A Smoothed Latent Dirichlet Allocation Model With Application to Business Intelligence

Zhihua Wei,Rui Zhao,Ying Wang,Duoqian Miao,
Wenbo Yuan
Department of Computer Science and Technology,
Tongji University
Shanghai, China
zhihua_wei@tongji.edu.cn, zhaorui1@126.com,
wying@hfuu.edu.cn, dqmiao@tongji.edu.cn,
yobo.ywb@gmail.com

Rui Zhao
The Third Research Institute of the Ministry of Public Security
Shanghai, China

***Abstract*—As a kind of intelligent component, text classification plays an important role in Business Intelligence (BI) application such as client opinion classification, market feedback analysis and so on. Latent Dirichlet Allocation (LDA) model, which is a kind of excellent text representation model, has been widely used in various document processing applications. However, its performance is affected by the data sparseness problem. Existing smoothing techniques usually make use of statistic theory to assign a uniform distribution to absent words. They don't concern the real word distribution or distinguish between words. In this paper, a method based on Tolerance Rough Set Theory (TRST) is proposed, which makes use of upper approximation and lower approximation theory in Rough Set to assign different values for absent words in different approximation regions. Theoretically, our algorithms can estimate smoothing value for absent words according to their relation with respect to existing words. Text classification experiments on public corpora have shown that our algorithms greatly improve the performance of LDA model, especially for unbalanced corpus.**

Keywords- Business Intelligence (BI); Text Classification; Latent Dirichlet Allocation (LDA); Smoothing; Tolerance Rough Set.

I. INTRODUCTION

Business Intelligence (BI) aims to support better business decision-making. Thus a BI system can be called a decision support system (DSS). BI technologies provide historical, current, and predictive views of business operations. With more and more serious business competition, traditional BI functions such as reporting, online analytical processing are hard to satisfy market needs. Intelligent components such as text mining and predictive analytics have attracted much attention. They orient not only on structured data but also on unstructured data. This paper proposes a new text classification algorithm based on smoothed Latent Dirichlet Allocation (LDA) model which could be used in text mining tasks, for example opinion classification, market feedback analysis and so on.

In recent years, statistical topic models have been successfully applied in many text mining tasks. These models can capture the word correlations in the corpus with a low-dimensional set of multinomial distribution, called ‘‘topics’’, and find a relatively short description for the documents. Latent Dirichlet Allocation (LDA) model which is proposed by D. Blei et al. is a widely used statistical topic model [1]. Its basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

However, the difference between the vocabularies of each class results in sparseness problem in LDA model. LDA model learned from documents of a certain class could only compute generative probability of a new document according to the vocabulary of this class. Unknown word in object documents will bring zero probability; as a result, the model is invalid. In fact, “absence” is not really “not exist”. If the corpus is large enough, these words should be present themselves. In order to simulate the reality and improve classification performance, some smoothing strategies are necessary to be conducted.

In [2], S. Chen & J. Goodman regarded that not only do smoothing methods generally prevent zero probabilities, but they also attempt to improve the accuracy of the model as a whole. Previous researches on smoothing LDA model mainly follow the statistic theory and assign a uniform distribution on all absent words. Although this kind of method is effective to prevent zero probability, it treats all terms as the same and couldn't emphasis the class-specific terms. This paper proposes a new smoothing strategy based on Tolerance Rough Set Theory (TRST) which makes use of upper approximation and lower approximation theory in Rough Set to assign different values for absent words in different approximation regions. Theoretically, it could effectively avoid uniform distribution in smoothing process and emphasizes the terms which have much more similarity with some class.

The rest of the paper is organized as follows. The second section presents LDA model and related research works in smoothing LDA model. The third section presents Tolerance Rough Set Theory (TRST). The fourth section presents smoothing LDA algorithm based on TRST. The fifth section presents experiments and discussion. The sixth section conclusions and presents the future research direction.

978-1-61284-109-0/11/$26.00 ©2011 IEEE

This paper is sponsored by the National Natural Science Foundation of China (No.60970061 and No. 61075056)

II. LATENT DIRICHLET ALLOCATION MODEL AND ITS SMOOTHING PROBLEM

Formally, [1] give the definition of the following terms.

A ***word*** is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1,\ldots,V\}$. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the v th word in the vocabulary is represented by a V -vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.

A ***document*** is a sequence of N words denoted by $w = \{w_1; w_2; ...; w_N\}$, where w_n is the n th word in the sequence.

A ***corpus*** is a collection of M documents denoted by $D = \{w_1; w_2; ...; w_M\}$.

Latent Dirichlet Allocation (LDA) model is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document w in a corpus D:

(1) Choose $N \sim Poisson(\xi)$.

(2) Choose $\theta \sim Dir(\alpha)$.

(3) For each of the N words w_n:

a) Choose a topic $z_n \sim Multinomial(\theta)$.

b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

In text classification task, the difference between the vocabularies of each class results in sparseness. In this case, LDA model (α^i, β^i) learned from documents of class *i* could only compute generative probability of a new document according to the vocabulary of class *i*. Unknown word in object documents will bring zero probability, as a result, the model is invalid. LDA model (α^i, β^i) learned from documents of class *i* has two parameters α^i and β^i. The initial value of α^i could be manually assigned. However, β^i which is related to vocabulary should be smoothed.

In [1], D. Blei designed a direct smoothing method which sets an exchangeable prior Dirichlet distribution $Dir(\eta)$ for β. Bayes inference is performed on this smoothed model and the estimation of parameter $\beta^{'}$ is as follows.

$$\beta_{ij}^{'} = \eta + \beta_{ij} \tag{1}$$

In fact, this process only adds a positive value on original estimate. Obviously, every word in β^i is not zero, as a result, zero probability is avoided in LDA model. Wherever, there are some disadvantages in this smoothing strategy.

It is likely to modify latent parameter β arbitrarily. β is related to latent variation (θ, z) and α. Directly modifying the β may neglect its effect on other parameters. The selection of exchangeable Dirichlet prior is in consideration of the conductible of LDA model. It confines the distribution of smoothing method. The assignment of parameter η is according to experience. There is no principle for its assignment.

Some other smoothing strategies for Latent Dirichlet Allocation (LDA) model are proposed based on statistic theory. Data-driven smoothing strategy is provided by W. Li et al. in which probability mass is allocated from smoothing-data to latent variables by the intrinsic inference procedure of LDA. Following this data-driven strategy, two concrete methods, Laplacian smoothing and Jelinek-Mercer smoothing, are employed in LDA model [3]. A feature-enhanced smoothing method is brought by D.X. Liu et al. in the idea that words not appeared in the training corpus can help to improve the classification performance [4].

III. TOLERANCE ROUGH SET

Rough Set Theory (RST) is put forward by Poland mathematician Z. Pawlak [5]. As a kind of soft computing method, it can effectively analyze and process incomplete, inconsistent, inaccurate data. It gets widespread international concern as it has been successfully applied in areas such as Knowledge Discovery (KD) in recent years [6]. The classical RST is based on equivalence relation that divides the universe of objects into disjoint classes. Tolerance Rough Set Theory (TRST) which is put forward by W. Ziarko extends the equivalence relation in classic RST to tolerance relation [7].

Let $S = (U, A, V, f)$ be an information system, where U is a nonempty finite set of objects called universe of discourse, A is a nonempty finite set of conditional attributes; and for every $a \in A$, such that $f : U \to V_a$, where V_a is called the value set of attribute a.

Definition 3.1. *If some of the precise attribute values in an information system are not known, i.e., missing or known partially, then such a system is called an incomplete information system. Otherwise the system is called a complete information system [8].*

Definition 3.2. *Let $S = (U, A, V, f)$ be an information system and the sign * denote null value, a tolerance relation T is defined as:*

$$\begin{aligned} T(B) = \{(x, y) \in U \times U \mid \forall b \in B, \\ b(x) = b(y) \vee b(x) = * \vee b(y) = *\}. \end{aligned} \tag{2}$$

Where, $B \subseteq A$. Obviously, T is reflexive and symmetric, but not transferable. Let $I_B(x) = \{y \in U \mid (x,y) \in T(B)\}$, and then $I_B(x)$ is called the tolerance class of the object x with respect to the set $B \subseteq A$.

Definition 3.3. *Let $S = (U, A, V, f)$ be an information system, $X \subseteq U$, $B \subseteq A$, the upper approximation and lower approximation of X with regard to attribute set B under the tolerance relation T can be defined as:*

$$U_B(X) = \{x \in U \mid I_B(x) \cap X \neq \varnothing\} \tag{3}$$

$$L_B(X) = \{x \in U \mid I_B(x) \subseteq X\} \tag{4}$$

IV. A SMOOTHED LATENT DIRICHLET ALLOCATION MODEL

A smoothed LDA model based on TRS is defined as follows.

An incomplete information system for a document set is represented as $WS = (U, TS \cup \{class\}, f)$, where U is the set of documents, each document is an object $d \in U$; TS is the set of total terms which occur in the document set, $class$ is the decision attribute, i.e., the class label of the documents. The weights of those terms which do not occur in a document are considered missing information and denoted by sign * instead of zero.

In document space, the tolerance relation and tolerance class of document are defined as [9]:

Definition 4.1. *For a subset of TS, $B \subseteq TS$, a tolerance relation $T(B)$ on U is defined as:*

$$\begin{aligned} T(B) = \{(d_x, d_y) \in U \times U \mid \forall b \in B, \\ \mid b(d_x) - d(d_y) \mid \leq \delta \vee b(d_x) = * \vee b(d_y) = *\}. \end{aligned} \tag{5}$$

Because weights are real values, the requirement $b(d_x) = b(d_y)$ is too strict. Here it is replaced with $\mid b(d_x) - b(d_y) \mid \leq \delta$, where, $\delta \in [0,1]$. Consequently, tolerance class of a document d_x with respect to $B \subseteq TS$, $I_B(d_x)$ is the set of documents which are indiscernible to d_x, i.e., $I_B(d_x) = \{d_y \in U \mid (d_x, d_y) \in T(B)\}$.

On the other hand, correlation between terms is valuable for complementing missing information. Thus, the tolerance class of term is also defined in term space. Let $U = \{d_1, ..., d_M\}$ be a set of documents and $TS = \{t_1, ..., t_N\}$ set of terms for U. The tolerance space of term is defined over a universe of all terms for U.

Definition 4.2. *Let $f_D(t_i, t_j)$ denotes the number of documents in U in which both term t_i and t_j occurs. The uncertainty function I with regards to co-occurrence threshold θ is defined as:*

$$I_\theta(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \bigcup \{t_i\} \tag{6}$$

Clearly, the above function satisfies conditions of being reflexive: $t_i \in I_\theta(t_i)$ and symmetric: $t_j \in I_\theta(t_i) \Leftrightarrow t_i \in I_\theta(t_j)$ for any $t_i, t_j \in T$. Thus, $I_\theta(t_i)$ is the tolerance class of term t_i. Tolerance class of terms is generated to capture conceptually related terms into classes. The degree of correlation of terms in tolerance classes can be controlled by varying the threshold θ.

In tolerance space of term, an expanded representation of document can be acquired by representing document as set of tolerance classes of terms it contains. This can be achieved by simply representing document with its upper approximation, e.g., the document $d_i \in U$ is represented by:

$$U_R(d_i) = \{t_i \in T \mid I_\theta(t_i) \cap d_i \neq \varnothing\} \tag{7}$$

This approach to document representation takes into consideration not only terms actually occurring document but also other related terms with similar meanings.

Definition 4.3. *Assume that v_i is vocabulary of class c_i, the out-of-vocabulary (oov) words of class c_i is represented as follows.*

$$oov_i = \bigcup_{i=1}^{c} v_i - v_i \tag{8}$$

Firstly, the oov_i word list is added to class i and LDA model could be trained with an extended vocabulary. The inference mechanism in LDA could be maturely applied. That is, parameter (θ, z) and α could be served in smoothing process with the changing of w.

Secondly, we make use of the idea of Laplacian transform to add virtual vocabulary to each class and assign its value as follows.

$$P(w_t \mid c_i) = \frac{n_i(w_t) + \lambda}{n_i(c_i) + \lambda \mid V \mid} \tag{9}$$

$$\lambda = \begin{cases} tf_{ij} + 1 & if\ t_j \in d_i, \\ \varphi + \rho min_{t_k \in d_i} TF_{ik} & if\ t_j \notin U_R(d_i), \\ \varphi & if\ t_j \in U_R(d_i) \wedge t_j \notin d_i \end{cases} \tag{10}$$

Where, tf_{ij} is the frequency of term t_j in document d_i, $\varphi, \rho \in [0,1]$. In this chapter, we assign the paprameter $\varphi = \rho = 0.2$ based on experience. That is, for an unknown word, we assign its prior value according to three cases: appearing in d_i, being absent in d_i, but appearing in $U_R(d_i)$ and being absent in $U_R(d_i)$.

V. EXPERIMENTS AND DISCUSSIONS

Experiments are performed on both English corpus 20Newsgroup and Chinese corpus TanCorp-12. 20Newsgroup was collected by Ken Lang for text classification research [10]. All documents are distributed across 20 categories evenly (balanced), altogether 20,000 texts. TanCorp-12, which is collected by Songbo Tan, includes 14,150 texts, distributing in 12 categories (unbalanced) [11].

For both corpora, we extracted 70% texts in each class as training set and the rest 30% as testing set. We repeated our experiments five times and got the average results for each experiment setting.

In our experiments, the Micro-F1 and the Macro-F1 measures which are introduced in [12] are used to evaluate the synthesis classification performance. We perform the comparison between our algorithm TRS-LDA and traditional LDA model. For LDA model, the number of latent topics is an important factor which defines the granularity of the model. So, we evaluate models on different topic numbers: 5, 10, 20, 50, 100.

A. Experiments on Analysis Synthesis Performance on 20Newsgroup

Firstly, we use English corpus 20Newsgroup to testify our algorithm. There are 700 texts used for training classifier in each class and the other (300 texts) for testing process. The results are shown in FIG. 1. Some phenomena could be observed as follows.

1) On the whole, TRS-LDA model has an evident higher performance of about 6%~7% absolutely than LDA model both on Micro-F1 and Macro-F1. This occurs across all values of topic number.

2) When different latent topic number is selected, the behaviors of the two models are different. With the increase of topic number, the performance of LDA model changes little. However, the TRS-LDA model improves little by little.

3) Micro-F1 and Macro-F1 have the very consistent results.

B. Experiments on Analysis Synthesis Performance on TanCorp-12

1) We also use Chinese corpus TanCorp-12 to testify our algorithm. There are 70% texts used for training classifier in each class and the other (30% texts) for testing process. The results are shown in FIG. 2. Some phenomena could be observed as follows.

2) TRS-LDA model has an evident higher performance of about 7% absolutely than LDA model on Micro-F1 across all values of topic number.

3) TRS-LDA model has an evident higher performance of nearly 20% absolutely than LDA model on Macro-F1 across all values of topic number.

4) When different latent topic number is selected, the behaviors of the two models are different. With the increase of topic number, the Micro-F1 value of LDA model changes little. However, that of TRS-LDA model improves little by little.

5) Micro-F1 and Macro-F1 have inconsistent results. The possible reason is that TanCorp-12 is an unbalanced corpus.

From the experiments, we could find that TRS-LDA model could improve the classification performance obviously on unbalanced corpus. For an unbalanced text classification task, the classification performance on small classes are usually not satisfied because there are too little training documents in small classes. TRS-LDA algorithm could improve the classification performance on small classes to some extent.

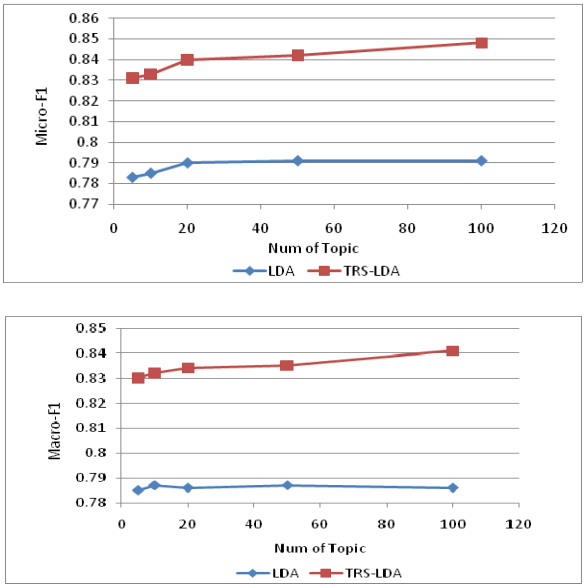

Figure 1. Classification results on 20Newsgroup

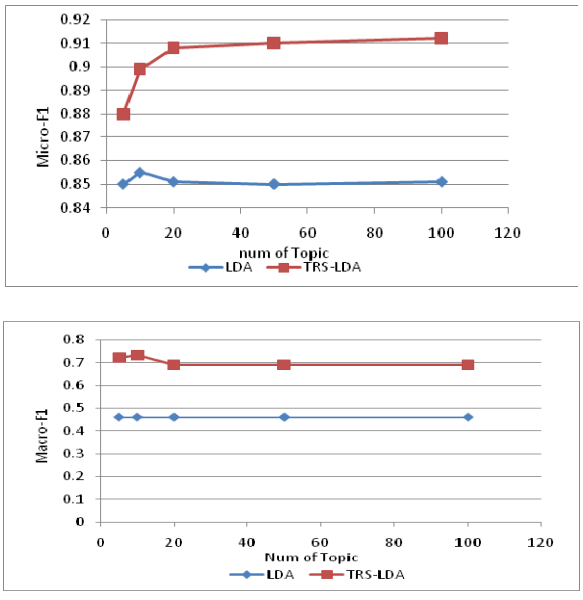

Figure 2. Classification results on TanCorp-12

C. *Analysis on performance of each class on unbalanced corpus*

Let us observe the effectiveness of TRS-LDA from the more refined granularity. From the distribution of TanCorp-12, we could find that the largest class includes nearly 3,000 documents while the smallest one includes only 150 documents.

We still adopt F1 metrics to observe the classification performance on each class. FIG..3 shows the distribution of TanCorp-12 and the F1 performance on each class on ths corpus. The class sequence on the top of FIG.3 is consistent with that on the bottom. That is, the classes are ordered ascending according to the number of documents in class.

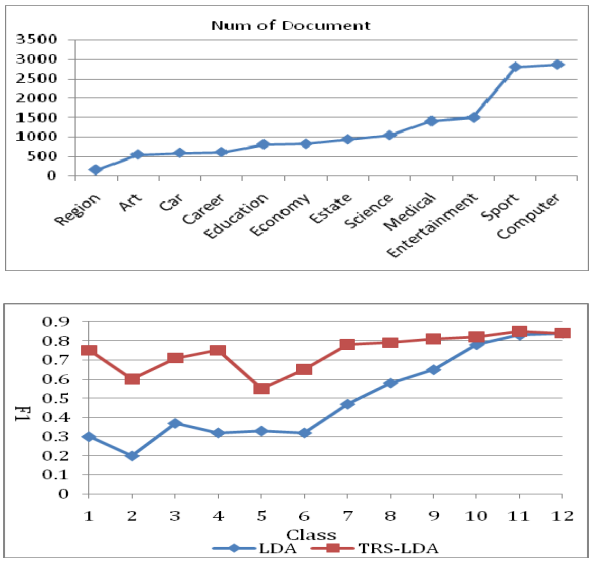

Figure 3. Relation between class size and classification performance

From FIG.3, we could find that F1 values of TRS-LDA model are always higher than those of LDA model. The increase of F1 values on big classes are less than those on small classes. It indicates that the contribution of TRS-LDA algorithm lies in improving classification effectiveness on small classes, in the same time, not degrading the performance on big classes. The effect of this smoothing algorithm lies in relieving the over-fitting on small class.

From all these experiments, we could find that the smoothing algorithm TRS-LDA has different effect on big classes and small classes. Normally, there is much more vocabulary in big class, as a result, the virtual vocabulary added has little change on its original one, vice verse, for a small class. Consequently, a perfect smoothing algorithm could effectively improve the classification performance on small classes.

We could also find that this strategy avoids the arbitrariness in modification parameters in LDA model. In the same time, we could adjust the scale of tolerance class to meet different situation.

VI. CONCLUSION AND PERSPECTIVES

As a kind of intelligent component of BI system, text classification plays important role in BI application such as client opinion classification, market feedback analysis and so on. Latent Dirichlet Allocation model is a widely used text representation model for various text processing applications. However, data sparseness problem limits its performance. The main contribution of this paper is proposing a smoothed LDA model based on Tolerance Rough Set Theory (TRST), which constructs term tolerance class according to term co-occurrence and compensates the probability of unknown words based on tolerance class information. Effectiveness of the algorithm has been verified theoretically and experimentally. Experiments on public corpora show that the algorithm could improve classification performance greatly, especially when the corpus is unbalanced.

Algorithms proposed in this paper only make use of the co-occurrence information for the purpose of smoothing. Further research is expected that linguistic semantic information could be incorporated into constructing tolerance class to enhance text representation models.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.

[2] S. F. Chen and J. T. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, Computer Speech and Language, vol.13, no.4, pp.359-393, 1999.

[3] W. B. Li, L. Sun, Y.Y. Feng and D. K. Zhang, Smoothing LDA Model for Text Categorization, The 4th Asia Information Retrieval Symposium, Harbin, China, pp.83-94, 2008.

[4] D. X. Liu, W. R. Xu, and J. N. Hu, A feature-enhanced smoothing method for LDA model applied to text classification, Intl. Conf. on Natural Language Processing and Knowledge Engineering, Dalian, China, pp.1- 7, 2009.

[5] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic publisher, Dordrecht, 1991.

[6] D. Miao, G. Wang and Q. Liu, Granular Computing: Past, Present and future, Science Publisher, Beijing, 2007.

[7] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences, vol.46, no.1, pp.39-59, 1993.

[8] Y. Y. Yao and J. T. Yao, Granular computing as a basis for consistent classification problems, Proc. of PAKDD'02 Workshop on Toward the Foundation of Data Mining, Taipei, Taiwan, pp.101-106, 2002.

[9] Q. Duan, D. Miao and M. Chen, Web Document Classification Based on Rough Set, The 11th Intl. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Toronto, Canada, pp. 240-247, 2007.

[10] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html.

[11] S. Tan, X. Cheng, Moustafa M. Ghanem, B. Wang and H. Xu, A novel refinement approach for text categorization. Proc. of the 14th ACM intl. conf. on Information and knowledge management. Bremen, Germany, pp.469 -476, 2005.

[12] C. J. Van Rijsbergen. Information Retrieval, Butterworths, London, 1979.