Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation

Levent Bolelli[1], Şeyda Ertekin[2], and C. Lee Giles[3]

[1] Google Inc., 76 9^{th} Ave., 4^{th} floor, New York, NY 10011, USA
[2] Department of Computer Science and Engineering,
Pennsylvania State University, University Park, PA, 16802, USA
[3] College of Information Sciences and Technology,
Pennsylvania State University, University Park, PA, 16802, USA

Abstract. Algorithms that enable the process of automatically mining distinct topics in document collections have become increasingly important due to their applications in many fields and the extensive growth of the number of documents in various domains. In this paper, we propose a generative model based on latent Dirichlet allocation that integrates the temporal ordering of the documents into the generative process in an iterative fashion. The document collection is divided into time segments where the discovered topics in each segment is propagated to influence the topic discovery in the subsequent time segments. Our experimental results on a collection of academic papers from CiteSeer repository show that segmented topic model can effectively detect distinct topics and their evolution over time.

1 Introduction and Related Work

Automatic identification of semantic content of documents has become increasingly important due to its effectiveness in many tasks, including information retrieval, information filtering and organization of documents collections in digital libraries. In collections where the documents do not exhibit temporal ordering, investigating a snapshot of the collection is sufficient to gather information about various topics in the collection. However, many document collections, including scientific literature, exhibit temporal relationships that can help the topic discovery process. A topic detection algorithm, thus, can utilize the temporal ordering of documents and improve the accuracy of detected topics.

Latent Dirichlet Allocation (LDA) [1] has been shown to be a highly effective unsupervised learning methodology for finding distinct topics in document collections. It is a generative process that models each document as a mixture of topics where each topic corresponds to a multinomial distribution over words. The document-topic and topic-word distributions learned by LDA describe the best topics for documents and the most descriptive words for each topic. An extension of LDA is the author-topic model (ATM) [2,3], which is based on the author-word model [4]. In ATM, a document is represented as a product of the mixture of topics of its authors, where each word is generated by the activation of one of the topics of an author of that document, but the temporal ordering is

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 776–780, 2009.

discarded. Topics over Time (TOT) [5] is an LDA-based generative process that models time jointly with word co-occurence patterns for topic discovery. Blei and Lafferty [6] capture topic dynamics through defining an iterative process that learns the mixture parameters of topics for each time slice in the collection, and propagates the topic distributions to the next iteration by evolving the distributions with Gaussian noise. Both algorithms discard the authorship information of documents, which has been shown to be an effective ingredient for topic discovery in document collections [7].

In this paper, we propose Segmented Author-Topic Model (S-ATM), a generative model of documents that utilizes the temporal ordering of documents to improve topic discovery process. S-ATM is based on the Author-Topic Model and extends it to integrate the temporal characteristics of the document collection into the generative process.

2 Segmented Author-Topic Model for Text Collections

In S-ATM, each topic has a multinomial distribution over words and each author has a multinomial distribution over topics. A document with multiple authors has a distribution over topics that is a mixture of the topic distributions of authors. For each word w in document d, an author of d is chosen uniformly from the set of authors a_d of the document, and a word is generated through sampling a topic from the multinomial distribution of the chosen author over all topics. In the model, author-topic distributions θ have a symmetric Dirichlet prior with a hyperparameter α and word distributions of topics ϕ have a symmetric Dirichlet prior with a hyperparameter β. The generative process in S-ATM is conceptually similar to ATM, extending it to maintain a "memory" of learned distributions from past observations and utilizing θ and ϕ distributions from earlier iterations as prior knowledge for subsequent iterations. For each word w_i, the topic z_i and the author x_i responsible for that word are assigned based on the posterior probability conditioned on all other variables: $P(z_i, x_i | w_i, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}, \mathbf{a_d})$. z_i and x_i denote the topic and author assigned to w_i, while $\mathbf{z_{-i}}$ and $\mathbf{x_{-i}}$ are all other assignments of that topic and author, excluding their current assigment. $\mathbf{w_{-i}}$ represents other observed words in the document set and $\mathbf{a_d}$ is the observed author set for the document.

In order to estimate the model parameters, we use Gibbs sampling, which approximates the joint distribution of multiple variables by drawing a sequence of samples. A key issue in using Gibbs sampling for distribution approximation is the evaluation of the conditional posterior probability. That is, given T topics and V words, $P(z_i, x_i | w_i, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}, \mathbf{a_d})$ is estimated by:

$$P(z_i = j, x_i = k | w_i = m, \mathbf{z_{-i}}, \mathbf{x_{-i}}, \mathbf{w_{-i}}, \mathbf{a_d}) \propto \tag{1}$$

$$P(w_i = m | x_i = k) P(x_i = k | z_i = j) \propto \tag{2}$$

$$\frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha} \tag{3}$$

where $m' \neq m$ and $j' \neq j$, α and β are prior parameters for topic and word Dirichlets, C^{WT}_{mj} represents the number of times that word $w_i = m$ is assigned to topic $z_i = j$, C^{AT}_{kj} represents the number of times that author $x_i = k$ is assigned to topic j. The transformation from Eq. 1 to Eq. 2 drops the variables $\mathbf{z_{-i}}$, $\mathbf{x_{-i}}$, $\mathbf{w_{-i}}$ and $\mathbf{a_d}$, making the assumption that each instance of w_i is independent for simplicity. For any sample from this Markov chain, we can then estimate $P(w_i = m|z_i = r)$ and $P(z_i = r|x_i = q)$ from the topic-word distribution ϕ and author-topic distribution θ, respectively:

$$P(w_i = m|z_i = r) \propto \frac{C^{WT}_{mr} + \beta}{\sum_{m'} C^{WT}_{m'r} + V\beta} \tag{4}$$

$$P(z_i = r|x_i = q) \propto \frac{C^{AT}_{rq} + \alpha}{\sum_{r'} C^{AT}_{r'q} + T\alpha} \tag{5}$$

The iteration at time $\mathfrak{t}_0$ starts with random initialization of author-topic assignments C^{AT} and topic-word assignments C^{WT} which, at the end of the training, yields us the author-topic distributions $\theta^{\mathfrak{t}_0}$ and and topic-word distributions $\phi^{\mathfrak{t}_0}$. Each subsequent iteration then utilizes the distributions obtained in the previous iterations to initialize the assignments for the current time segment. That is, initialization of author-topic assignments for a time segment $\mathfrak{t}_k, k > 0$ becomes

$$C^{AT}_{rq,\mathfrak{t}_k} = \lambda\mathfrak{R}(C^{AT}) + (1-\lambda)\sum_{i=0}^{k-1}(\frac{1}{2})^{k-i}\theta^{t_i}_{rq} \tag{6}$$

Similarly, the initialization of the topic-word assignments is computed as

$$C^{WT}_{mr,\mathfrak{t}_k} = \lambda\mathfrak{R}(C^{WT}) + (1-\lambda)\sum_{i=0}^{k-1}(\frac{1}{2})^{k-i}\phi^{t_i}_{mr} \tag{7}$$

where $\mathfrak{R}(\cdot)$ adds random noise to the initialization by assigning topics to authors in Eq. 6 and words to topics in Eq. 7 independent of the prior knowledge obtained from $(\theta^0, \theta^1, \cdots, \theta^{k-1})$ and $(\phi^0, \phi^1, \cdots, \phi^{k-1})$, respectively. The initialization places higher emphasis to recent distributions than earlier ones through the decay component $(\frac{1}{2})^{k-i}$. This enables the learner to integrate all prior knowledge it has gathered so far with varying levels of confidence based on the influence that they may have, based on the temporal distance of the distribution to the current time segment. Since we train the model on each time segment while propagating knowledge from previous segments, the distributions $\theta^{\mathfrak{t}_k}$ and $\phi^{\mathfrak{t}_k}$ only contain the topic-probabilities of authors and topic probabilities of words seen so far. Hence, at the start of the initialization of a new segment $\mathfrak{t}_{k+1}$, the model may find a new author a', or a new word w', in which case the distributions $\theta^{\mathfrak{t}_i}_{a'm}$ and $\phi^{\mathfrak{t}_i}_{mw'}$, $i = [0, \cdots, k]$ $m = [1, \cdots, T]$ will be zero. This is a realistic representation of the corpus and denotes that we don't have prior knowledge for that particular author or word at that time segment.

3 Experiments on the CiteSeer Collection

The application of S-ATM to CiteSeer dataset provides insight into the distinct topics in the collection, the most influential authors for those topics and the popularity trends of the topics over time. Our collection contains a subset of papers from the CiteSeer repository published between 1990 and 2004 in ACM conferences. There are a total of 41,540 documents published by 35,314 authors. We used the title, abstract and keywords fields from the documents and preprocessed the text by removing all punctuation and stop words, yielding a vocabulary size of 25,101 distinct words. Due to space constraints, we show two example topics that are learned by S-ATM for the CiteSeer dataset in Table 1. The topics are extracted from a single sample at the 1000th iteration of the Gibbs sampler with a model distribution propagation parameter $\lambda = 0.5$. For the topics, we provide the top 5 topical words most likely to be generated conditioned on the topic, and the top 5 most likely authors to have generated a word conditioned on the topic, at the beginning and end of the 15 year period.

We also show the popularity trends of sample topics discovered by S-ATM in Figure 1. The popularity of topics are calculated by the fraction of words assigned to each topic for a year for all topics and for each year from 1990 to 2004. It can be seen that the popularity of machine learning topic has been steadily increasing over those years and the popularity of Digital Library and Processor Architectures topics have been stabilizing. On the other hand, the topics Programming Languages and Operating Systems have been declining in popularity in our dataset, which also agrees with the findings in [2], where our results show a more smooth decline for the popularity of these topics. One of the reasons might be attributed to the fact that the knowledge propagation in S-ATM causes the model to be less sensitive to the minor fluctuations in the topic popularities at each year and presents a smoothed topic trend analysis.

Table 1. Evolution of Two Sample Topics for S-ATM

Topic 1				**Topic 2**			
1990		**2004**		**1990**		**2004**	
memory	.1125	dynamic	.0809	graph	.1494	networks	.1220
random	.0719	memory	.0799	process	.0987	search	.0894
disk	.0654	access	.0677	routing	.0691	graph	.0848
access	.0636	random	.0463	architecture	.0668	routing	.0754
consistency	.0501	low	.0379	computation	.0485	process	.0677
Author	**Prob.**	**Author**	**Prob.**	**Author**	**Prob.**	**Author**	**Prob.**
Patterson_D	.0403	Kandemir_M	.0227	Kaiser_G	.0319	Wang_J	.0421
Chen_P	.0381	Dubois_M	.0188	Perry_D	.0271	Sen_S	.0418
Soffa_M	.0247	Jouppi_N	.0181	Gupta_R	.0188	Morris_R	.0263
Gibson_G	.0235	Pande_S	.0170	Gupta_A	.0167	Estrin_D	.0198
Reed_D	.0215	Zhuang_X	.0134	Rothberg_E	.0161	Liu_J	.0192

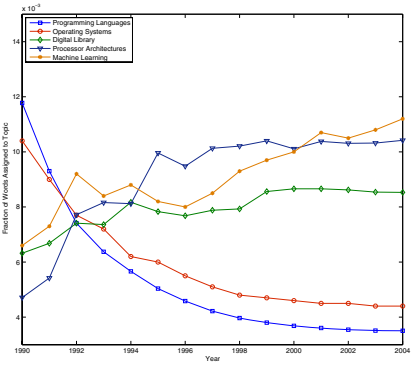

Fig. 1. Topic trends for five research topics discovered in CiteSeer collection

4 Conclusions

Many real-world text collections exhibit temporal relationships where the temporal aspects of these collections present valuable insight into the topical structure of the collections. Temporal topic discovery requires an understanding of the characteristics of the data based on the temporal evolution of the topics in the collection. In this paper, we present S-ATM, a generative model of documents that iteratively learns author-topic and topic-word distributions for scientific publications while integrating the temporal order of the documents into the generative process. The application of S-ATM to a sample dataset from CiteSeer repository shows that we can effectively discover scientific topics and most influential authors for the topics, as well as the evolution of topics over time.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: KDD 2004, pp. 306–315 (2004)
3. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI 2004, pp. 487–494 (2004)
4. McCallum, A.: Multi-label text classification with a mixture model trained by EM. In: AAAI Workshop on Text Learning (1999)
5. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: KDD 2006, pp. 424–433. ACM Press, New York (2006)
6. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML 2006: Proceedings of the 23rd international conference on Machine learning, pp. 113–120. ACM Press, New York (2006)
7. Bolelli, L., Ertekin, S., Zhou, D., Giles, C.L.: A clustering method for web data with multi-type interrelated components. In: WWW 2007, pp. 1121–1122. ACM Press, New York (2007)