Who Wrote This? Textual Modeling with Authorship Attribution in Big Data

Naruemon Pratanwanich
Computer Laboratory
University of Cambridge
William Gates Building
JJ Thomson Avenue, CB3 0FD
Email: np394@cam.ac.uk

Pietro Lio’
Computer Laboratory
University of Cambridge
William Gates Building
JJ Thomson Avenue, CB3 0FD
Email: pl219@cam.ac.uk

Abstract—By representing large corpora with concise and meaningful elements, topic-based generative models aim to reduce the dimension and understand the content of documents. Those techniques originally analyze on words in the documents, but their extensions currently accommodate meta-data such as authorship information, which has been proved useful for textual modeling. The importance of learning authorship is to extract author interests and assign authors to anonymous texts. Author-Topic (AT) model, an unsupervised learning technique, successfully exploits authorship information to model both documents and author interests using topic representations. However, the AT model simplifies that each author has equal contribution on multiple-author documents. To overcome this limitation, we assume that authors give different degrees of contributions on a document by using a Dirichlet distribution. This automatically transforms the unsupervised AT model to Supervised Author-Topic (SAT) model, which brings a novelty of authorship prediction on anonymous texts. The SAT model outperforms the AT model for identifying authors of documents written by either single authors or multiple authors with a better Receiver Operating Characteristic (ROC) curve and a significantly higher Area Under Curve (AUC). The SAT model not only achieves competitive performance to state-of-the-art techniques e.g. random forests but also maintains the characteristics of the unsupervised models for information discovery i.e. word distributions of topics, author interests, and author contributions.

I. INTRODUCTION

Characterizing the content of documents is generally conducted in order to search, organize, and classify a large collection of documents effectively. Some documents often come with a variety of side information such as authors, keywords, and publishers, while such information is missing in others and needs to be predicted. Authorship attribution involves assigning authors to anonymous texts, which plays an important role in areas such as criminal investigation, social science, text analysis, cognitive systems, to name but a few. Since different authors have different interests in writing, learning their interests based on textual data brings many advantages such as matching authors and reviewers in publication systems, recommending favorite media to users, and promoting personalized advertisement. However, the complication in this issue is the increasing dimensionality in the number of words and authors used in the analysis.

Typically, text data can be represented as a count bag-of-words vector. As the size of vocabulary in the corpus is increasing, the task is to learn the low frequency count data in the high-dimensional setting for information discovery and prediction. Conventional machine learning approaches to authorship prediction have largely relied on discriminative modeling techniques that depend crucially on a variety of features such as word functions, word length distributions, and word contents [1]. The main drawback is that they generate a “black box” that makes it hard to understand why they give high performance on prediction.

To discover the meaningful structures underlying documents, probabilistic generative models which employ the abstract definition of topics as a fundamental concept to generate words have gained popularity for document analysis as unsupervised learning techniques. In topic-based generative models, a document is described by a particular topic proportion, where a topic is defined as a distribution over words. After latent Dirichlet allocation (LDA), a mixed-membership topic model, was introduced [2], many studies have proposed a great number of model variations [3]–[7]. The primary goal of such extensions is to incorporate side information or meta-data together with words in the texts for better characterization of the content of documents in unsupervised learning settings. In contrast, the author model proposed by McCallum [8] uses authorship information of each document to model author interests with word distributions instead of the latent topics. Accordingly, the model assumes that a word in a document is written by an author of the document. Combining both models, the author-topic (AT) model represents documents and authors by topic distributions [9]. Thus, characterizing author interests and modeling documents can be achieved simultaneously with concise, meaningful representations.

Regarding authorship attribution, the original AT model focuses on the unsupervised learning environment where words are assigned into topics and author interests better than those being learned by the LDA model and the author model [9]. Nonetheless, recent studies showed that the topic-based generative models have been effectively applied for authorship prediction [10], [11]. The AT model is found superior to the LDA model to predict authors for anonymous texts [11]. In addition, the performance of authorship prediction can be improved if the model distinguish words generated from a topic that belongs to either documents or author interests, resulting in the disjoint author-document topic (DADT) model [11].
However, the DADT performance strongly depends on the specification of many parameters regarding to the ratio between documents and author words, which are different from corpus to corpus. It is also noted that only single-author texts have been studied. Moreover, even though the LDA model extensions can accommodate various types of the response variable for prediction, they rely on the assumption that the response variable arises from the same underlying topics that generate words. This hypothesis is not suitable for modeling authorship where authors can have their own topics of interests differ from documents they have written.

In this study, we have developed a new generative method, Supervised Author-Topic (SAT) model, with the aim of identifying authorship of anonymous documents and concurrently extracting author interests based on topic-word distributions. Particularly, we introduce a new definition of author contributions modeled by using an additional Dirichlet distribution layer in addition to the two Dirichlet priors that are used to model topics and author interests. With the novelty of putting the Dirichlet prior on author contributions in a document level, the SAT model can automatically assign authors to anonymous documents written by one or multiple authors.

### II. Supervised Author-Topic Model

Even though the AT model takes advantage of authorship information to model documents and author interests, it does not provide an explicit framework for author prediction of anonymous texts. Moreover, the AT model assumes the equity of author contributions for documents written by many authors using a uniform distribution. To overcome these two limitations, the SAT model uses an additional Dirichlet distribution to model author contributions. Therefore, a document is represented by a particular proportion of author contributions, where each author is described by a topic distribution and each topic is defined as a distribution over vocabulary. With this implementation, the SAT model can also perform a straightforward prediction of authors on new anonymous documents. In the following subsections, we illustrate a mathematical description, an inference method, and an approach for author prediction of our proposed model, discussing its structure in comparison to the AT model. The notations are summarized in Table I.

#### TABLE I: Notations used in the SAT model.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V, T, A, D$</td>
<td>Number of distinct words, topics, authors, and documents.</td>
</tr>
<tr>
<td>$N_d, A_d$</td>
<td>Number of words and authors in a document $d$ respectively.</td>
</tr>
<tr>
<td>$w_{d,n}, x_{d,n}, z_{d,n}$</td>
<td>A word and its topic and author assignment of a document $d$ at position $n^{th}$.</td>
</tr>
<tr>
<td>$\phi_t$</td>
<td>An $V$-dimensional probability vector of word distribution under a topic $t$, per-topic word distribution.</td>
</tr>
<tr>
<td>$\theta_a$</td>
<td>An $A$-dimensional probability vector of topic distribution under an author $a$, per-author topic distribution.</td>
</tr>
<tr>
<td>$\psi_d$</td>
<td>An $A$-dimensional probability vector of author distribution under a document $d$, per-document author distribution.</td>
</tr>
<tr>
<td>$a_{d,a}$</td>
<td>An $A$-dimensional binary vector of authors under a document $d$, where $a_{d,a} = 1$ when $a$ is an author of the document $d$, and 0 otherwise.</td>
</tr>
<tr>
<td>$C^{TV}$</td>
<td>A $T \times V$ count matrix of words for each topic.</td>
</tr>
<tr>
<td>$C^{AT}$</td>
<td>A $A \times T$ count matrix of topic assignments for each author.</td>
</tr>
<tr>
<td>$C^{DA}$</td>
<td>A $D \times A$ count matrix of author assignments for each document.</td>
</tr>
<tr>
<td>$\alpha, \beta, \gamma$</td>
<td>Concentration parameters of Dirichlet distributions.</td>
</tr>
</tbody>
</table>

Fig. 1: Graphical representation of the Supervised Author-Topic (SAT) model, where shaded nodes are observed data, transparent nodes are model parameters and latent variables, the others are hyperparameters. All notations are described in Table I.

#### A. Model description

Dirichlet distribution is a distribution over distributions. Thus, we posit three Dirichlet priors in order to model the author contribution per document, the topic interest of each author, and the distribution of words per topic. The generative process of our SAT model for learning a corpus is fully described as follows:

1) For each topic $t = 1, \ldots, T$:
   a) $\phi_t \sim \text{Dirichlet}(\alpha)$;
2) For each author $a = 1, \ldots, A$:
   a) $\theta_a \sim \text{Dirichlet}(\beta)$;
3) For each document $d = 1, \ldots, D$:
   a) $\psi_d \sim \text{Dirichlet}(\gamma \cdot a_d)$;
4) For each word $n = 1, \ldots, N_d$ and $d = 1, \ldots, D$:
   a) $x_{d,n} \sim \text{Multinomial}(\psi_d)$;
   b) $z_{d,n} \sim \text{Multinomial}(\theta_{x_{d,n}})$;
   c) $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$.

For each position $n$ in each document $d$, the process begins with choosing an author $x_{d,n}$ based on their contributions $\psi_d$ modeled by an asymmetric Dirichlet distribution with a parameter $\gamma a_d$. The sparsity pattern of this parameter is the same as that of the observed author vectors $a_d$ where $a_{d,a} = 1$ when $a$ is an author of the document $d$, and 0 otherwise. Then, a topic $z_{d,n}$ is selected from the interests of the author $\theta_{x_{d,n}}$ modeled by a distribution drawn from a symmetric Dirichlet with a parameter $\beta$. Finally, a word is picked up from the distribution over words of that topic $\phi_{z_{d,n}}$.

The main differences from the AT model are the assumption of author contributions modeled by a Dirichlet distribution in Step 2a and the consequence of author assignment in Step 4a. This generalization on the author contributions also
brings the extra ability to predict authorship of anonymous documents. Figure 1 exhibits the graphical representations the SAT model.

B. Inference of model parameters and latent variables

Given a training set of documents where words and authors of each document are observed, we perform an inverse of the generative process in order to make inference on the latent variables and the model parameters \((x_{d,n}, z_{d,n}, \phi_t, \theta_a, \psi_d)\). Since the exact joint posterior distribution of the model parameters is intractable, approximate algorithms such as variational EM (deterministic) and Gibbs sampling (stochastic) have achieved the parameter estimation. We apply a collapsed Gibbs sampling to find only the joint posterior distribution of \(x_{d,n} \) and \(z_{d,n} \) while \(\phi_t, \theta_a, \psi_d\) are integrated out (Equation (1)). Due to the conjugacy of the Dirichlet and multinomial models, the collapsed Gibbs sampling can be implemented through the updating of count matrices \(C^{TV}, C^{AT}, \) and \(C^{DA} \) which denote the number of words for each topic, the number of topics for each author, and the number of authors for each document respectively. The steps of posterior derivation can be analogous to those described in the LDA model [12]. In each iteration of the collapsed Gibbs sampler, the probability to assign an author \(a \) and a topic \(t\) to a word \(w_{d,n}\) is updated by Equation (1). Note that the subscript \(\{d,n\}\) indicates that all positions excluding the current position \(n\) in the current document \(d\) are regarded.

\[
P(x_{d,n} = a, z_{d,n} = t | w_{d,n} = w, \Phi_t, \Theta_a, \Psi_d) \propto \frac{C^{TV}_{t,w} + \alpha}{\sum_{w'} C^{TV}_{t,w'} + V \alpha} \cdot \frac{C^{AT}_{a,t} + \beta}{\sum_{t'} C^{AT}_{a,t'} + \beta} \cdot \frac{C^{DA}_{d,a} + \gamma d,a}{\sum_{a'} (C^{DA}_{d,a'} + \gamma d,a')} \cdot \prod_{x \in \{d,n\}} \psi_{d} \quad (1)
\]

Intuitively, the probability of assigning a word \(w\) to a topic \(t\) written by an author \(a\) depends on three probabilities - how likely the word \(w\) belongs to the topic \(t\), how likely the topic \(t\) is written by the author \(a\), and how likely the author \(a\) contributes to the document \(d\). All count matrices \((C^{TV}, C^{AT}, C^{DA})\) in Equation (1) are used in the Gibbs sampler to represent the aforementioned probabilities. Accordingly, the contribution of an author in a document is proportional to the number of author assignments to words in the document with respect to all topics.

Owing to a sufficient statistic, the point estimates of the model parameter values \((\phi_t, \theta_a, \text{and } \psi_d)\) are finally calculated as in Equation (2)-(4):

\[
\Phi = \frac{C^{TV} + \alpha}{\sum_{w'} C^{TV}_{t,w'} + V \alpha} \quad (2)
\]

\[
\Phi = \frac{C^{AT} + \beta}{\sum_{t'} C^{AT}_{a,t'} + \beta} \quad (3)
\]

\[
\Psi = \frac{C^{DA} + \gamma d,a}{\sum_{a'} (C^{DA}_{d,a'} + \gamma d,a')} \quad (4)
\]

Our R implementation of the pathway-based Gibbs sampling is available upon request to the corresponding author.

C. Authorship attribution

Given an unseen anonymous document and the learned model, we aim to predict a set of authors who wrote the document. By folding the initial parameters of the new document into the learned model, the SAT model identifies authors and their contributions using the aforementioned Gibbs sampler. At the initialization step, we randomly generate the count matrices \(C^{TV}, C^{AT}, \) and \(C^{DA}\) of the new document and integrate with those from the latest iteration of the Gibbs sampler obtained in the training stage. Since the authors of the new document are not observed, every entry in the binary vector \(a_d\) is set to one, that is, every author has equal possibility to write the document. The sampling process is similar to the aforementioned inference algorithm, but only samples for the new document are drawn. Hence, the chain would converge in tens of iterations. We finally identify the potential authors of the new document from the probability distribution \(\psi_{d}\) estimated by the Dirichlet distribution.

Likewise, we apply the predictive Gibbs sampler which relies on the probabilistic structure of the AT model. However, there is no parameter of the AT model that we can interpret directly for author identification. Based on the latent author assignment, we therefore count the number of author assignments in the entire document and normalize so as to indicate the possibility of each author belonging to the document (Equation (5)).

\[
\tilde{\psi}_d = P(a \in a_d | \tilde{w}_d, M) = \frac{\sum_{n=1}^{N_d} \delta(\tilde{x}_{d,n} = a)}{N_d} \quad (5)
\]

With a straightforward interpretation on the inferred contributions \((\psi_d)\), we can identify a subset of authors given an anonymous text. We remark that other methods that use the inferred variables from the AT model for author prediction may yield better results, but we use the results that are confined to its assumption of the data generation for the model comparison purpose.

III. RESULTS AND DISCUSSION

We collected 2484 articles of the NIPS conferences from 1987 to 2003. The corpus was made up of 2865 unique authors, 14036 distinct words and 3280697 word tokens in total [13]. Figure 2 depicts the histograms of the number of authors (a) and words (b) of the NIPS data. Of those, 90% were used as a training set and the rest as a test set, where every author in the test set appeared at least once in the training set.
For the Gibbs sampler settings in this paper, the smoothing hyper-parameters of both models are conventionally fixed to the same values at 0.01, 50/T, and 0.01 for α, β, γ respectively. However, we initialize the count matrices of a Markov chain with random initial assignments. To ensure that the Gibbs sampler converged to a stationary state, 2000 iterations were run along with convergence analysis.

A. Convergence analysis and model fitness

To fit a probabilistic model, we aim to achieve the highest likelihood on the training data set. We used the perplexity [2], which is a monotonically decreasing function of the data likelihood (Equation (6)), as a metric of fitness for comparison between the two models.

\[ \text{perplexity} = \exp\left(-\frac{\sum_{d=1}^{D} \log p(w_d|\alpha, \beta, \gamma, \mathbf{a}_d)}{\sum_{d=1}^{D} N_d}\right) \]  

Specifically, we computed the perplexity at each iteration of the Gibbs sampler obtained from the SAT model and the AT model. The 100-topic models were selected to display as a representative in Figure 3 (a). Varying number of topics, we calculated the perplexity at the stationary state (Figure 3 (b)). A lower perplexity value suggests a relatively better model fitness. Notably, both models reach the stationary state quickly after 1000 iterations. However, the SAT model achieves the invariant state at the lower perplexity, confirming that the concept of contributions can improve the fitness for the research article data. Even though increasing the number of topics can enhance the model fitness, it should be traded off with the computational cost, especially in the author prediction. As a larger number of topics causes the Gibbs sampler to run slower per iteration, we explored the performance of both models based on 100 topics for further evaluation on author interests and authorship attribution.

In order to investigate the model convergence in the case of multiple-author documents, we excluded those documents that did not meet the condition and trained the models again. Figure 3 shows that the perplexity at the stationary state of the AT model seems unchanged, while the SAT yields a better fitness. This could be because the AT model executes single-author documents in the same way as learning multiple-author ones, while the SAT model treats these two types of texts differently by implementing author contributions. Despite the modest difference in the perplexity of model fitness, the SAT model has a distinguished performance over the AT model in the authorship prediction.

B. Topics, author interests, and author contributions

In this section, we display examples of the inferred \( \phi_t, \theta_a, \) and \( \psi_d \) that indicate the word distribution in a topic \( t \), the topics in which an author \( a \) is interested, and the degrees of all authors contributing in a document \( d \). The selected probability distributions of these parameters were obtained from a single chain at the last iteration (2000\( T^{th} \)) of the Gibbs sampler. Table II shows the most interesting topics of selected five authors.
TABLE II: Representatives of author interests obtained from the SAT model. Each author is shown with the top interesting topic IDs that have probabilities more than the threshold of 0.01 (1/T). The topic IDs also correspond to those shown in Table III.

<table>
<thead>
<tr>
<th>Author: Cole R</th>
<th>Author: Agin P</th>
<th>Author: Ghabramani Z</th>
<th>Author: MacKay D</th>
<th>Author: Bishop C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topics</td>
<td>Probability</td>
<td>Topics</td>
<td>Probability</td>
<td>Topics</td>
</tr>
<tr>
<td>64</td>
<td>0.059</td>
<td>22</td>
<td>0.584</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>0.070</td>
<td>78</td>
<td>0.152</td>
<td>63</td>
</tr>
<tr>
<td>52</td>
<td>0.098</td>
<td>22</td>
<td>0.183</td>
<td>22</td>
</tr>
<tr>
<td>78</td>
<td>0.071</td>
<td>10</td>
<td>0.093</td>
<td>86</td>
</tr>
<tr>
<td>87</td>
<td>0.036</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE III: Representatives of 12 topics obtained from the 100-topic SAT model. The most frequent 10 words are shown in each topic.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>Probability</td>
<td>Words</td>
<td>Probability</td>
</tr>
<tr>
<td>experts</td>
<td>0.028</td>
<td>classifier</td>
<td>0.045</td>
</tr>
<tr>
<td>basis</td>
<td>0.024</td>
<td>class</td>
<td>0.039</td>
</tr>
<tr>
<td>expert</td>
<td>0.021</td>
<td>training</td>
<td>0.031</td>
</tr>
<tr>
<td>gating</td>
<td>0.019</td>
<td>classifiers</td>
<td>0.026</td>
</tr>
<tr>
<td>network</td>
<td>0.018</td>
<td>classes</td>
<td>0.016</td>
</tr>
<tr>
<td>radial</td>
<td>0.017</td>
<td>feature</td>
<td>0.015</td>
</tr>
<tr>
<td>networks</td>
<td>0.017</td>
<td>pattern</td>
<td>0.014</td>
</tr>
<tr>
<td>mixture</td>
<td>0.016</td>
<td>decision</td>
<td>0.012</td>
</tr>
<tr>
<td>gaussian</td>
<td>0.013</td>
<td>nearest</td>
<td>0.011</td>
</tr>
</tbody>
</table>

using the probability threshold is 0.01 (1/T). The probabilities of each topic conditioned on a given author can suggest the ranking of author interests, which can be used as a similarity measure between authors. Conditioned on these topics, the words that were most frequently generated are illustrated in Table III.

Table II and Table III illustrate examples of author interests and topic distributions discovered by the model on the NIPS data. To begin with, Topic 58 is related to Bioinformatics while Topic 64 is pertinent to speech recognition, which characterize Cole R from Agin P. From these results, we can conclude that Cole R is interested in applying classification techniques (Topic 11) for speech recognition, and Agin P focuses also on machine learning (Topic 78) but applies to Bioinformatics. We can also see that the other three authors, whose interests lie on the core of machine learning techniques, have many sub areas of interest in common, each of which has a set of different dominant words that can be manually annotated (Table III).

Similar to the AT model [9], there were approximately 30% of all topics that are generic such as Topic 78. This could be dominated by the nature of the NIPS corpus. These topics may be used as the representative topics to compare across different types of corpora in order to gain more insights into the characteristics of each database.

It is problematic to make a direct evaluation of the inferred contributions $\psi_d$ based on the order of authorship appearance in documents. In other words, it is not necessary that the preceding authors appearing in the document contributes on writing more than the authors coming later in the order. For example, the first author sometimes conducts the experiments but the one who wrote the paper is the last author.

In order to validate our assumption of the author contributions we therefore created a pseudo document by combining six single-author documents - one document from Abu-Mostafa, Y (0.17%), two documents from Anastasios_T
contains six single-author documents - one document from Abu-Mostafa, and three documents from Linsker R (0.50%), and three documents from Linsker R (0.50%), and tested how much the SAT model can describe the contributions of those authors. Figure 4 (a) shows that our model not only predicts those writers from 2865 candidates correctly, but also ranks the contributions precisely.

More importantly, the inferred contribution probabilities modeled by a Dirichlet distribution prove their benefit to the authorship prediction. Figure 4 (b-c) exemplifies that the characteristic of the uniform assumption in the AT model fails to perform the authorship prediction based on its author assignment in Equation (5) using the fold-in Gibbs sampler. In contrast, the Dirichlet distribution modeled by the SAT model can distinguish the true authors from 2865 candidate authors given an anonymous documents.

\[ TPR_k = \frac{TP_k}{A} \] (7)
\[ FPR_k = \frac{D \times k - TP_k}{D \times A} \] (8)

where \( TP_k \) is the number of true positives in the top-\( k \) ranks and other notations are described in Table I.

As a result, the receiver operating characteristic (ROC) curves of each method for both experiment scenarios as shown in Figure 5 (top) as well as the area under the ROC curve (AUC) in Table IV. Furthermore, for the single-author test documents, the SAT model shows the best performance, while the random forest with five trees comes for the first place.

<table>
<thead>
<tr>
<th>No. of authors</th>
<th>SAT</th>
<th>AT</th>
<th>RF: single tree</th>
<th>RF: 5 trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>0.76</td>
<td>0.71</td>
<td>0.57</td>
<td>0.71</td>
</tr>
<tr>
<td>Multiple</td>
<td>0.67</td>
<td>0.62</td>
<td>0.55</td>
<td>0.71</td>
</tr>
</tbody>
</table>
for those documents written by multiple authors. Even though the predictive power of random forests is high, even higher when growing more trees, the main disadvantage is that they cannot provide information discovery i.e. author interests and underlying topics. Noticeably, of all test documents, AUC scores of the SAT model were significantly greater than the AT model for both scenarios (p-value < 0.05).

Indeed, we are interested in the inferred contributions at the top-K ranks for author identification, regardless the order at the lower ends of the probability vector. In this case, we therefore emphasize the percentage of true positives at the top 35% of the rank as shown in Figure 5 (bottom). The SAT model is superior to the AT model for both scenarios. Importantly, the SAT model identified true authors at the first rank approximately 25%, whereas the AT model achieved with very tiny percentage (6%). This difference also occurs in the multiple-author case.

IV. CONCLUSION

We have presented the Supervised Author-Topic (SAT) model in order to infer topic-specific word distributions, author interests and author contributions. We also derived the posterior of all model parameters and implemented the collapsed Gibbs sampling for inference on the model parameters and for authorship attribution to anonymous documents.

The SAT model can be viewed as a generalized model of the AT model, a successful unsupervised learning technique for analyzing documents and author interests based on topic representations. Since the AT model uses a uniform distribution to assign an author to a word, this implies that the model holds the assumption of equal contributions among the authors on a document. As a result, it has limited performance on authorship prediction in a supervised learning setting. We relax this restriction by imposing an additional Dirichlet distribution as a prior on author contributions. The underlying assumption

Fig. 5: ROC curves of authorship prediction (top) and the percentage of true positives as a function of top-k ranks (bottom) in single-author documents (a), multiple-author documents (b).
is that different authors contribute to a document with different degrees. Furthermore, the AT model can be considered a special case when the parameter of the Dirichlet prior on contributions in the SAT model is set to one.

With regard to author prediction, we define the contribution of an author as the number of words in the given document assigned to the author. Given an anonymous document, the SAT allocates each author with the equal contribution. Having observed more words with the learned SAT model, the Gibbs sampler updates the contribution probability according to word-author assignments. Such probability distribution across all authors at the invariant stage can indicate the likeliness of those who wrote the document. Consequently, the SAT model provides an explicit framework for authorship attribution to anonymous documents with knowledge discovery of topics and author interests.

Integrated with a scoring system, the inferred contributions may be used for measuring the impact of publication articles of scholars. Moreover, executing the SAT model on single sections e.g. introduction, methodology, and conclusion separately may reveal the dynamic of writing within the research articles. Performing the SAT model with respect to the level of the authors, e.g. full professor, assistant professor, and PhD student at the time of submission would also provide insights into the relationships between the academic hierarchy and the writing contribution. In addition, this research could be in theory used with sentiment analysis in the message content in the facebook/twitter of a person with suicide intention [15], [16]. This would introduce the ability to detect the degrees contributions of the suspects.

The SAT model can also be applied on any other kind of dyadic data that have the same relationship as authors and words. For example, analyzing clinical data annotated by physicians and nurses such as prescriptions and diagnostic texts for individual patients using the SAT model can establish the system of “patient like me” (http://www.patientslikeme.com). This helps patients compare treatments, symptoms and experiences with people like them and take control of their health. In addition, this system allows tracking their health by characterizing them into the groups like them over time and also contributes to research that can advance medicine for all.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their useful suggestions for improving the paper. NP acknowledges the Royal Thai Government Scholarship.

REFERENCES