ISSN 1063-7850, Technical Physics Letters, 2016, Vol. 42, No. 8, pp. 837–839. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © S.N. Kol'tsov, S.I. Nikolenko, E.Yu. Kol'tsova, 2016, published in Pis'ma v Zhurnal Tekhnicheskoi Fiziki, 2016, Vol. 42, No. 16, pp. 21–25.

Gibbs Sampler Optimization for Analysis of a Granulated Medium

S. N. Kol'tsov*, S. I. Nikolenko, and E. Yu. Kol'tsova

National Research University Higher School of Economics, St. Petersburg, 190008 Russia

**e-mail: skoltsov@hse.ru*

Received February 26, 2016

Abstract—A new variant of the method of probability density distribution recovery for solving topical modeling problems is described. Disadvantages of the Gibbs sampling algorithm are considered, and a modified variant, called the "granulated sampling method," is proposed. Based on the results of statistical modeling, it is shown that the proposed algorithm is characterized by higher stability as compared to other variants of Gibbs sampling.

DOI: 10.1134/S1063785016080241

The rapid advancement of computer technology is accompanied by the development of methods of statistical modeling (e.g., Markov chain Monte Carlo algorithms). Bayesian approach to the estimation of hidden parameters of multidimensional distributions is among the main methods of analysis in many fields including high-energy physics [1], mass spectrometry and bioinformatics [2], astrophysics [3], and statistical physics [4]. One of the most widely used algorithms for determining hidden parameters from observation data is based on the Gibbs sampling. A remarkable feature of this algorithm is that it does not require an explicitly expressed general distribution and employs only one-dimensional conditional probabilities involved in the distribution.

In recent years, methods developed in physics are frequently used for analysis of big data (data mining) characterized by multidimensional distributions. Although big data can represent a set of various objects, such as mass spectra [2], sound tracks and photographs [5], or news in social networks [6], the task of recovering initial probability density distribution is generally the same despite this diversity.

The problem of recovery of the initial multidimensional distribution in the form of a mixture of distributions with hidden parameters is called "topic modeling (TM) [7, 8]. TM is based on the following assumptions: (i) the a posteriori distribution based on the Bayes rule is found by estimating expected values via sampling and (ii) distributions of topics in documents and words (terms) in topics are multinominal, representing a priori Dirichlet distributions with parameters α and β. The approach to estimating probability density distribution in TM with allowance for the Dirichlet functions is called "latent Dirichlet allocation" (LDA) [7]. The present work describes a modified Gibbs sampler for text data [8], which has been previously used for soft clusterization of mass spectra [2], detection and identification of nuclear isotopes [9], and many other applications.

In the framework of TM, observed variables in text data are represented by documents d and words w from a given collection. It is also assumed that there exist a finite set of topics T and the collection of documents is generated by discrete distribution $p(d, w, t)$, where d, w, and t are the document, word, and topic variables, respectively. Hidden variable t characterizes the countable set of topics and implies a one-dimensional Dirichlet distribution of words. Accordingly, each document represents a mixture of latent topic distributions and every topic is determined by its probabilistic distribution on the set of words. To construct a topic model of data means to find hidden word distributions in a document with respect to topics based on observed variables, i.e., to establish a set of one-dimensional conditional distributions $p(w|t) \equiv \varphi(w, t)$ (matrix Φ, distribution of words with respect to topics) and set of one-dimensional distributions $p(t|d) \approx \theta(t, d)$ (matrix Θ, distribution of documents with respect to topics) for each document d. Final distributions of words and documents based on the Gibbs sampling are calculated as follows [8]:

$$P(z_i = j|w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum_m C^{WT}_{m,j} + V\beta} \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j} + \alpha T}, \quad (1)$$

$$\theta_{dj} = \frac{C_{d,j}^{DT} + \alpha}{C_{d,j}^{DT} + T\alpha}, \quad (2)$$

$$\phi_{m,j} = \frac{C_{m,j}^{WT} + \beta}{\sum_{\dot{m}} C_{\dot{m},j}^{WT} + V\beta}, \quad (3)$$

where α and β are the parameters determining one-dimensional Dirichlet distributions and C are the counters obtained in the course of sampling: $C_{m,j}^{WT}$ is the number of times word w is encountered in topic t, $C_{d,j}^{DT}$ is the number of times word w in document d is related to topic t, $\Sigma_m C_{m,j}^{WT} = n_t$ is the number of words related to topic t, and $C_{d,j}^{DT} = n_d$ is the length of document in words. During the sampling, matrix $p(z = j|w_i = m, z_{-i}, w_{-i})$ is calculated that is used to determine counters by formula (1). Then, these counters are used to calculate final distributions $\theta_{d,j}$ and $\phi_{m,j}$ by formulas (2) and (3), respectively.

Solving the TM problem is equivalent to stochastic matric decomposition by which large matrix F containing documents d and words w is approximated by the product of two matrices, $\theta_{d,j}$ and $\phi_{m,j}$, of lower dimensions. However, the stochastic matric decomposition is not unique and can only be determined to within a nonsingular transformation [10]. In terms of the Gibbs sampling algorithm, the nonunique recovery of a multidimensional density of the mixture of distributions implies that the algorithm starting from various initial distributions would converge to different points in the set of solutions. This is manifested by the fact that, for different starts of the algorithm from same initial data, the content of matrices $\theta_{d,j}$ and $\phi_{m,j}$ will be different; in other words, the Gibbs sampler is not stable. Problems the solutions of which are not unique and/or stable are called "ill-posed." A general approach to solving such problems is provided by the regularization according to Tikhonov [11], which consists in adding a priori information that allows the set of solutions to be decreased.

In the present work, it is proposed to perform sampling by the granulated LDA technique, which differs from the conventional Gauss sampling [8] by using a modified topic concept. According to this, first, each document is treated as a granulated surface consisting of granules. Each granule represents a sequence of words of preset length. Second, all words belonging to same granule refer to the same topic. Granules are characterized by their size, which represents the width of the sampling window (regularization parameter). The granulated TM variant is aimed at the recovery of matrices $\theta_{d,j}$ and $\phi_{m,j}$ by averaging the topic content of granules over a large number of documents, i.e., by calculating the expected values (mathematical expectation).

The granulated variant of Gauss sampling was implemented as follows. After the initiation of matrices $\theta_{d,j}$ and $\phi_{m,j}$, two embedded cycles are organized so that the external cycle runs over the list of documents, while the internal cycle performs random sampling over granules. During this, all words in a randomly selected window are allocated to the same topic with randomly generated number. The number of random "samples" of words in a document is equal to the number of words in this document. During long-term sampling, words frequently encountered inside a granule will more frequently have the same topic number and their probability of belonging to same topic will be, on average, higher. At the last stage, final calculation of matrices $\theta_{d,j}$ and $\phi_{m,j}$ of the distributions of words and documents is performed using the corresponding counters.

The stability of TM in this work was evaluated by mutual comparison of a series of one-dimensional distributions (topics) obtained in various runs based on the Kullback–Leibler normalized measure (Kn) [12]. Investigation showed that two topics are identical provided that $Kn > 90\%$ [12]. In the present work, a topic was considered stably reproduced from one run to another if it was reproduced in three runs on a level of $Kn > 90\%$. In order to evaluate the stability of TM, we have analyzed three topic models based on the Gibbs sampling procedures: (i) LDA (standard variant); (ii) SLDA (model with learning) [6]; and (iii) GLDA (granulated LDA with granule size +1).

Each model was started three times from the same initial data. The models were tested on a set of 101 481 documents from the Livejournal social network using 200 topics with the same Dirichlet function parameters $\alpha = 0.1$ and $\beta = 0.5$. The TM results showed that the standard LDA variant provided 74 stable topics of 200, the SLDA variant yielded 87 stable topics, and the GLDA variant ensured 195 stable topics.

Thus, the results of statistical modeling showed that the proposed granulated LDA (GLDA) model significantly exceeds other analogous models in respect of stability and can be effectively used in solving problems related to the recovery of multidimensional distributions in physics and other fields (e.g., data mining) employing statistical physico-mathematical models.

REFERENCES

1. A. Caldwell, D. Kollar, and K. Kröninger, Comput. Phys. Commun. **180**, 2197 (2009); arXiv:0808.2552.
2. I. Chernyavsky, T. Alexandrov, P. Maass, and S. Nikolenko, *Proceedings of the German Conference on Bioinformatics (September, 2012)*, pp. 39–48.
3. *Handbook of Markov Chain Monte Carlo*, Ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Chapman & Hall/CRC Press, 2011), pp. 383–399.

4. B. A. Berg and A. Billoire, *Markov Chain Monte Carlo Simulations* (John Wiley & Sons, 2008).
5. S. Geman and D. Geman, IEEE Trans. Pattern Anal. Machine Intell. **6**, 721 (1984).
6. S. Bodrunova, S. Koltsov, O. Koltsova, S. Nikolenko, and A. Shimorina, *Proceedings of the 12th Mexican Int. Conf. on Artificial Intelligence (MICAI 2013)* (Springer Verlag, Berlin, 2013), Part I, pp. 265–274.
7. D. Blei, A. Ng, and M. Jordan (Ed. by J. Lafferty), J. Machine Learn. Res. **3**, 993 (2003).
8. T. Griffiths and M. Steyvers, Proc. Natl. Acad. Sci. USA **101** (Suppl. 1), 5228 (2004).
9. C. Nelson et al., *Proceedings of the IEEE Conference on Frequency Control* (2012).
10. K. V. Vorontsov, Dokl. Math. **89** (3), 301 (2014).
11. A. N. Tikhonov and V. Ya. Arsenin, *Methods of Solution of Ill-Posed Problems* (Nauka, Moscow, 1986) [in Russian].
12. S. Koltsov, O. Koltsova, and S. Nikolenko, *Proceedings of the ACM Web Science Conference (WebSci'14, June 23–26, 2014, Bloomington, IN, USA)*, pp. 161–165.

Translated by P. Pozdeev