2}^{\psi_{k2}-1}}{B(\psi_{k1}, \psi_{k2})} \times \frac{(1 - t_{w3})^{\psi_{p1}-1} t_{w3}^{\psi_{p2}-1}}{B(\psi_{p1}, \psi_{p2})}$$

We assume that the root topic is y_1. z_{w2} and z_{w3} correspond to super-topic and sub-topic assignments respectively. $\mathbf{z}_{-w}$ is the topic assignments for all other words. Excluding the current token, $n_i^{(d)}$ is the number of occurrences of topic y_i in document d; $n_{ij}^{(d)}$ is the number of times topic y_{ij} is sampled from its parent y_i in document d; n_i is the number of occurrences of sub-topic y_i in the whole corpus and n_{iw} is the number of occurrences of word w in sub-topic y_i. Furthermore, α_{ij} is the jth component in α_i. β_w is the component for word w in β. ψ_{i1} and ψ_{i2} are the two parameters in the Beta distribution of topic y_i. $B(a, b)$ is the Beta function.

Note that in the Gibbs sampling equation, we assume that the Dirichlet parameters α and the Beta parameters Ψ are given. Since they capture different topic correlations and different distributions over time, we cannot assume uniform distributions for them, and instead, we need to learn these parameters from the data during Gibbs sampling, e.g., using maximum likelihood or EM. However, since there are no closed-form solutions for these methods and we wish to avoid iterative methods for the sake of simplicity and speed, we estimate them by the method of moments.

Related Work

Several studies have examined topics and their changes across time. Rather than jointly modeling word co-occurrence and time, many of these methods simply use post-hoc or pre-discretized analysis (Griffiths & Steyvers 2004; Wang, Mohanty, & McCallum 2005; Song *et al.* 2005).

More recently, time series analysis rooted models have become popular, many of which are based on dynamic models, with a Markov assumption that the state at time $t + 1$ or $t + \Delta t$ is independent of all other history given the state at time t. Hidden Markov models and Kalman filters are two such examples. For instance, Blei and Lafferty (2006) present Dynamic Topic Model (DTM) in which the alignment among topics across time steps is modeled by a Kalman filter on the Gaussian distribution in the logistic normal distribution (Blei & Lafferty 2006). This approach is

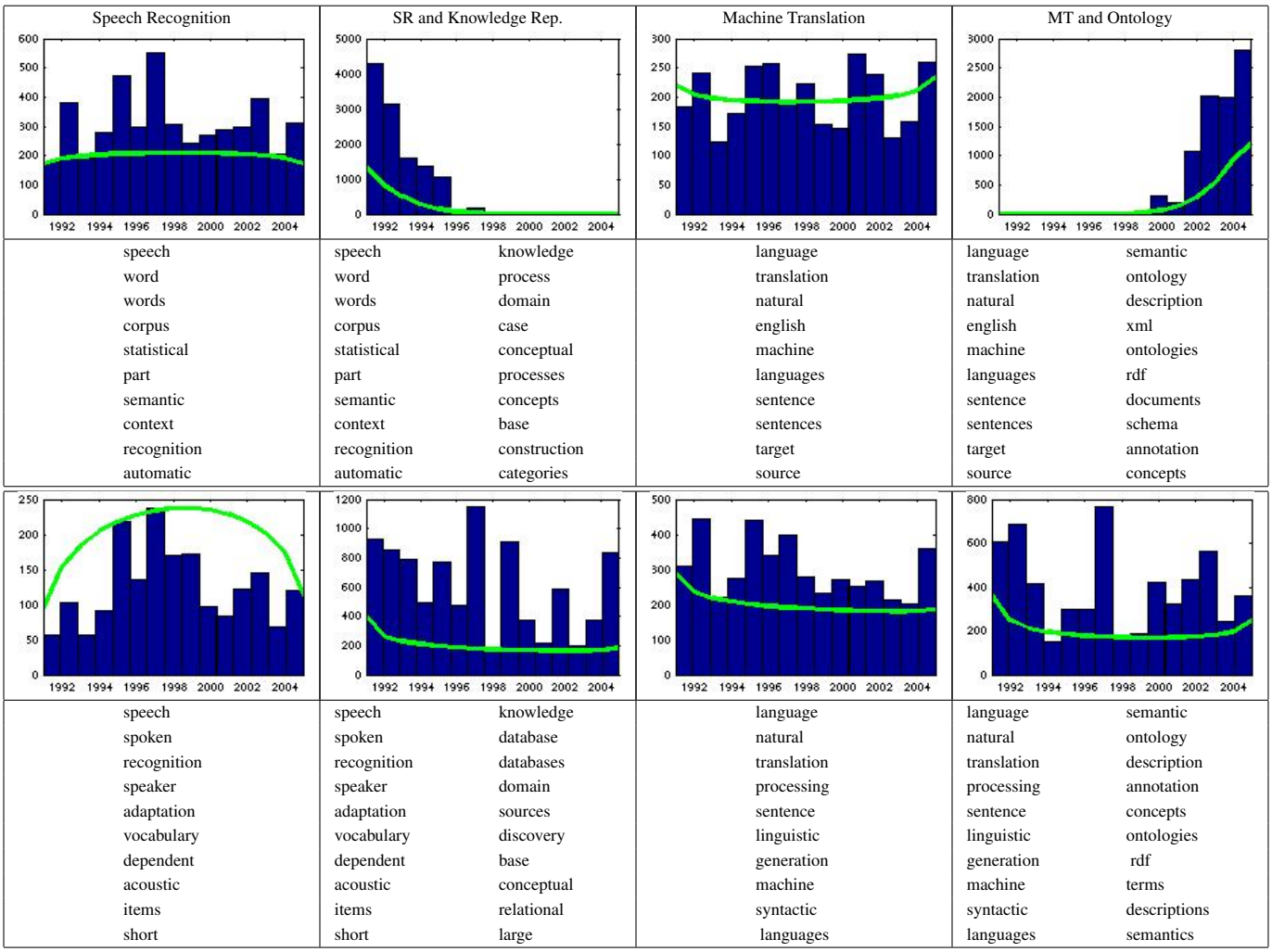

Figure 2: Two examples discovered by PAMTOT (above) and PAM (bottom) from the Rexa dataset. Each example consists of a sub-topic and a super-topic. The titles are our own interpretation of the topics. Histograms show how the topics are distributed over time; the fitted Beta PDFs is shown also. (For PAM, Beta distributions are fit in a post-hoc fashion). For sub-topics, we list the top words below the histograms. For super-topics, we list the top words for their child topics.

quite different from PAMTOT. First, it employs a Markov assumption over time; second, it is based on the view that the "meaning" (or word associations) of a topic changes over time.

Another Markov model that aims to find word patterns in time is Kleinberg's "burst of activity model" (Kleinberg 2002). This approach uses an infinite-state automaton with a particular state structure in which high activity states are reachable only by passing through lower activity states. Rather than leveraging time stamps, it operates on a stream of data, using data ordering as a proxy for time. Its infinite-state probabilistic automaton has a continuous transition scheme similar to Continuous Time Bayesian Networks (CTBNs) (Nodelman, Shelton, & Koller 2002). However, it operates only on one word at a time, whereas the PAMTOT model finds time-localized patterns in word *co-occurrences*.

Like TOT, PAMTOT uses time quite differently than the above models. First, PAMTOT does not employ a Markov assumption over time, but instead treats time as an observed continuous variable. Second, many other models take the view that the "meaning" (or word associations) of a topic changes over time; instead, in PAMTOT we can rely on topics themselves as *constant*, while topic co-occurrence patterns change over time.

Although not modeling time, several other topic models have associated the generation of additional modalities with topics. E.g., the aforementioned Group-Topic (GT) model (Wang, Mohanty, & McCallum 2005) conditions on topics for both word generation and relational links. As in TOT and PAMTOT, GT results also show that jointly modeling an additional modality improves the relevance of the discovered topics. Another flexible, related model is the Mixed Membership model (Erosheva, Fienberg, & Lafferty 2004), which treats the citations of papers as *additional* "words", thus the formed topics are influenced by both words and citations.

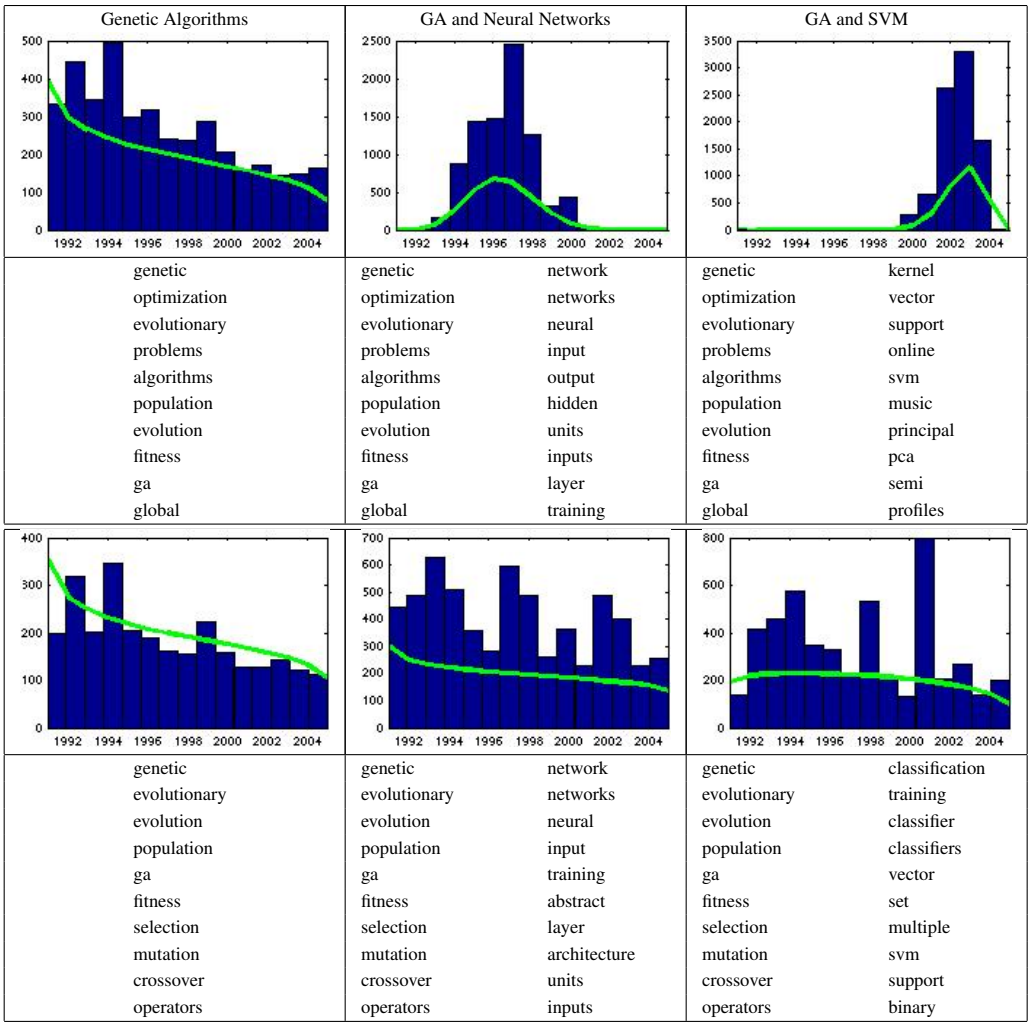

Figure 3: Another example showing a pattern discovered by PAMTOT. The first column is a sub-topic and the other two columns correspond to two parent super-topics that capture its correlations with other topics.

Experimental Results

In this section, we present example topics discovered by the PAMTOT model, focusing on the interesting patterns in the evolution of topics and their correlations.

The dataset we use in our experiments comes from Rexa, a search engine over research papers. Rexa has a large collection of papers from a wide range of research areas. We choose a subset of paper titles and abstracts that are mostly about machine learning and natural language processing. Then from this subset, we randomly draw 4454 documents spanning from the years 1991 to 2005. For each of the 15 years, there are exactly 300 documents except 1991, for which there were only 254 machine learning documents in the corpus. The overall distribution is therefore close to uniform. After down-casing and removing stopwords, we obtain a total set of 372936 tokens and 21748 unique words.

In our experiments, we use a fixed four-level hierarchical structure that includes a root, 50 super-topics, 100 sub-topics and a word vocabulary. For the root, we assume a fixed Dirichlet distribution with parameter 0.01. We can change this parameter to adjust the variance in the sampled multinomial distributions. We choose a small value so that the variance is high and each document contains only a small number of super-topics, which tends to make the super-topics more interpretable. We treat the sub-topics in the same way as LDA, i.e., assume they are sampled once for the whole corpus from a given Dirichlet with parameter 0.01. So the only parameters we need to learn are the Dirichlet parameters for the super-topics and the Beta parameters for both super-topics and sub-topics.

We show two example trends in Figure 2. Each column corresponds to one sub-topic or super-topic. The titles are our own interpretation of the topics. The Beta distributions over time and their actual histograms over time are displayed in the graphs. We also list the 10 most likely words for each sub-topic and the highly correlated children of each super-topic. As a comparison, we also show their most similar PAM topics at the bottom, decided by KL-divergence. The time analysis for PAM topics is done post-hoc.

The first column in Figure 2 demonstrates how the sub-

Table 1: Errors and accuracies of time (publishing year) predictions for PAMTOT

	L1 Error	E(L1)	Accuracy
PAMTOT	1.56	1.57	0.29
PAM	5.34	5.30	0.10

topic "Speech Recognition" changes over time. As we can see, this topic has a relatively smooth distribution between year 1991 and 2005. More interestingly, as shown in the second column, a super-topic that captures the correlation between "Speech Recognition" and another topic "Knowledge Representation" has more dramatic changes in the frequency. These two topics are well connected before 1994 and the correlation gradually ceases after 1998. Our understanding is that other techniques for speech recognition have become more popular than knowledge-based approaches. At the bottom of these two columns, we show the corresponding sub-topic and super-topic discovered by PAM without time information. While the time distribution of "Speech Recognition" remains almost the same, PAM alone cannot discover the correlation pattern between these two topics.

The two columns on the right show another example. "Machine Translation" has been a popular topic over the entire time period. On the other hand, "Ontologies" is a relatively new subject. We see increasing correlation between them from year 2000. Again, without time information, PAM does not pick up this trend clearly.

Figure 3 shows another pattern that becomes clearer when we analyze two super-topics at the same time. The first column corresponds to the sub-topic "Genetic Algorithms". Its frequency has been slowly decreasing from year 1991 to 2005. Its connection with other topics are more localized in time. As shown by the second and third columns, it co-occurs more often with "Neural Networks" around 1996, and from 2000, the focus has shifted to other topics like "Support Vector Machines". This pattern reflects the popularities of these techniques in different years. We cannot capture this trend by PAM and post-hoc analysis of time. As the graphs at the bottom show, we can only see slight decrease of the correlation between "Genetic Algorithms" and "Neural Networks" over the years, and also the connection with "Support Vector Machines" has too much noise to exhibit any interesting pattern over time.

One interesting feature of our approach (and one not shared by state-transition-based Markov models of topical shifts) is the capability of predicting the timestamp given the words in a document. This task also provides another opportunity to quantitatively compare PAMTOT against PAM.

On the Rexa dataset, we measure the ability to predict the publishing year given the text of the abstract of a paper, as measured in accuracy, L1 error (the difference between predicted and true years) and expected L1 distance to the correct year (average differences between all years and true year). As shown in Table 1, PAMTOT achieves almost triple the accuracy of PAM, and provides an L1 relative error reduction of 70%.

Conclusions

This paper has presented an approach combining the Pachinko allocation model and topics over time that jointly captures topic correlations and identifies their localization in time. We have applied this model to a large corpus of research papers and discovered interesting patterns in the evolution of topics and the connections among them. We also show improved ability to predict time given a document. Unlike some related work with similar motivations, PAMTOT does not require discretization in time or Markov assumptions on state dynamics. The relative simplicity provides advantages for future extensions to model more complex structures among not only topics, but other related information as well.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, and under contract number HR0011-06-C-0023.

References

Allan, J.; Carbonell, J.; Doddington, G.; Yamron, J.; and Yang, Y. 1998. Topic detection and tracking pilot study. final report. In *DARPA Broadcast News Transcription and Understanding Workshop*, 194–218.

Blei, D., and Lafferty, J. 2006. Dynamic topic models. In *International Conference on Machine Learning (ICML)*.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Erosheva, E.; Fienberg, S.; and Lafferty, J. 2004. Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences* 101(Suppl. 1).

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl. 1):5228–5235.

Gulli, A., and Signorini, A. 2005. The indexable web is more than 11.5 billion pages. In *The 14th Interntional World Wide Web Conference (WWW2005)*.

Kleinberg, J. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Nodelman, U.; Shelton, C.; and Koller, D. 2002. Continuous time bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 378–387.

Song, X.; Lin, C.-Y.; Tseng, B. L.; and Sun, M.-T. 2005. Modeling and predicting personal information dissemination behavior. In *The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wang, X., and McCallum, A. 2006. Topics over time: A nonmarkov continuous-time model of topical trends. In *Submitted to the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wang, X.; Mohanty, N.; and McCallum, A. 2005. Group and topic discovery from relations and text. In *SIGKDD Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-05)*, 28–35.