The Role of Semantic History on Online Generative Topic Modeling

Loulwah AlSumait, Daniel Barbará, Carlotta Domeniconi
Department of Computer Science
George Mason University
Fairfax - VA, USA
lalsumai@gmu.edu, dbarbara@gmu.edu, carlotta@cs.gmu.edu

Abstract

Online processing of text streams is an essential task of many genuine applications. The objective is to identify the underlying structure of evolving themes in the incoming streams online at the time of their arrival. As many topics tend to reappear consistently in text streams, incorporating semantics that were discovered in previous streams would eventually enhance the identification and description of topics in the future. Latent Dirichlet Allocation (LDA) topic model is a probabilistic technique that has been successfully used to automatically extract the topical or semantic content of documents. In this paper, we investigate the role of past semantics in estimating future topics under the framework of LDA topic modeling, based on the online version implemented in [1]. The idea is to construct the current model based on information propagated from topic models that fall within a "*sliding history window*". Then, this model is incrementally updated according to the information inferred from the new stream of data with no need to access previous data. Since the proposed approach is totally unsupervised and data-driven, we analyze the effect of different factors that are involved in this model, including the window size, history weight, and equal/decaying history contribution. The proposed approach is evaluated using benchmark datasets. Our experiments show that the embedded semantics from the past improved the quality of the document modeling. We also found that the role of history varies according to the domain and nature of text data.

1 Introduction

The huge advancement in databases and the explosion of the internet, intranet, and digital libraries have resulted in giant text databases. It is estimated that approximately 85% of world-wide data is held in unstructured formats with an increasing rate of roughly 7 million digital pages per day [9]. Within their domain, these electronic documents are not static. As they become available in streams over time, text documents are dynamic and interact among each other. Thus, their contents contain a strong temporal ordering that is essential to better understand the underlying structure of the text data and its evolution over time.

Moreover, there is a great demand from genuine applications (for example newswires and security organizations for crime detection and prevention) to analyze, summarize, and categorize text streams online at the time of their arrival. Among many approaches to solve this problem, Latent Dirichlet Allocation topic modeling (LDA) [6] was successfully implemented to process text streams and identify the underlying themes and their drifts over time [1, 5, 13, 18]. LDA is a statistical generative model that relates documents and words through latent variables which represent the topics [6]. OLDA is an online version of LDA that incrementally builds an up-to-date model (mixture of topics per document and mixture of words per topic) when a new document (or a set of documents) appears [1].

When a topic is observed at a certain time, it is more likely to appear in the future with a similar distribution over words. Unlike general data mining techniques, such assumption is trivial in the area of text mining. It is widely acceptable, for instance, to consider the documents and the words in the documents to be statistically dependent. A word that occurs in a document has a higher probability to appear again in the same document and in other documents of the same topic. Consequently, a similar implication can be made about the topic distribution over time. Despite their natural drifts, the underlying themes of any domain are, in general, consistent. Hence, the semantics that are discovered at a certain point of time hold important information about the underlying structure of data in general. Incorporating such information in future knowledge discovery can improve the learning process and, eventually, enhance the inferred topics.

This paper investigates the role of previously discovered topics in inferring future semantics in text streams under the framework of LDA topic modeling. The idea is to incrementally adjust the learned topics according to the dynamical changes in the data with no need to access the previously processed documents. This is achieved by utilizing a model that uses the estimated posteriors in previous time epochs to construct the parameters of the current generative topic model. The count of words in topics, resulted from running LDA at each time instance within a "*history window*", is used to construct the priors at the following time instance. To the

Table 1: Notation used in the paper

SYMBOL	DESCRIPTION
D	total number of documents
K	number of topics
V	total number of unique words
δ	size of sliding window
N_d	number of word tokens in document d
S^t	a stream of documents arriving at time t
D^t	number of documents in S^t
V^t	number of unique words in S^t
N^t	number of word tokens in S^t
w_{di}^t	the unique word associated with the i^{th} token in document d at time t
z_i^t	the topic associated with w_{di}^t
θ_d^t	the multinomial distribution of topics specific to the document d at time t
ϕ_k^t	the multinomial distribution of words specific to the topic k at time t
α_d^t	K-vector of priors for document d at time t
β_k^t	V^t-vector of priors for topic k at time t
$\mathbf{B}_k^t$	$V^t \times \delta$ evolution matrix of topic k with columns = ϕ_k^i, $i \in \{t-\delta, \cdots, t\}$
ω	δ-vector of weights of ϕ^i, $i \in \{t-\delta, \cdots, t\}$

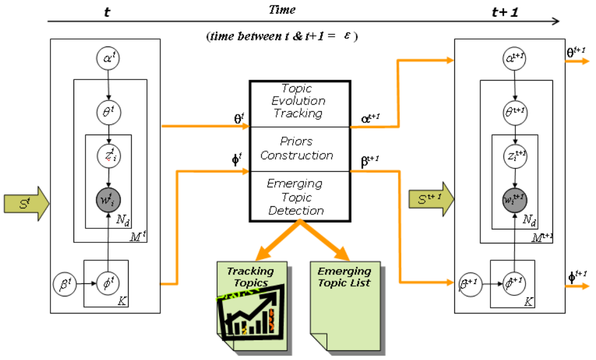

Figure 1: A flowchart of the online LDA

best of our knowledge, this is one of the first attempts to embed sematic information in the dynamics of online LDA topic modeling.

Our method is evaluated in the domain of online document modeling. An online LDA that embed semantic information from the past performed better than online LDA with no semantic embedding. In addition, with larger history windows, OLDA performs better in learning the topical content of streams, and, hence, in predicting the likelihood of unseen documents. Moreover, depending on the homogeneity and consistency of the data, balancing between the weight of embedded historic semantics and the current semantics of the newly arrived stream has an important role in the process of topic estimation.

This paper is organized as follows. Our Online LDA approach is introduced in Section 2, followed by the definition of the sliding history window and its parameters' configuration in Section 2.1. The generative process together with the algorithm is given in Section 2.2. In Section 3, we present the experiments we performed on NIPS and Reuters-21578 datasets and the results we obtained. Our final conclusions and suggestions for future work are discussed in Section 5, after giving a short review of related work (Section 4).

2 Online Latent Dirichlet Allocation.

A glossary of notations used in this paper is given in Table 2.

Probabilistic topic modeling is a relatively new approach that has been successfully applied to explore and predict the underlying structure of discrete data like text. Latent Dirichlet Allocation topic model (LDA) is a hierarchical Bayesian network that relates words and documents through latent topics. LDA is based on the assumption of *exchangeability* for the words in a document and for the documents in a corpus. Since the words are observed, the document and the topic distributions, θ and ϕ, are conditionally independent. Furthermore, the documents are not directly linked to the words. Rather, this relationship is governed by additional latent variables, z, introduced to represent the responsibility of a particular topic in using that word in the document, i.e. the topic(s) that the document is focused on. The generative process of the topic model specifies a probabilistic sampling procedure that describe how words in documents can be generated based on the hidden topics. By introducing the Dirichlet priors α and β over the document and topic distributions, respectively, the generative model of LDA is complete and generalized to process unseen documents.

Online LDA (OLDA) is an online version of the LDA model that is able to process text streams [1]. OLDA model considers the temporal ordering information and assumes that the documents are divided in time slices. At each time slice, a topic model with K components is used to model the newly arrived documents. The generated model, at a given time, is used as a prior for LDA at the successive time slice, when a new data stream is available for processing (see Figure 1 for illustration). The hyper-parameters β can be interpreted as the prior observation counts on the number of times words are sampled from a topic before any word from the corpus is observed ([15], [4]). So, the count of words in topics, resulted from running LDA on documents received at time t, can be used as the priors for the $t+1$ stream.

Formally, OLDA assumes that documents arrive in ascending order of their publication date. After each time slice, t, of a predetermined size ε, e.g. an hour, a day, or a year, a stream of documents, $S^t = \{d_1, \cdots, d_{D^t}\}$, of variable size, D^t, is received and ready to be processed. The size of the time slice, ε, depends on the nature of the corpus on which the model is applied, and on how fine or coarse the resulted description of data is expected to be. The indices of the documents within a stream, S^t, preserve the order by which the documents were received during the time slice t, i.e. d_1 is the first document to arrive and d_{D^t} is the latest document in the stream. A document d received at time t is represented as a vector of word tokens, $\mathbf{w}_d^t = \{w_{d1}^t, \cdots, w_{dN_d}^t\}$. It is naturally the case that stream S^t introduces new word(s) in the vocabulary. These words are assumed to have 0 count in ϕ for all topics in previous streams. This assumption is important to simplify the definition the topic evolution matrix and the related computation in the subsequent section.

2.1 Sliding Window of Historic Semantics Our approach allows to incorporate inferred semantics from the past data to be used to guide the inference process of the upcoming streams. This is achieved by considering all the topic-word distributions learned within a sliding "*history window*" when constructing the current priors. As a result, OLDA provides many alternatives for keeping track of history at any time t, including:

- a full memory of history which sums up all the word-topic co-occurrences from the first time instance until present.

- a short memory of history which keeps the counts of the model associated with time $t-1$ only.

Between these two ends, many intermediate options can be implemented. One approach is to set the history window parameter to some constant in which only the models that fall within this window frame are taken into consideration when constructing the priors. Another approach is to let the priors hold a decaying history information such that the contribution of history in the priors' computation declines with time. Such variety of solutions suit the structure of text repositories, since the flow and nature of document streams differ according to the type of the corpus and, hence, the role of history would be different too. In this paper, we implement these alternatives of historic semantics and investigate their effect on the performance.

By updating the priors as described above, we keep the structure of the model simple, as all the historic knowledge patterns are printed in the priors, rather than in the structure of the graphical model itself. In addition, the learning process on the new stream of data takes off from what has been learned so far, rather than starting from arbitrary settings that do not relate to the underlying distributions.

To formulate the problem, let $\mathbf{B}_k^t$ denotes an *evolutionary matrix* of topic k in which the columns are the word-topic counts ϕ_k^j, generated for streams received within the time specified by the sliding history window δ, i.e. $j \in \{t-\delta, \cdots, t\}$. Let ω be a vector of δ weights each of which is associated with a time slice from the past to determine its contribution in computing the priors for stream S^{t+1}. Hence, the parameters of a topic k at time $t+1$ are determined by a weighted mixture of the topic's past distributions as follows:

$$\beta_k^{t+1} = \mathbf{B}_k^t \omega \tag{2.1}$$

Under this definition of β, topic distributions in consecutive models are aligned so that the evolution of topics in a sequential corpus is captured. For example, if a topic distribution at time t corresponds to a particular theme, then the distribution that has the same ID number in the consecutive models will relate to the same theme, assuming it appears consistently over time. Thus, the topic drift and topic importance can be monitored by examining the change over time in the probability of words given the topic and the probability of the topic, respectively.

In addition, the overall influence of history in topic estimation is an important factor that can effect the semantic description of the data. For example, some text repositories, like scientific literatures, persistently introduce novel ideas and, thus, topic distributions change faster compared to other datasets. On the other hand, a great part of news in news feed, like sports, stock markets, and weather, are steady over time. Thus, for such consistent topic structures, assigning higher weight for historic information, compared to the weight of current observations, would improve topic discovery, while the settings should be reversed in fast evolving datasets.

By adjusting the total value of elements of the weight vector, ω, our model provides a direct way to deploy and tune the influence of history in the inference process. If the sum of history weights $\sum_{c=1}^{\delta} \omega_c$ is equal to one, this would (relatively) balance the weights of historic and current observations. When the total history weights is less than one, the historic semantic has less influence than the semantic of the current stream.

For more fine tuning of history information, the weights of individual models can be adjusted too. Equal contributions of historic models correspond to assigning equal values to the elements of ω, while decaying history information is represented by a vector of weights in an ascending order. The weights can have an accumulative value so that important words in a topic would have an increasing influence. On the other hand, only the previous knowledge about the "relative" importance of words is preserved and used to guide the prior computation. This can be achieved by assigning the weights in ω to the probabilities computed from the past word-topic frequencies.

2.2 OLDA Generative Process and Algorithm Thus, the generative model for time slice t of the proposed online LDA model is given as follows:

1. For each topic $k = 1, \cdots, K$
    2. Compute $\beta_k^t = \mathbf{B}_k^{t-1}\omega$
    3. Draw $\phi_k^t \sim \text{Dir}(\cdot|\beta_k^t)$
4. For each document, d,

5. Draw $\theta_d^t \sim \mathrm{Dir}(\cdot|\alpha^t)$
6. For each word token, w_i, in document d
   7. Draw z_i from multinomial θ_d^t; $(p(z_i|\alpha^t))$
   8. Draw w_i from multinomial ϕ_{z_i}; $p(w_i|z_i, \beta_{z_i}^t)$

Maintaining the models' priors as Dirichlet is essential to simplify the inference problem by making use of the conjugancy property of Dirichlet and multinomial distributions. In fact, by tracking the history as prior patterns, the data likelihood and, hence, the posterior inference in the static LDA are left the same, and applying them to our proposed model is straightforward. Our model uses Gibbs sampling as an approximate inference method to estimate the word-topic assignments, as in [10]. Hence, the posterior distribution over the assignments of words to topics at time t, $P(\mathbf{z}_i^t = j|\mathbf{w^t})$, is conditioned on the topic assignments to all other word tokens in stream t ($\mathbf{z}_{\neg i}^t$) as follows:

$$P(z_i^t = j|\mathbf{z}_{\neg i}^t, w_{di}^t, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{C_{w_{\neg i},j}^{V^tK} + \beta_{w_i,j}}{\sum_{v=1}^{V^t}(C_{v_{\neg i},j}^{V^tK} + \beta_{v,j})} \times \frac{C_{d_{\neg i},j}^{D^tK} + \alpha_{d,j}}{\sum_{k=1}^{K}(C_{d_{\neg i},k}^{D^tK} + \alpha_{d,k})}$$

where $C_{w_{\neg i},j}^{V^tK}$ is the number of times word w is assigned to topic j, not including the current token instance i; and $C_{d_{\neg i},j}^{D^tK}$ is the number of times topic j is assigned to some word token in document d, not including the current instance i. Unlike the batch settings, the sampling in our online approach is performed over the current stream only, which makes the time complexity and memory usage of OLDA efficient and doable in practice. In addition, the βs under OLDA are constructed from (a linear combination of) historic observations, rather than fixed values.

An overview of the proposed Online LDA algorithm is shown in Algorithm 1. In addition to the text streams, S^t, the algorithm takes as input the weight vector ω, and fixed Dirichlet values, a and b, for initializing the priors α and β, respectively, at time slice 1. Note that b is also used to set the priors of new words that appear for the first time in any time slice. The output of the algorithm is: the generative models, and the evolution matrices $\mathbf{B}_k$ for all topics.

Algorithm 1 Online LDA

```
INPUT: $b; a; CL; \omega; S^t, t \in \{1, 2, \cdots\}$
for (ever) do
    if $t = 1$ then
        $\beta_k^t = b, k \in \{1, \cdots, K\}$
    else
        $\beta_k^t = \mathbf{B}_k^{t-1}\omega, k \in \{1, \cdots, K\}$
    end if
    $\alpha_d^t = a, d = 1, \cdots, D^t$
    initialize $\phi^t$ and $\theta^t$ to zeros
    initialize topic assignment, $\mathbf{z}^t$, randomly for all word tokens in $S^t$
    [$\phi^t$,$\theta^t$, $\mathbf{z}^t$] = GibbsSampling($S^t, \beta^t, \alpha^t$)
    $\mathbf{B}_k^t = \mathbf{B}_k^{t-1} \cup \phi_k^t, k \in \{1, \cdots, K\}$
end for
```

3 Experimental Design

Online LDA (OLDA) with historic semantic is evaluated in the problem domain of document modeling. The objective is to estimate the density of the underlying structure of data. *Perplexity* is a canonical measure of goodness that is used in language modeling. It evaluates the generalization performance of the model on previously unseen documents. Lower perplexity means a better generalization performance, and, hence, a better estimation of density. Formally, for a test set of M documents, the perplexity is [6]:

$$(3.2) \quad perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

We tested OLDA under different configurations of historic semantic embedding. The datasets that were used in our experiments are described in Subsection 3.1, while a summary of the conducted models and their parameter settings are listed in Table 2. The window size, δ, was set to values from 0 to 5. OLDA model with history window of size 0 ignores the history and processes the text stream using fixed symmetric Dirichlet prior. Under such model, the estimation is influenced by the semantics of the current stream only. This model, named OLDAFixed, and OLDA model with $\delta = 1$ are considered as baselines to which the rest of tested models are compared to. At every time instance, we use the documents of the next stream as the test set of the model currently generated, and compute the perplexity.

We also tested the effect of the overall weight of history contribution on future inference. This was achieved by ranging the total sum of weights in ω from 0.05 to δ. Models with total weights greater than 1 were implemented to represent incremental history information. In addition, the elements of ω were either set to equal values (equal contributions), or to values in an ascending order (decaying history).

All models were run for 500 iterations and the last sample of the Gibbs sampler was used for evaluation. The number of topics, K, is fixed across all the streams. K, a, and b are set to 50, $50/K$, and 0.01, respectively. All experiments are run on 2GHz Pentium(R) M-processor laptop using "Matlab Topic Modeling Toolbox", authored by Mark Steyvers and Tom Griffiths[1].

[1] The Topic Modeling Toolbox is available at: psiexp.ss.uci.edu/research/programs_data/toolbox.htm

3.1 Datasets In the following we provide a short description of the datasets used in our experiments.

Reuters-21578[2]. The corpus consists of newswire articles classified by topic and ordered by their date of issue. There are 90 categories with some articles classified in multiple topics. The ApteMod version of this database has been used in many papers. This version consists of 12,902 documents, with approximately 27,000 features in total.

For our experiments, only articles with at least one topic were kept for processing. For data preprocessing, stopwords were removed while the remaining words were down-cased and stemmed to their root source. The resulting dataset consists of 10337 documents, 12112 unique words, and a total of 793936 word tokens. For simplicity, we partitioned the data into 30 slices and considered each slice as a stream.

NIPS dataset[3]. The NIPS set consists of the full text of the 13 years of proceedings from 1988 to 2000 Neural Information Processing Systems (NIPS) Conferences. The data was preprocessed for down-casing, removing stopwords and numbers, and removing the words appearing less than five times in the corpus. The data set contains 1,740 research papers, 13,649 unique words, and 2,301,375 word tokens in total. Each document has a timestamp that is determined by the year of the proceedings. Thus, the set consisted of 13 streams in total. The size of the streams, D^t, varies from 90 to 250 documents.

Table 2: Name and parameter settings of OLDA models. The * indicates that the model was applied on the specified data

Reuters	NIPS	Model Name	δ	ω
*	*	OLDAFixed	0	NA($\beta = 0.05$)
*	*	1/ω(1)	1	1
*	*	2/ω(1)	2	1, 1
*		2/ω(0.8)	2	0.2, 0.8
*	*	2/ω(0.7)	2	0.3, 0.7
*	*	2/ω(0.6)	2	0.4, 0.6
*	*	2/ω(0.5)	2	0.5, 0.5
*	*	3/ω(1)	3	1, 1, 1
*	*	3/ω(0.8)	3	0.05, 0.15, 0.8
*	*	3/ω(0.7)	3	0.1, 0.2, 0.7
*		3/ω(0.6)	3	0.15, 0.25, 0.6
*	*	3/ω(0.33)	3	0.33, 0.33, 0.34
*	*	4/ω(1)	4	1, 1, 1, 1
	*	4/ω(0.9)	4	0.01, 0.03, 0.06, 0.9
*		4/ω(0.8)	4	0.03, 0.07, 0.1, 0.8
*	*	4/ω(0.7)	4	0.05, 0.1, 0.15, 0.7
*		4/ω(0.6)	4	0.05, 0.15, 0.2, 0.6
*	*	4/ω(0.25)	4	0.25, 0.25, 0.25, 0.25
	*	5/ω(1)	5	1, 1, 1, 1, 1
	*	5/ω(0.7)	5	0.05, 0.05, 0.1, 0.15, 0.7
	*	5/ω(0.6)	5	0.05, 0.1, 0.15, 0.2, 0.6
*	*	5/ω(0.2)	5	0.2, 0.2, 0.2, 0.2, 0.2

3.2 Results When OLDA was run on the Reuters dataset, we found that by increasing the window size, δ, OLDA resulted in lower perplexity than the baselines. Figure 2 plots the perplexity of OLDA and OLDAFixed at every stream of Reuters under different settings of window size, δ, and the weight vector, ω, was fixed on $1/\delta$. Figure 3 illustrates the average perplexity, over 30 streams, of OLDA on Reuters for δ set to 1, 2,3, and 4 and different ω settings. Note that for $\delta = 1$ (the left most bar group), the figure shows one result only because experimenting with different ω is not applicable. Both figures clearly show that embedding semantics enhanced the document modeling performance. In addition, incorporating semantics from more models, i.e. using a window size greater than 1, improves further the perplexity with respect to OLDA with short memory ($\delta = 1$).

But, as shown in Figure 3, models with incremental history information, $\delta/\omega(1), \delta > 1$, did not improve topic estimation at all. The same behavior was recorded with NIPS data. This indicates that summing un-weighted models agitates the document modeling process. Yet, normalizing the contributed word-topic distributions to sum to one has always shown better performance.

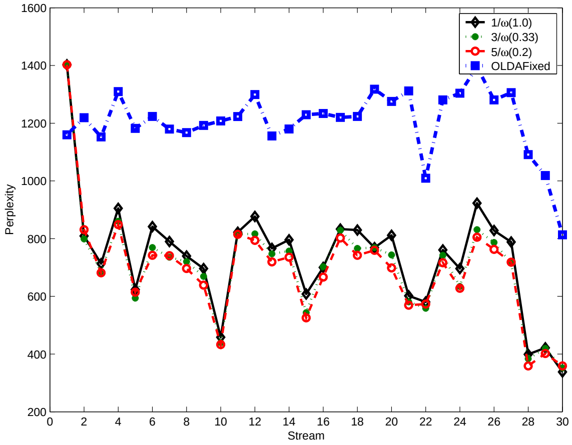

Figure 2: Perplexity of OLDA on Reuters for various window sizes compared to OLDAFixed

Testing with NIPS resulted in a slight different behavior. When ω was fixed, increasing the window size did show a reduction in the model's perplexity, compared to OLDA with short memory. This is illustrated in Figure 4. The larger the window, the lower the perplexity of the model. Nonetheless, the OLDA model only showed improvements with respect to the OLDAFixed when the window size was larger than 3. In addition to the window size, previous experiments on NIPS suggested the effect of the total weight of history in estimating the topical semantics of heterogenous and fast evolving domains like scientific research [1]. The experiments explained next provides the evidence of such

[2] The original dataset is available to download from the UCI Knowledge Discovery in Databases Archive. http://archive.ics.uci.edu/ml/

[3] The original dataset is available at the NIPS Online Repository. http://nips.djvuzone.org/txt.html.

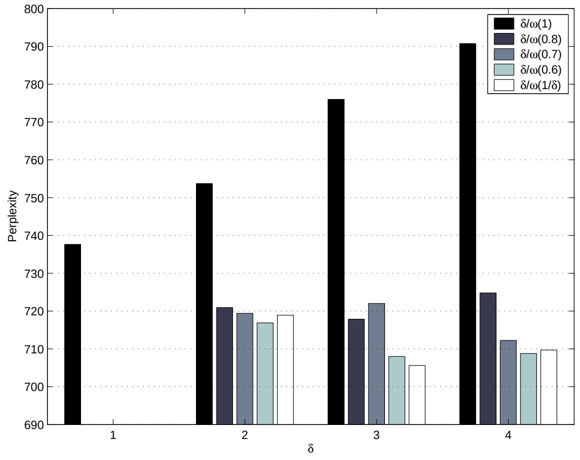

Figure 3: Average perplexity of OLDA on Reuters for various history weightings

justification. Nonetheless, it is worth mentioning here that the OLDA model outperforms OLDAFixed in its ability to automatically detect and track the underlying topics.

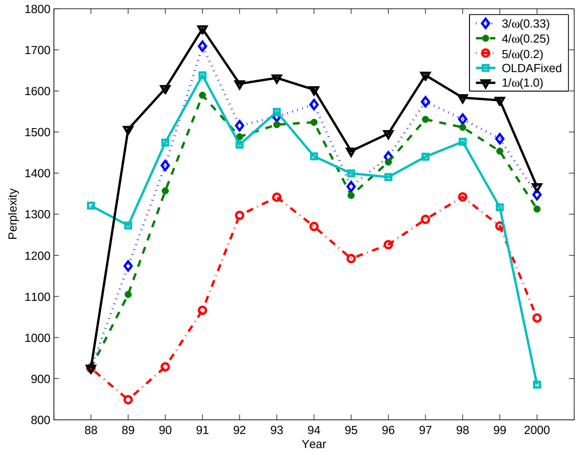

Figure 4: Perplexity of OLDA on NIPS for various window sizes compared to OLDAFixed

To investigate the role of the total history weight, we tested OLDA on NIPS and Reuters under a variety of ω settings. Figure 5 illustrates the perplexity of OLDA on NIPS with δ fixed to 2 and the total sum of ω set to 0.05, 0.1, 0.15, 0.2, and 1. Figure 6 shows the average perplexity of the same models for both datasets. In both figures, both baselines, OLDAFixed and OLDA with short memory, are shown. We found that the contribution of history in NIPS is completely the opposite of Reuters. While increasing the weight for history resulted in a better topical description of Reuters news, lower perplexities were reported with NIPS only for topic models that assigns lower weight for history. In fact, the history weight and perplexity in NIPS (Reuters) are negatively (positively) correlated.

The topics of Reuters dataset are homogenous and more stable. So, letting the current generative model be heavily influenced by the past topical structure would eventually result in a better description of the data. On the other hand, although there is a set of predefined publication domains in NIPS, like algorithm, applications, and visual processing, these topics are very broad and interrelated. Furthermore, research papers usually span over more topics and continuously introduce novel ideas and topics. Hence, the influence of previous semantics should not exceed the topical structure of the present.

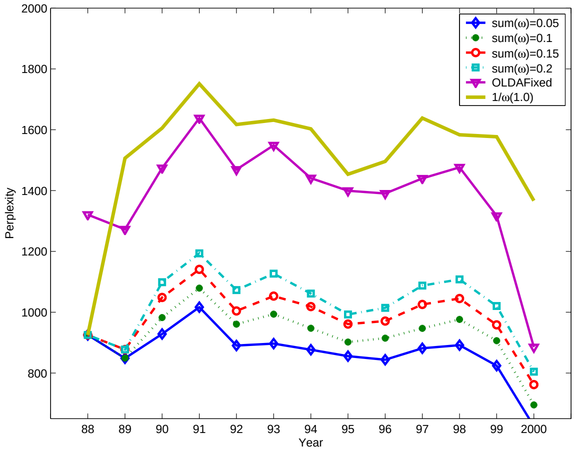

Figure 5: Perplexity of OLDA on NIPS under different weights of history contribution

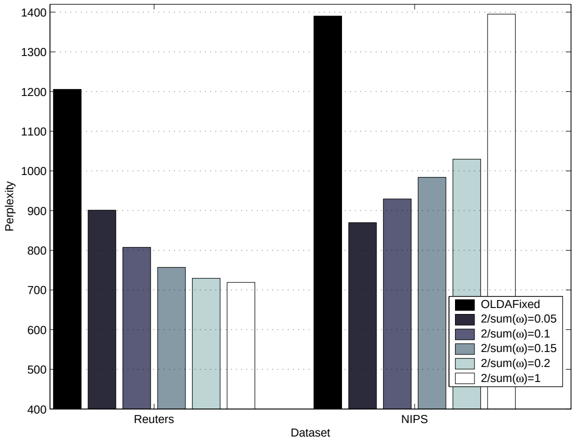

Figure 6: Average perplexity of OLDA on Reuters and NIPS under different weights of history contribution compared to OLDA with fixed β

We also experimented with different settings of weights of individual models, mainly, to investigate the role of equal versus decaying contribution. Figures 7, 8, 9, and 10 show the perplexity of OLDA trained on NIPS streams with δ set to 2,3,4, and 5, respectively, and under different settings of weights, ω. The plotted baseline is the perplexity of OLDA with short memory. Figure 3 also shows the average perplexity of OLDA on Reuters under different settings of weights ω. The best model that resulted in the lowest perplexity among all was model $(5/\omega(0.2))$ for Reuters and model $(4/\omega(0.7))$ for NIPS. In general, for both datasets, we found that equal contribution of past models, although naive, either performed better than, or comparable to, the rest of the models with the same window size. It is important to note that the Reuters data was divided into streams in a naive way too; this fact can provide some explanation for the previous observation. Assigning a distinct weight for individual models would be particularly significant if the rate of stream arrival and domain perplexity are known.

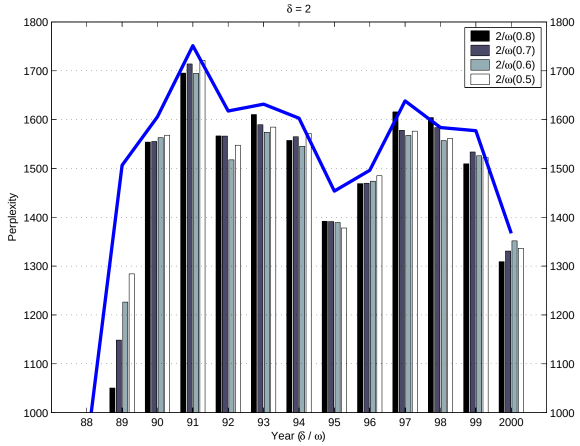

Figure 7: Perplexity of OLDA on NIPS with window size $(\delta) = 2$ and various settings of ω

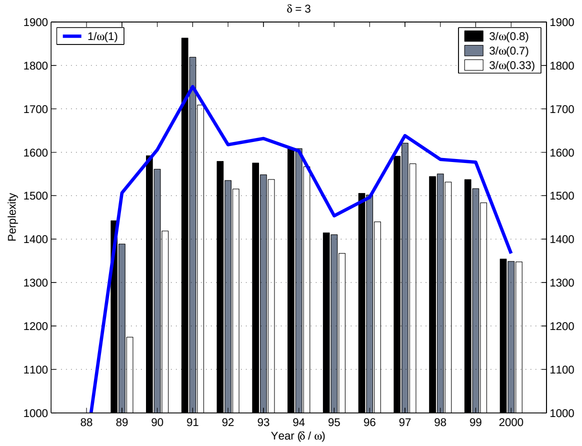

Figure 8: Perplexity of OLDA on NIPS with window size $(\delta) = 3$ and various settings of ω

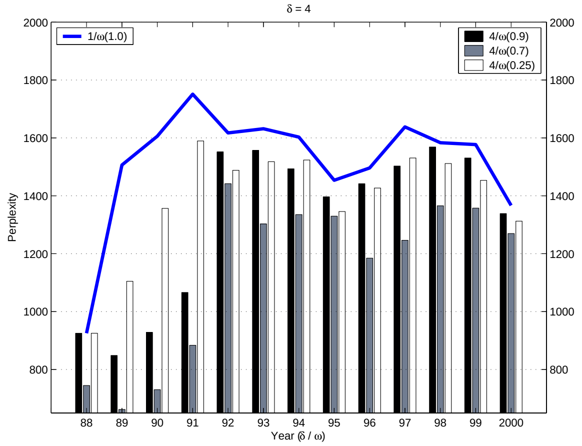

Figure 9: Perplexity of OLDA on NIPS with window size $(\delta) = 4$ and various settings of ω

4 RELATED WORK

Statistical modeling have been recently deployed to solve the problem of identifying and tracking topics in time-stamped text data, e.g. using PLSI model ([7]) and LDA model ([5, 10, 13, 16, 18, 19]). However, most of the work either processes archives in an off-line fashion (e.g. [19]), post-discretizes the time ([16]), or uses unconjugated priors to multinomial distributions and trained on all the previous data (e.g. [5, 18]). Our topic model, however, processes small subsets of data in an online fashion while making use of the conjugacy property of the Dirichlet distribution to keep the model's structure simple, and to enable sequential inference.

The problem of embedding semantic information within the document representation and/or distance metrics is another emerging problem in text mining that has been heavily investigated in the domain of text classification and clustering, e.g. [2], [3], [8], and [12]. However, to the best of our knowledge, our paper presents one of the earliest attempts, if not the first, to embed semantic information to enhance online document modeling within the framework of LDA.

A number of papers in the literature have used LDA topic modeling to represent some kind of semantic embedding. In the domain of text segmentation, the work in [17] used an LDA-based Fisher kernel to measure the text semantic similarity between blocks of documents in the form of latent semantic topics that were previously inferred using LDA. The kernel is controlled by the number of shared semantics and word co-occurrences. Phrase discovery is another area that aims to identify phrases (n-grams) in text. Wang et. al [20] presented Topical N-Gram model that au-

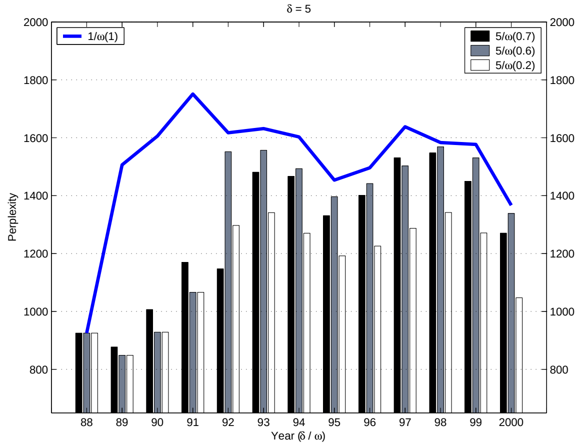

Figure 10: Perplexity of OLDA on NIPS with window size $(\delta) = 5$ and various settings of ω

tomatically identified feasible N-grams based on the context that surround it. Moreover, there are some research work to incorporate prior-knowledge from large universal datasets, like Wikipedia. The work in [14] built a classifier on both a small set of labeled documents in addition to an LDA topic model estimated from Wikipedia.

5 Conclusion And Future Work

In this paper, the effect of embedding semantic information in the framework of probabilistic topic modeling of text streams was investigated. Particularly, we introduced an online LDA topic model (OLDA) that constructs its current parameters based on the topical semantics that have been inferred by the past generated models. The proposed solution is completely unsupervised and requires no external or prior knowledge. Our experiments showed that OLDA did enhance the performance of document modeling.

Moreover, a variety of related factors were analyzed, and their effect on the performance of document modeling was investigated. These factors include the total influence of history, the history window size, and the effect of equal or decaying contributions.

To extend this work, we are considering the use of prior-knowledge to learn (or enhance the construction of) the parameters. In addition, the effect of the embedded historic semantics on detecting emerging and/or periodic topics is also part of our future work.

References

[1] L. AlSumait, D. Barbará, and C. Domeniconi, Online LDA: Adaptive Topic Model for Mining Text Streams with Application on Topic Detection and Tracking, *Proceedings of IEEE International Conference on Data Mining (ICDM08)*, (2008).

[2] L. AlSumait and C. Domeniconi, Text Clustering with Local Semantic Kernels, *Survey of Text Mining: Clustering, Classification, and Retreival*, 2nd Ed., M. W. Berry and M. Castellanos, Ed. Springer, (2008).

[3] Roberto Basili, Marco Cammisa and Alessandro Moschitti,*A Semantic Kernel to exploit Linguistic Knowledge*, In proceedings of the 9th Conference of the Italian Association for Artificial Intelligence (AI*IA 2005), Lecture notes in Computer Science - Springer Verlag, (2005).

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] D. M. Blei and J. D. Lafferty, "Dynamic topic models," *In International conference on Machine learning*, pp. 113-120, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, (2003).

[7] T. Chou and M. Ch. Chen, "Using Incremental PLSI for Threshold-Resilient Online Event Analysis," *the IEEE Transactions On Knowledge And Data Engineering*, vol. 20, no. 3, 2008.

[8] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi, *Latent Semantic Kernels*, Journal of Intelligent Information Systems, Springer Netherlands. 18(2-3), (2002), pp. 127-152.

[9] Gartner Group. www.gartner.com.

[10] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceeding of the National Academy of Sciences*, pp. 5228–5235, 2004.

[11] T. Hofmann, "Probablistic Latent Semantic Indexing," *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

[12] Chung-Hong Lee, Hsin-Chang Yang.*A Classifier-based Text Mining Approach for Evaluating Semantic Relatedness Using Support Vector Machines*, IEEE International Conference on Information Technology: Coding and Computing (ITCC'05), I, (2005), pp. 128-133.

[13] R. Nallapati, S. Ditmore, J.D. Lafferty, and K. Ung, "Multiscale topic tomography," *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery in data mining*, 2007.

[14] X. Phan, L. Nguyen, and S. Horiguchi, "Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-scale Data Collections," *The International World Wide Web Conference Committee (IW3C2)*, 2008.

[15] M. Steyvers and T. L. Griffiths, "Probabilistic Topic Models," In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (ed), *Latent Semantic Analysis: A Road to Meaning*, Laurence Erlbaum, 2005.

[16] X. Song, C.-Y. Lin, B. L. Tseng, and M.-T. Sun, "Modeling and predicting personal information dissemination behavior," *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

[17] Q. Sun, R. Li, D. Luo, and X. Wu, "Text Segmentation with LDA-Based Fisher Kernel," *Proceedings of Association for Computational Linguistics*, 2008.

[18] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," *The 23rd Conference on Uncertainty in Artificial Intelligence*, 2008.

[19] X. Wang and A. McCallum, “Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends,” *ACM SIGKDD international conference on Knowledge discovery in data mining*, 2006.

[20] X. Wang, A. McCallum,, and X. Wei, “Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval,” *Seventh IEEE International Conference on Data Mining*, 2007.