WordNet-Enhanced Topic Models

Hsin-Min Lu
National Taiwan University
Management Building II, Room 509
1 Roosevelt Rd. Sec. 4,
Taipei, Taiwan 106
+886-2-33661184
lu@im.ntu.edu.tw

ABSTRACT
Wordnet is one of the most popular human-generated knowledge base that contains the relationships of synsets. Each synset is a group of words expressing a unique concept. Compared to the semantic structures discovered automatically by topic models, Wordnet synsets may be more coherent for human interpretation. This study explores the idea of incorporating Wordnet synsets as the prior knowledge to help discovering the latent topic structures in a document collection. A novel Wordnet-enhanced topic model (WNTM) is developed to incorporate the synset information at token-level using the multinomial Probit regression prior. During the learning phase, WNTM expands and combines existing synsets to form topics. WNTM also creates new topics unrelated to existing synsets when appropriate. Experiments show that the newly developed WNTM performs better compared to the latent Dirichlet allocation (LDA) in terms of perplexity.

Categories and Subject Descriptors
G.3 [**Probability and Statistics**]: Probabilistic algorithms; I.2.7 [Artificial Intelligence]: Natural Language Processing—Text analysis

General Terms
Algorithms, Performance, Design.

Keywords
Wordnet-enhanced topic model, multinomial Probit regression, Gibbs sampling, text mining, prior knowledge.

1. INTRODUCTION
Topic models such latent Dirichlet allocation (LDA) [3] model word occurrences in a document as the mixture of latent topics, which are distributions of the vocabulary. Efficient estimation approaches such as collapsed Gibbs sampling [7] and variational Bayes [1] can be used to learn the latent structure as specified by LDA. While the topics estimated based on the word co-occurrence structure are typically meaningful for human interpretation, there may be situations when the prior knowledge such as the synsets in Wordnet may contribute to the learning process.

One example is when certain words with similar meanings happened to have low frequencies. Topic models may fail to group them together due to weak statistical evidence. Wordnet synsets may be useful for topic models to better discover the hidden topical structure. On the other hand, relying only on Wordnet [12] may be unsuitable because it may not cover every aspect of a document collection. Concepts evolve constantly and updating a large-scale knowledge base takes time. A topic modeling approach that has the flexibility to incorporate and expand human-generated knowledge may improve upon existing approaches.

In this paper, we propose a Wordnet-enhanced topic model (WNTM) to incorporate Wordnet synsets as the prior knowledge of topic models. Wordnet is selected as the representative human-generated knowledge base in this study. The proposed WNTM, however, is a general algorithm that can be applied to other knowledge base with similar types of information.

Unlike LDA and some variations, WNTM moves away from the conjugate Dirichlet prior and adopted a multinomial Probit regression-like structure to incorporate token-level synset information. Each token is associated with a vector that indicates the synsets it belongs to. Note that a word may belong to more than one synset if it has multiple senses. WNTM determines the mapping between synsets and topics during model learning. The learned coefficients, which indicate synset-topic mappings, allow a topic to be associated with one or more synsets. Words not exist in the synsets may also receive high conditional probability in the topic if it is favored by the data. New topics unrelated to existing synsets may also be created by WNTM. In essence, the multinomial Probit-like structure allows WNTM to determine, based on the observed data, how Wordnet synsets are incorporated in the discovered topics.

One advantage of adopting the regression structure for the prior information in topic modeling is the flexibility of combining information. The learned synset coefficients determine the tendency that a word is linked to a topic. WNTM also includes a document-specific vector that represents the document-specific topic tendency. These coefficients for each word jointly determine the prior probability it belongs to a topic.

The flexible token-level regression structure creates a challenging model inference problem. The non-conjugate setting does not allow document-specific topic probability to be collapsed. We developed an augmented Gibbs sampling algorithm to approximate model posterior. The latent topic assignments, coefficients for synsets, and the document-specific topic tendency are updated in sequence. A set of intermediate latent variables that links the regression coefficients to discrete latent topic assignments are created (i.e. "augmented") in order to facilitate the Gibbs sampling procedure.

Conference'10, Month 1–2, 2010, City, State, Country.

Our work is an extension of the on-going research stream that incorporates meta-data and context information into topic models. We review related work in the next section, followed by a detailed discussion of WNTM. Experimental results that compare our approach with LDA are then presented.

2. RELATED WORK

Topic models are a family of algorithms that can identify latent topical structures in a document collection. One well-known topic model is the LDA that shows better performance compared to the probabilistic latent semantic indexing (pLSI) [3]. Assume that there are J topics in a collection of D documents. Each topic j ($0 \leq j \leq J-1$) is associated with ϕ_j , a distribution over vocabulary following the Dirichlet distribution with the hyperparameter β. Let z_{di} be the latent topic of the i-th word ($1 \leq i \leq N_d$) in document d ($1 \leq d \leq D$), then the observed word of document d at position i, w_{di}, is sampled from $\phi_{z_{di}}$.

LDA assumes that each document d is associated with θ_d , a distribution over the potential topics in the document following the Dirichlet distribution with the hyperparameter α. For each position i in document d, the data generating process starts by sampling z_{di} from Discrete(θ_d), followed by sampling w_{di} from Discrete($\phi_{z_{di}}$).

One weakness of LDA is that there is no convenient ways to include domain knowledge into the model. To address this issue, Andrejewski, Zhu and Craven [5] adopt the Dirichlet forest priors (DFP) to enforce constraints on topic distributions. The goal is to provide a convenient way for users to interactively construct a topic model by providing feedbacks to the model. The feedbacks are must-link and cannot-link relationships between words that constraints the model inference process. The Dirichlet forest priors, a collection of Dirichlet tree distributions, are used to represent these constraints.

It is important to note that the topic model needs to follow the domain knowledge encoded in the Dirichlet forest prior to the degree specified by the hyperparameters of the prior. In other words, the Dirichlet forest priors are meant to be constraints instead of suggestions to the topic models. The model cannot choose to "turn off" the constraints when hyperparameters enforce rigid priors.

Concept topic model (CTM) [4; 6] achieves similar goals with different model specifications. It assumes that the observed words are either generated from a set of hidden topics or a set of fixed concepts. The concepts are human-defined groups of words with similar meanings. For a CTM with T topics and C concepts, a word at position i of document d is generated by first drawing z_{di} from $Discrete(\theta_d)$, where θ_d is a vector of length T+C that represent the topic and concept mix of document d. If $z_{di} \leq T$, then the word is generated based on topic z_{di}. Otherwise, the concept $c_{di} = z_{di} - T$ is used to generate the word. It is clear from the data generating process of CTM that concepts are a special kind of topics. The difference is that only those words preexist in a concept can receive positive probability.

Another related model is the latent Dirichlet allocation with WordNet (LDAWN) [14]. LDAWN assumes that words in a document are generated by walking down the tree structure of Wordnet synsets. One goal is to leverage the hyponym structure among synsets to better disambiguate the senses of a word. This model, however, cannot handle words that do not exist in Wordnet.

Other interesting extensions exist. Examples include the hierarchical concept topic model (HCTM) [4], Dirichlet-multinomial regression (DMR) [13], author-topic (AT) models [16], correlated topic models [2], and time-dependent topic models [18].

Previous studies on incorporating human-generated knowledge into topic models mostly take one of the following three approaches. The first approach focuses on constraining the topic model so that the learned topic models are consistent with human interpretation. The second approach includes preexisting concepts as additional "topics" for topic models to choose during model learning. Finally, LDAWN relies completely on the Wordnet tree structure for word sense disambiguation. In contrast to previous methods, the proposed WNTM incorporates Wordnet synsets as additional features of each latent topic so that the model can decide how to link the synsets to the hidden topics. The model may decide to ignore all synsets by assigning coefficients close to zero. On the other hand, the model may decide, for a given topic, to assign large coefficients to one or more synsets so that the topic is a combination of several existing synsets. The decisions are done automatically during the model inference process to better discover the hidden structure of a given text collection.

3. WORDNET-EHNANCED TOPIC MODEL

The Wordnet-enhanced topic model (WNTM) incorporates Wordnet synsets to improve topic modeling outcomes. We first discuss the synset representation adopted in WNTM, followed by the model setting to incorporate the synset data into topic models.

3.1 Concept Construction

Wordnet organize nouns, verbs, adjectives, and adverbs via different structures. Synsets for nouns are organized as a tree structure while verbs are organized as a semantic net [12]. This study focuses on nouns. Other part of speech can be included using similar approaches.

The intension is to incorporate higher level of abstraction through synset information. In the following discussion, the term concept is defined as a group of words with similar meanings constructed from Wordnet synsets. For a word in a document, related concepts are constructed by first mapping to its base form using the Wordnet morphy tool. Matching noun synsets are then collected. For each collected synsets, the corresponding concept is constructed by including words in this synset, its descendants, its parent, it siblings, and descendants of siblings. The parent synset name is used to identify this concept.

As an illustrative example, consider the word debt. There are three senses in Wordnet. To construct the concept for the second sense "money or goods or services owed by one person to another," we include words in this sense (debt), its descendants (arrears, loan, principle, score, national debt, public debt, etc.), its parent (liabilities), its siblings (tax liability, payables, deficit, charge, and accounts payable), and descendants of siblings (budget deficit, trade deficit, levy, etc.). The concept is the union of all unigrams under consideration. Unigram tokens not observed in the text collection are deleted. The parent synset name, liabilities.n.01, is used to identify this concept. The other two concepts for debt are constructed in a similar manner.

The process of concept construction is repeated for every word in the text collection. Note that the above process may accidently include irreverent concepts. For example, the word "cts" in Reuters-21578 is the abbreviation for "cents." Our concept

construction process, however, convert cts to "ct" using the morphy tool and maps to a sense of computed tomography scan. This concept is unrelated to the original meaning of cts in the text collection.

To identify concepts that are most relevant to the text collection, we develop a set of filtering and scoring methods. A useful concept is those that provide enough synonyms and is also relevant to the text collection. A concept with few words is not useful since the model is not able to generalize topic assignments through these concepts. The first step, as a result, is to remove concepts with less than five distinct tokens.

We then compute the average co-occurrence length for each remaining concepts. For a concept, the co-occurrence length in a document is the number of unique tokens appearing in the document. The average co-occurrence length is the mean co-occurrence length for all documents with positive co-occurrence length. The second step is to delete all concepts with average co-occurrence length less than 1.15.

The remaining concepts are sorted in descending order by a relevance score defined as the average co-occurrence length divided by the number of unique tokens in the concept. The concepts with a relevance score in the last 25% percentile are deleted.

Finally, we remove similar concepts by computing the pairwise Jaccard similarity coefficient and delete a concept in the pair with a similarity coefficient larger than 0.1. The remaining concepts form the Wordnet background knowledge to be included into WNTM.

3.2 Incorporating Wordnet Concepts into Topic Models

To incorporate the Wordnet concepts constructed from synsets, WNTM adopts a multinomial Probit regression-like structure for the prior of topic assignment. Figure 1 plots the WNTM.

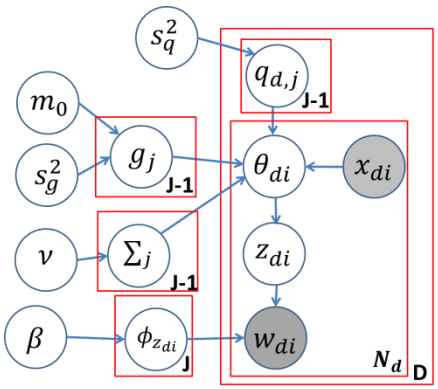

Figure 1. Wordnet Enhanced Topic Model (WNTM).

WNTM has a similar structure as the LDA. The most important difference is how z_{di}, the latent topic for the word at position i of document d, is generated. In LDA, z_{di} is generated from a multinomial distribution with a document-specific topic mix θ_d. WNTM adopted a more flexible structure to allow each z_{di} to be generated by its own topic mix θ_{di}. This vector has a length J (the number of latent topic in WNTM). Each element of θ_{di} is the prior probability that the word at position i of document d belongs to a topic j (j$= 0, 1, 2, \dots, J-1$) conditional on the document-specific topic tendency $q_d = (q_{d,1}, q_{d,2}, \dots, q_{d,J-1})$, the concept features x_{di}, and its coefficients g_1, g_2, …, g_{J-1}. The elements in θ_{di} sum to one.

The dependency between θ_{di} and other variables is defined by introducing an augmented variable $H_{di,j}$. Specifically, let $H_{di,j} = q_{d,j} + x'_{di}g_j + e_{di,j}$, for $j = 1,2, \dots, J-1$. The document-specific effect $q_{d,j}$ captures the topic tendency of topic j for document d. The concept feature x_{di} is a vector of zeros and ones that indicates whether a word belongs to a Wordnet concept. A constant 1 is attached at the beginning of the vector for the corpus-wide topic tendency. The coefficient g_j is to be estimated using the observed document collection. The white noise $e_{di,j} \sim N(0, \Sigma_j)$ has the same distribution for topic j across all words in the document collection.

The topic assignment z_{di} is determined by comparing $H_{di} \equiv (H_{di,1}, H_{di,2}, \dots, H_{di,J-1})$ so that:

$$z_{di} \equiv Y(H_{di}) = \begin{cases} 0, & \text{if } \max(H_{di}) \le 0 \\ j, & \text{if } \max(H_{di}) = H_{di,j} > 0, \end{cases}$$

Note that the index of latent topic starts from zero. The above process can be seen as a way to parameterize θ_{di}, which then determines z_{di}. That is, $\theta_{di} = \big(P(z_{di} = 0|q_d, g, x_{di}), P(z_{di} = 1|q_d, g, x_{di}), \dots, P(z_{di} = J-1|q_d, g, x_{di})\big)$, and $z_{di} \sim Discrete(\theta_{di})$. While this setup is not the only way to parameterize θ_{di}, it provides a convenient way to draw coefficients (g, q_d, and Σ) via Gibbs sampling from their conditional posteriors. Model inference will be discussed in the following section.

As an illustrative example, consider a model with three topics (J=3). Assume that there are three words in two documents. The first document contains "tender" and "loan"; the second document contains "ahead." The document-specific topic tendency vectors are $q_1 = (0.4, 0.1)'$ and $q_2 = (-1, -1.5)'$. Other things being equal, the first document tends to have topic 1 since the first element is larger and the second document tends to have topic 0 since both elements are negative.

For simplicity, assume that the Wordnet concept construction procedure leads to the inclusion of concepts "offer.n.02" and "medium_of_exchange.n.01" for "tender," and "liabilities.n.01" for "loan." "Ahead" is not associated with any concepts. Arranging the concepts by the order of "offer.n.02," "medium_of_exchange.n.01," and "liabilities.n.01," the feature vector is $(1, 1, 1, 0)'$ for "tender," $(1, 0, 0, 1)'$ for "loan," and $(1, 0, 0, 0)'$ for "ahead." Note that the first element is always one. The coefficients g_1 and g_2 determines how the features contribute to the realization of hidden topics. This example assumes that topic 1 is related to the concept "offer.n.02" and "liabilities.n.01" with $g_1 = (-0.21,\ 1,\ 0,\ 1)'$ and topic 2 is marginally related to the concept "medium_of_exchange.n.01" with $g_2 = (-0.21,\ 0,\ 0.5,\ 0)'$.

The while noise, $e_{di,j}$, determines how rigid the document-specific tendency and Wordnet features determines the topic assignment. This example assumes that $e_{di,1} \sim N(0,1)$ and $e_{di,2} \sim N(0.1)$ for all words.

The above discussion suggests that the word "tender" in the first document has $(H_{11,1}, H_{11,2}) = \big(q_{1,1} + x'_{11}g_1 + e_{11,1},\ q_{1,2} + x'_{1,1}g_2 + e_{11,2}\big) = (1.19 + e_{11,1},\ 0.39 + e_{11,2})$. Since both $e_{11,1}$ and $e_{11,2}$ have unit variance, there is a large chance that $H_{11,1} > H_{11,2}$ and $H_{11,1} > 0$. In fact $\theta_{11} = (0.030,\ 0.720,\ 0.250)$; that is,

$P(z_{11} = 1|q_1, x_{11}, g) = 0.720$. Similarly, for the second word in document 1, $(H_{12,1}, H_{12,2}) = (1.19 + e_{12,1}, -0.11 + e_{12,2})$ and $\theta_{12} = (0.047,\ 0.801,\ 0.152)$; $(H_{21,1}, H_{21,2}) = (-1.21 + e_{21,1}, -1.71 + e_{21,2})$ and $\theta_{21} = (0.855,\ 0.105,\ 0.040)$.

Note that both θ_{11} and θ_{12} have a larger probability for topic 1 since these two positions share the same document-specific tendency q_1. However, since the concept features for these two words are different, the prior probability for topic 1 is larger for the second position. The first word in the second document, on the other hand, has a large probability (0.855) of picking topic 0 since the negative document-specific tendency makes it easy to have negative $H_{21,1}$ and $H_{21,2}$.

We conclude the discussion with a summary of the data generating process:

(1) For each topic $j = 0, 1, \dots, J-1$
  a. Draw ϕ_j (word distribution for topic j) from $Dirichlet(\beta)$.
  b. If $j > 0$, draw g_j (regression coefficients) from $MultivariateNormal(m_0, s_g^2)$.
  c. If $j > 0$, draw Σ_j from $InverseChiSquare(v)$.

(2) For j=$J-1, J-2, \dots, 1$, set $\Sigma_j = \Sigma_j/\Sigma_1$.

(3) For each document $d = 1, 2, \dots, D$,
  a. For each topic $j = 0, 1, 2, \dots, J-1$,
    i. Draw q_{dj} (document-specific topic tendency) from $q_{d,j} \sim N(0, s_q^2)$.
  b. For each position $i = 1, 2, \dots, N_d$,
    i. Draw $e_{di,j}$ (token-level noise) from $e_{di,j} \sim N(0, \Sigma_j)$.
    ii. Conditional on x_{di} compute $H_{di,j} = q_{d,j} + x'_{di} g_j + x_{di,j}$.
    iii. Assign z_{di}:
$$z_{di} = \begin{cases} 0, & \text{if } \max(H_{di}) \le 0 \\ j, & \text{if } \max(H_{di}) = H_{di,j} > 0, \end{cases}$$
    iv. Choose a word w_{di} from $w_{di} \sim Discrete\ (\phi_{z_{di}})$.

Note that at Step (2) the variance of white noise is normalized so that $\Sigma_1 = 1$. This step is essential for the model to be identifiable. Consider the latent topic z_{di} and the corresponding H_{di}:

$$H_{di,j} = q_{d,j} + x'_d g_j + e_{di,j},\ e_{di,j} \sim N(0, \Sigma_{jj}),\ j = 1,2, \dots, J-1.$$

If $H_{di,j}$ is scaled by a, then $\tilde{H}_{di,j} = aH_{di,j} = \tilde{q}_{d,j} + x_d^T \tilde{g}_j + \tilde{e}_{di,j}$, $\tilde{e}_{di,j} \sim N(0, \tilde{\Sigma}_{jj})$, $j = 1,2, \dots, J-1$, where $\tilde{q}_{d,j} = aq_{d,j}$, $\tilde{g}_j = ag_j$, and $\tilde{\Sigma}_j = a^2\Sigma_j$. Since the topic assignment based on H_{di} and $\tilde{H}_{di}$ is the same, the original H_{di} is not a unique model. The normalization for $\Sigma_1 = 1$ avoid the problem and make sure that the model is identifiable [10].

4. MODEL INFERENCE

Model inference for WNTM can be achieved by viewing the model as the hybrid of LDA and multinomial Probit regression. Figure 2 plots an alternative setting for WNTM. The difference is that θ_{di} is removed from the diagram and the regression coefficients determine the latent topic directly. Note that under this model setting, the upper half of the model (how $q_{d,j}$, g_j, x_{di}, and Σ_j influence z_{di}) is very similar to the multinomial Probit regression if z_{di} can be observed. On the other hand, if all regression coefficients are given, then the lower-half of the model is very similar to the LDA based on the model setting in Figure 1.

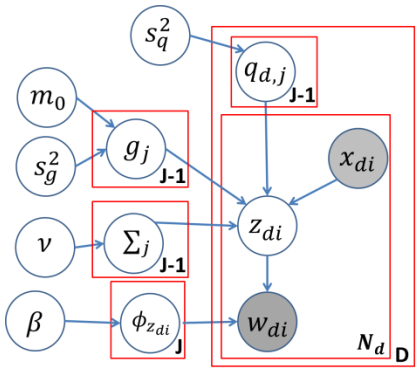

Figure 2. Alternative View of Wordnet Enhanced Topic Model.

Based on the observation, the Gibbs sampling for WNTM can roughly be divided into two parts:

- Draw latent topic z from $p(z|w, X, q, g, \Sigma, \cdot)$.
- Draw regression coefficients q, g, Σ from $p(q, g, \Sigma|w, X, z, \cdot)$.

Here $z = \{z_{di}\}$ is the collection of all latent topics, $q = \{q_{d,j}\}$ is document-specific topic tendency for all documents, $g = \{g_j\}$ is regression coefficients, $\Sigma = \{\Sigma_j\}$ is the variance of white noise, $w = \{w_{di}\}$ is all observed words, and $X = \{x_{di}\}$ is all Wordnet concept features. The dot ("·") represents the parameters of prior distributions. We discuss these two steps below.

4.1 Drawing Latent Topic z

Updating the latent topic z is done sequentially by drawing each z_{di} conditional on other variables. The procedure is very similar to the collapsed Gibbs sampling for LDA [7]. The posterior of z_{di} conditional on other variables is:

$$\begin{aligned} &p(z_{di} = j | z_{-di}, , w_{di}, w_{-di}, X, q, g, \Sigma) \\ &\propto p(w_{di} | z_{di} = j, z_{-di}, w_{-di}) p(z_{di} = j | q, g, \Sigma, X) \\ &= \frac{n_{-di,j}^{(w_{di})} + \beta}{n_{-di,j}^{(\cdot)} + W\beta} p(z_{di} = j | q_d, g, \Sigma, x_{di}), \end{aligned} \quad (1)$$

where $n_{-di,j}^{(\cdot)}$ is the number of assignment to topic j, excluding the assignment at position i of document d; $n_{-di,j}^{(w_{di})}$ is the instance word type w_{di} assigned to topic j, excluding the instance at position i in document d; W is the number of unique words in the corpus.

The first term in (1) is derived by integrating out ϕ_j and can be readily computed based on word-topic assignments. The second term, however, involves intractable integrals. We propose to evaluate this term using a simulation method.

Let $\hat{\theta}_{di,j}$ denote the estimated $p(z_{di} = j | q_d, g, \Sigma, x_{di})$, probability of assigning to topic j conditional on the document meta data x_d and regression parameters. The idea is to compute $\hat{\theta}_{di,j}$ by generating $e_{di,j} \sim N(0, \Sigma_j)$, computing $H_{di,j}$,and then determining topic assignment. The process is repeated for G times to compute $\hat{\theta}_{di,j}$ by computing the probability of reaching each topic.

Direct implementation of this approach introduces independent simulation error into $\hat{\theta}_{di,j}$, which may be undesirable in our setting. To address the problem, we adopt the common random

variable method [8] and cache $G(J-1)$ draws of random variable from the standard normal distribution. The probability estimation $\hat{\theta}_{di,j}$ for each word is computed based on the same set of random variables. This approach provides the internal consistency for the estimated probability.

4.2 Drawing Regression Coefficients

The problem of drawing regression coefficients from $p(q, g, \Sigma | w, X, z, \cdot)$ is very similar to the inference problem of multinomial Probit model [9; 11]. We follow marginal augmentation approach [9] and include two augmented variables for model inference. The first variable is H_{di} that bridges the regression coefficients and latent topic z_{di}. The other augmented variable is a scaling variable a ($a > 0$) that addresses the technical difficulties caused by the identification constraint $\Sigma_1 = 1$. Compared to alternative approaches such as conditional data augmentation [11] or McCulloch's approach [10], this approach achieves higher convergence speed and only need to sample from standard distributions [9].

As discussed in the previous section, the latent topic z_{di} is determined by $H_{di,j} = q_{d,j} + x'_{di} g_j + e_{di,j}$, for $j = 1,2,\dots,J-1$. The topic is assigned to the largest positive $H_{di,j}$ and is assigned to topic 0 if every $H_{di,j}$ is negative. The model with $\Sigma_1 = 1$ is referred to as the restricted model in subsequent discussion.

The advantage of including $H_{di,j}$ is clear: the posterior of regression coefficients $(q_{d,j}, g_j, \Sigma_j)$ conditional on $H_{di,j}$ are well-known and can be updated via Gibbs sampling. The cost, nonetheless, is the additional computing effort to update $H_{di,j}$.

The conditional posterior of $H_{di,j}$, $H_{di,j}|H_{di,-j}$, z_{di},$\cdot$, follows a truncated normal distribution with mean $m_{di,j} = q_{d,j}^{(old)} + \mathrm{x}'_{\mathrm{d}} g_j^{(old)}$ and variance $\tau_{di,j}^2 = \Sigma_j^{(old)}$ (before truncation). The vector $H_{di,-j}$ denotes the elements in H_{di}, excluding $H_{di,j}$. If $z_{di} = j > 0$, then the normal distribution is truncated below at $\max(H_{di,-j}, 0)$. Otherwise if $z_{di} \neq j > 0$, the distribution is truncated above at $\max(H_{di,-j}, 0)$. Finally if $z_{di} = 0$, then the normal distribution is truncated above at 0.

The scaling variable a transfer the restricted model to an alternative representation by scaling up $H_{di,j}$:

$$\widetilde{H}_{di,j} = aH_{di,j} = aq_{d,j} + x'_{di}\, ag_j + ae_{di,j}$$

Note that in this alternative representation, $Var(a\, e_{di,1}) = a^2$. The marginal data augmentation approach transfers the restricted model to an alternative augmented-data model, draw the regression coefficients, then scale down the updated coefficients.

Specifically, the first step is to draw a^{2^*} from $(\sum_{j=1}^{J-1} 1/\Sigma_j^{(old)})/\chi^2_{(J-1)^2}$, where $\Sigma_j^{(old)}$ is the white noise variance for topic j from the previous sweep. This step determines the augmented-data model used in this sweep. The second step is to determine $\widetilde{H}^*_{di,j}$ conditional on z and old regression coefficients. It is achieved by drawing $H^*_{di,j}$ from the restricted model (with $\Sigma_1 = 1$) and then set $\widetilde{H}^*_{di,j} = a^* H^*_{di,j}$.

The third step is to update regression coefficients q, and g. For a topic j, we first update g_j, followed by the random effect q_j. To update g_j conditional on other parameters, note that $\widetilde{H}^*_{di,j} - a^* q_{d,j}^{(old)} = x'_d \tilde{g}_j + \tilde{e}_{di,j}$. Let $y_{di,j} = \widetilde{H}^*_{di,j} - a^* q_{d,j}^{(old)}$ and stack the variable from the first word in the first document to the last word in the last document to form $y_j = \left(y_{11,j}, y_{12,j}, \dots, y_{1N_1,j}, y_{21,j}, \dots y_{2N_2,j}, \dots, y_{D1,j}, \dots, y_{DN_D,j}\right)'$. Similarly, stack x_d in a similar manner to define $X = (x'_1, \dots, x'_1,\ x'_2, \dots, x'_2, \dots,\ x'_D, \dots, x'_D)'$, where each x'_d is repeated for N_d times. There are $\sum_{d=1}^{D} N_d$ rows and K columns in X. Following Bayesian regression theory, the posterior of $\tilde{g}_j$ follows a multivariate normal distribution with mean m_{g_j} and variance V_{g_j}:

$$\mathrm{m}_{g_j} = \left(\frac{X'X}{\Sigma_j^{(old)}} + A\right)^{-1} \left(\frac{X' y_j}{\Sigma_j^{(old)}}\right)$$

$$\mathrm{V}_{g_j} = \left(\frac{\mathrm{X'X}}{a^{2^*}\Sigma_j^{(old)}} + \frac{\mathrm{A}}{a^{2^*}}\right)^{-1},$$

where A is the prior precision of g_j. The matrix A is a K by K scale matrix with diagonal elements equal $1/s_g^2$. After drawing g_j^* from $N(m_{g_j}, V_{g_j})$, we are ready to update the random effect of each document. The random effect of document d for topic j, $q_{d,j}$, also follows a normal distribution with mean m_{q_j} and variance V_{q_j}:

$$m_{q_j} = \frac{\sum_{i=1}^{N_d} (\tilde{H}^*_{di,j} - x_d^T g_j^*)}{N_d + \Sigma_j^{(old)}/s_q^2}$$

$$V_{q_j} = \frac{a^{2^*} s_q^2 \Sigma_{jj}^{(old)}}{N_d s_q^2 + \Sigma_j^{(old)}}.$$

To update document random effects, draw $q^*_{d,j} \sim N(m_{qj}, V_{qj})$ for $d = 1,2,\dots,D$. The above process is repeated for each topic $j = 1,2,\dots,J-1$ to draw every g_j^* and $q^*_{d,j}$. The next step is to draw another a^{2^*} that is associated with g_j^* and $q^*_{d,j}$. This is achieved by drawing $a^{2^{**}} \sim U/\chi^2_{(\sum_d N_d + J - 1)(J-1)}$, where $U = \left[\sum_{d=1}^{D}\sum_{J=1}^{J-1}\sum_{i=1}^{N_d} \frac{(\tilde{H}^*_{di,j} - x_d^T g_j^* - q^*_{d,j})^2}{\Sigma_{jj}^{(old)}}\right] + [\sum_{J=1}^{J-1} g_j^{*T} g_j^* s_g^2] + \sum_{j=1}^{J-1} 1/\Sigma_j^{(old)}$. The new coefficients are $g_j^{(new)} = g_j^*/a^{**}$ and $q_{d,j}^{(new)} = q^*_{d,j}/a^{**}$.

The last step is to draw $\tilde{\Sigma}_j^*$. It can be achieved by drawing individual diagonal element from $\tilde{\Sigma}_j^* \sim \left[1 + \sum_{d=1}^{D}\sum_{i=1}^{N_d} \left(\tilde{H}^*_{di,j} - x_d^T g_j^* - q^*_{d,j}\right)^2\right] / \chi^2_{J-1+\sum_{d=1}^{D} N_d}$. Set $\Sigma_j^{(new)} = \tilde{\Sigma}_j^*/\tilde{\Sigma}_1^*$ and $H_{di,j}^{(new)} = \tilde{H}^*_{di,j}/\sqrt{\tilde{\Sigma}_{11}^*}$. The process is repeated for a fixed number of sweeps to collect samples for subsequent inference task.

5. EXPERIMENTS

The Reuters-21578 dataset is used to evaluate the proposed WNTM. Stopwords were removed in a preprocessing step. The testbed contains 11,771 documents and 782,739 words. There are 22,098 unique tokens. Thirty percent of documents (3,532) are reserved for testing. The Wordnet 2.1 was used to construct concepts adopted in WNTM. Our concept construction procedure produced 155 concepts for WNTM.

A LDA model with symmetric Dirichlet prior is selected as the baseline model, with hyperparameters $\beta = 0.01$ and $\alpha = 50/J$ as suggested by prior studies [15]. Three thousand Gibbs sampling

sweeps were used to train the model. The perplexity was computed by randomly selecting half the words in a testing document and sampling topics conditional on these words. The perplexity for the other half of words was then computed conditional on the sampled topics [16].

A similar procedure is used to train and evaluate WNTM. The document-specific topic tendency $q_{d,j}$ has a prior mean zero and prior precision $\overline{N_d}/80$, where $\overline{N_d}$ is the average document length. The first element in g_j is set to a prior mean that corresponds to equal probability of all topics; the rest of g_j has a zero prior mean and a prior precision of $\sum N_d/4000$. Three thousand sweeps were executed to train a WNTM. The perplexity was computed by sampling q_d conditional on half the words in a testing document and then compute the perplexity of the other half conditional on the sampled q_d.

Table 1 lists concepts with largest average frequencies. It is not surprising to see concepts like proportion.n.01, security.n.04, and funds.n.01 since Reuters-21578 contains many articles about financial market updates. The offer.n.02 is related to articles about mergers, acquisitions, and business unit purchases. The concept fossil fuel.n.01 is related to news about energy. The sum.n.01 and slope.n.01 concepts are related to articles that provides quantitative information for investors. Most of irreverent concepts have been excluded by the concept construction routine.

Table 1. List of Selected Wordnet Concepts

Concept	# Unique Words / Avg. Freq. / Avg. Co-occur. Len.	Words in the Concept (List at Most 10 Words)
proportion.n.01	6/2372.7/1.24	scale, percent, pct, content, rate, percentage.
security.n.04	7/1856.9/1.47	scrip, debenture, share, treasury, convertible, stock, bond.
offer.n.02	9/842.4/1.24	price, question, proposition, prospectus, tender, proposal, reward, bid, special.
fossil fuel.n.01	6/838.8/1.64	oil, jet, gas, petroleum, coal, crude.
funds.n.01	7/806.0/1.15	exchequer, pocket, till, trough, treasury, roll, bank.
sum.n.01	49/736.3/2.21	figure, revenue, pool, win, purse, sales, profits, rent, proceeds, payoff (list truncated).
social science.n.01	5/688.2/1.17	econometrics, politics, economics, finance, government.
slope.n.01	15/616.7/1.42	decline, upgrade, descent, waterside, rise, coast, uphill, steep, brae, fall (list truncated).
gregorian calendar month.n.01	20/612.3/1.51	february, feb, mar, march, august, aug, september, sept, december, dec (list truncated).

Table 2 list the summary statistics of Wordnet concepts. Among the 782,739 words in the testbed, 45% of them are not associated with any Wordnet concepts. The rest of words are mapped to one or more Wordnet concepts. Twenty seven percent of words are mapped to one Wordnet concept. Fewer words are mapped to two or more concepts. Only eight percent of words are map to four or more concepts. The summary statistics suggests that about half of words may benefit from the additional Wordnet concepts.

Table 2. Summary Statistics of Wordnet Concepts

# of Wordnet Concepts Per Word	0	1	2	3	4 or more
Proportion	45%	27%	12%	8%	8%

One practical issue of conducting Gibbs sampling inference is to decide the total number of sweeps to be executed. While the theory of Markov chain Monte Carlo [17] shows that the collected sweeps will eventually converge to the joint posterior, we are looking for a practical guideline to choose the total number of sweeps.

Figure 3 plots the perplexity of a WNTM model at different number of sweeps (solid blue line). The number of topics is set to 25. A LDA model with the same number of topics (dotted red line) is also plotted for comparison. It is clear that the perplexity of the LDA model decreases as more Gibbs sampling sweeps are executed. The LDA perplexity fluctuates around 1160 after passing 1000 sweeps. The WNTM perplexity, on the other hand, continues to decrease until about 3000 sweeps. The perplexity then fluctuates around 530.

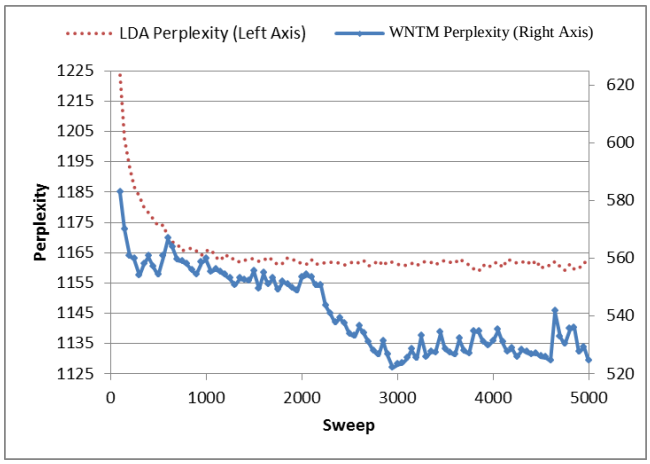

Figure 3. WNTM Perplexity with Different Number of Wordnet Concepts.

Note that while both WNTM and LDA benefit from more sweeps, WNTM perplexity has a larger variance across sweeps while LDA perplexity is more stable. One possible reason is that the collapsed Gibbs sampling of LDA integrated out the document-topic tendency and topic-word distribution and is more efficient. The augmented Gibbs sampling only integrates out the topic-word distribution and relies on additional augmented variables to approximate the joint posterior of regression parameters and latent topics. More sweeps are requires for WNTM to converge compared to the LDA. The total number of sweeps is set to 3,000 in the following experiments.

5.1 Estimated WNTM Topics

We report the estimation result of a WNTM model with 25 topics. Only selected topics are reported to save space. Table 3 summarizes the estimated WNTM topic "Statement." The topic name was assigned manually based on top keywords and related concepts. The second row lists estimated top keywords and the third row lists Wordnet concepts with largest coefficients, together with their high-probability words. The first two concepts, commercial document.n.01 and proceedings.n.01, are associated with large coefficients (5.43 and 4.29). Many words in these two concepts appear in the top keywords of this topic. In fact, 10 of the 15 top keywords are from these two concepts. The third concept, relationship.n.03, has a moderate coefficient (0.86) and contributes one word to the top keyword list. Note that some top

keywords such as coupon and usair do not appear in concepts with large coefficients. These keywords are included by WNTM based on the co-occurrence pattern in a similar manner as the LDA. The last row lists the top keywords of the LDA topic that is the most similar (based on symmetric Kullback-Leibler divergence) to the WNTM topic. One point the worth mentioning is the size of regression coefficients. Since the WNTM model is normalized to have $\Sigma_1 = 1$, a coefficients with a value larger than 2 has a very strong inference on the latent topic.

It is clear from the above observation that WNTM is able to form topics by combining related Wordnet concepts and word co-occurrence patterns. The matching LDA topic shows a very different keywords compared to the one estimated by WNTM, suggesting that WNTM may capture a latent topic that is otherwise missed by the LDA.

Table 3. WNTM Topic "Statement."

WNTM Topic "Statement"
Top Keywords: estimate statement bill account action order coupon intervention review case usair accounting suit pass transfer
Wordnet Concepts: commercial document.n.01 (5.43)* **estimate statement bill account order** proceeding.n.01 (4.29) * **action intervention review case suit** relationship.n.03 (0.86) * **account** hold restraint trust confinement advantage.n.01 (0.69)* **account** leverage profitability expediency privilege fact.n.01 (0.51) * **case** observation score specific item
Matching LDA Topic
Top Keywords: ct net loss shr profit rev note oper avg shrs mths qtr sales exclude gain

*Estimated coefficients for Wordnet concepts.

One important difference between WNTM and an LDA model with Dirichlet forest prior is that the WNTM determines whether the Wordnet concepts are suitable for the training corpus by selecting a set of regression coefficients. The Dirichlet forest prior, on the other hand, has a set of fixed parameters that determines whether two words should be grouped in the same topic. In other words, WNTM can choose to "turn off" some Wordnet concepts by assigning coefficients close to zero. It is not allowed if Dirichlet forest prior is used.

Table 4 summarizes the estimated topic "Earnings" from the same WNTM model. This topic has only one Wordnet concept (advantage.n.01) with positive coefficient. This concept contributes two of the top keywords. Other concepts have negative coefficients. Negative coefficients allow WNTM to avoid associating these Wordnet concepts with the topic. For example, the word "congress" belongs to the concept legislature.n.01 with a coefficient of -0.06. Based on the DGP presented in the previous section, this word contributes to a smaller $H_{di,j}$, which gives other topics a better chance to dominate. As a result, a Wordnet concept with a negative coefficient can be interpreted as "not" associating the concept with the topic.

Table 4. WNTM Topic "Earnings."

WNTM Topic "Earnings"
Top Keywords: mln ct net loss dlrs shr profit rev note year gain oper include avg shrs
Wordnet Concepts: advantage.n.01 (3.59) * **profit gain** good leverage preference subject.n.01 (-0.02) * puzzle head precedent case question push.n.01 (-0.03) * pinch crunch nudge mill boost legislature.n.01 (-0.06) * diet congress house senate parliament
Matching LDA Topic
Top Keywords: mln note net stg include profit tax extraordinary pretax operate full item making turnover income

*Estimated coefficients for Wordnet concepts.

Table 5 summarizes a WNTM topic that combines Wordnet concepts such as quantity.n.03, part.n.90, word time.n.01, economic process.n.01, and merchandise.n.01. All of the top keywords are from these five concepts, which provide an intuitive idea about the nature of this topic. Note that the matching LDA topic looks quite different from the one estimated by WNTM.

Table 5. WNTM Topic "Market Update."

WNTM Topic "Market Update"
Top Keywords: week total end product period average amount demand supply line inflation term shipment number release
Wordnet Concepts: quantity.n.03 (5.33) * **total product average amount term** part.n.09 (4.66) * **end period** factor top beginning work time.n.01 (4.38) * **week** turn hours shift turnaround economic process.n.01 (4.34) * **demand supply inflation** consumption spiral merchandise.n.01 (4.26) * **line shipment number release** inventory cargo
Matching LDA Topic
Top Keywords: union south area spokesman city ship strike port worker africa line week affect state southern

*Estimated coefficients for Wordnet concepts.

Table 6 summarizes the WNTM topic "Macroeconomics." Note that all Wordnet concept coefficients are negative, suggesting that this topic is different from all included Wordnet concepts. Moreover, the matching LDA topic looks quite similar to the WNTM topic, suggesting that both models identified a common topic.

Table 6. WNTM Topic "Macroeconomics."

WNTM Topic "Macroeconomics"
Top Keywords: dollar market currency west yen economic dealer central growth cut japan economy expect policy interest
Wordnet Concepts: semite.n.01 (-0.03) * palestinian arab saudi omani arabian rational_number.n.01 (-0.11) * thousandth fraction fourth eighth half

seed.n.01 (-0.12) *
soybean coffee hazelnut nut cob
fact.n.01 (-0.22) *
observation score specific item case
Matching LDA Topic
Top Keywords:
dollar currency yen west exchange market rates japan dealer central german germany intervention finance paris

*Estimated coefficients for Wordnet concepts.

To summarize, WNTM identify latent topics based on Wordnet concepts and word co-occurrence structures. The identified topics may be related to several Wordnet concepts with large coefficients. Words not in these concepts may also receive high conditional probability if supported by training data. It is possible to have topics with all negative Wordnet concept coefficients. This type of topics is unrelated to the Wordnet concepts provided to WNTM.

5.2 The Effect of Wordnet Concepts

To further understand the impact of Wordnet concepts to topic models, we conducted two experiments that compared WNTM and LDA in terms of perplexity. The first experiment estimated WNTM models with different number of Wordnet concepts. The number of latent topics was fixed at 25. The perplexity of a LDA model with the same number of latent topics was estimated for comparison.

Figure 4 plots the estimation results. The LDA model has a perplexity of 1150.1, which is slightly higher than that of a WNTM model with no Wordnet concept (perplexity = 965.7). The WNTM perplexity steadily decreased as new concepts were included. The perplexity dropped to 496.4, about half the WNTM perplexity with no Wordnet concept, when 54 concepts were included. The perplexity dropped to 246.2 when all 155 Wordnet concepts were included. It is clear from the results that adding Wordnet concepts to the WNTM reduces perplexity and improves prediction ability.

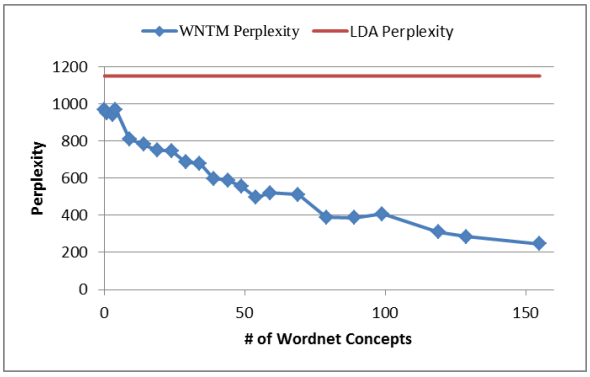

Figure 4. WNTM Perplexity with Different Number of Wordnet Concepts.

The second experiment is designed to understand the interaction between Wordnet concepts of the number of latent topics. In this experiment, the WNTM was provided with 25 Wordnet concepts. WNTM models with different number of latent topics were estimated. The perplexity of LDA model with matching number of topics was also computed for comparison purpose.

Figure 5 plots the estimation results. Both LDA and WNTM have a perplexity steadily decreases as the number of latent topics increases. The perplexity of WNTM, however, decreases faster compared to that of LDA especially when the number of topics is small. The WNTM perplexity curve, as a result, is consistently lower compared to the LDA perplexity curve. The overall pattern suggests that Wordnet concepts are the most effective in decreasing perplexity when there are enough number of latent topics. When the number of topics are large compared to the number of included Wordnet concepts, the marginal benefit levels out.

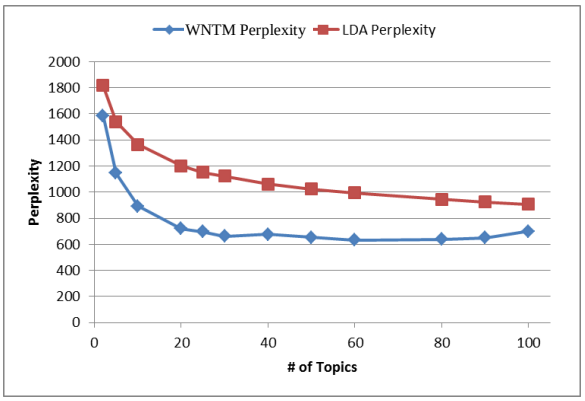

Figure 5. Comparing WNTM Perplexity and LDA Perplexity with Different Number of Topics.

Another interesting observation is that the gap between LDA and WNTM decreases as the number of topics increases. One interpretation is that additional topics in LDA make up the lack of Wordnet concepts. It is, nonetheless, a different question to decide whether a topic model with larger amounts of topics can provide additional benefit for human interpretation.

6. CONCLUSIONS

This paper reports a Wordnet-enhanced topic model (WNTM) that combines Wordnet concepts constructed from noun synsets to improve topic modeling outcomes. WNTM adopts a multi-nominal probit prior to link Wordnet concepts to the prior probability of individual words in a document. A document-specific topic tendency is also included so that each document is able to develop its own mixture of topics. The prior probability of a word is determined jointly by the document-specific topic tendency and the Wordnet concepts it belongs to.

One important difference between WNTM and existing topic models that incorporate contextual information is the flexibility to "turn-off" some Wordnet concepts if preferred by the training data. WNTM achieves this by assigning coefficients that are negative or close to zero. The flexible structure of WNTM allows forming a topic by combining multiple Wordnet concepts and based on the word co-occurrence structure. We developed an augmented Gibbs sampling algorithm to estimate WNTM. Experiments show that Wordnet concepts are useful in discovering latent topical structures.

WNTM shows promising results in adopting word-level features into topic models. One future direction is to apply WNTM under different context such as the health insurance claim data. The drug orders and diagnoses have rich token level features that could be included to improve discovering the latent structure. The other direction is to extend WNTM for documents from multiple sources.

7. ACKNOWLEDGMENTS

This work was supported in part by the National Science Council of Taiwan under the grant NSC100-2410-H-002-025-MY3.

REFERENCES

[1] Blei, D.M., 2012. Introduction to probabilistic topic models. *Communications of the ACM Forthcoming*.

[2] Blei, D.M. and Lafferty, J.D., 2006. Correlated Topic Models. In *Proceedings of the Neural Information Processing Systems (NIPS)* (2006).

[3] Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993-1022. DOI= http://dx.doi.org/http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

[4] Cameron, A.C. and Trivedi, P.K., 2005. *Microeconometrics: Methods and Applications*. Cambridge.

[5] Gelfand, A.E., 2000. Gibbs Sampling. *Journal of the American Statistical Association 95*, 452, 1300-1304. DOI= http://dx.doi.org/10.2307/2669775.

[6] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., 2003. *Bayesian Data Analysis*. Chapman & Hall/CRC.

[7] Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101*, 5228-5235.

[8] Hammersley, J.M. and Handscomb, D.C., 1964. *Monte Carlo Methods*. Halsted, New York.

[9] Imai, K. and Dyk, D.a.V., 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics 124*, 311-334.

[10] Mcculloch, R. and Rossi, P.E., 1994. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics 64*, 1-2, 207-240.

[11] Meng, X.-L. and Dyk, D.a.V., 1999. Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation. *Biometrika 86*, 2, 301-320.

[12] Miller, G.A., 1995. WordNet: A lexical database for English. *Communications of the ACM 38*, 11, 39-41.

[13] Mimno, D. and Mccallum, A., 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence (UAI)*.

[14] Minka, T.P., 2000. *Estimating a Dirichlet distribution.*

[15] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M., 2010. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst. 28*, 1, 1-38. DOI= http://dx.doi.org/10.1145/1658377.1658381.

[16] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T., 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the KDD* (Seattle, Washington, USA2004).

[17] Tierney, L., 1994. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics 22*, 4, 1701-1728. DOI= http://dx.doi.org/10.2307/2242477.

[18] Wang, X. and Mccallum, A., 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the KDD* (Philadelphia, Pennsylvania, USA2006), 424-433.