ISSN 1063-7850, Technical Physics Letters, 2017, Vol. 43, No. 6, pp. 586–588. © Pleiades Publishing, Ltd., 2017.
Original Russian Text © S.N. Koltcov, 2017, published in Pis'ma v Zhurnal Tekhnicheskoi Fiziki, 2017, Vol. 43, No. 12, pp. 90–95.

A Thermodynamic Approach to Selecting a Number of Clusters Based on Topic Modeling

S. N. Koltcov

National Research University Higher School of Economics, St. Petersburg, 190008 Russia
e-mail: skoltsov@hse.ru

Received January 31, 2017

Abstract—A thermodynamic approach has been applied to solving the problem of selecting the number of clusters/topics in topic modeling. The main principles of this approach are formulated and the behavior of topic models during temperature variations is studied. Using thermodynamic formalism, the existence of the entropy phase transition in topic models is shown and criteria for the choice of optimum number of clusters/topics are determined.

DOI: 10.1134/S1063785017060207

A thermodynamic formalism based on minimization of the free energy has been successfully applied in various fields, such as processing of images [1], neural networks [2], and cluster analysis [3]. Substantial development of clustering methods has been achieved in the framework of topic modeling (TM) [4, 5]. TM solves the task of recovering the initial multidimensional distribution in the form of a mixture of multinomial distributions with hidden parameters. One unsolved problem in TM is related to selecting the number of distributions in the mixture. This problem is also encountered in cluster analysis, network analysis [6], and investigation of phase transitions in substances with various spatial structures [7].

Since TM is intended to work with big data, the manifold of documents and words (terms) can be treated as a mesoscopic system with a large number of particles (millions of documents and words in these) characterized by thermodynamic values such as energy, entropy, and free energy. In view of this, a thermodynamic approach to the problem of selecting the number of clusters/topics in the TM can be formulated as follows.

(i) The total number of words and documents in the information thermodynamic system under consideration is constant (i.e., the system volume is not changed).

(ii) The topic is a state (analog of the spin direction) that each word and document in a given collection can take.

(iii) The information thermodynamic system is open and can only exchange energy with the external medium by means of temperature variation. In the proposed approach, the system temperature is defined as number T of topics (or clusters) determined from outside. A change in the number of topics leads to changes in the number of states $N(T)$, energy $E(T)$, and entropy $S(T)$;

(iv) The system entropy can be expressed in the form of a logarithm of the number of states with energy E: $S = \ln(N(E))$ [2].

(v) The energy of each element in the system can be expressed as the probability of finding this element in the system [2] or, in the context of this work, as the probability for a word to belong to a given topic: $E_{nt} = -\ln(P_{nt})$, where n is the number of this word in the vocabulary and t is the topic number.

One commonly accepted approach to investigation of the thermodynamic properties of a system is based on calculations of the statistical sum (partition function). The knowledge of this sum allows various thermodynamic quantities to be calculated as functions of the temperature. For a microcanonical ensemble, it is assumed that the system in a nonequilibrium state would occur on some of highly probable states [8]. The partition function of such a system can be expressed as follows [2]:

$$Z = \int de^{E-TS(E)} = \int de^{F}$$

where F is the free energy of the nonequilibrium system. Using numerical experiments on the TM, it is possible to directly calculate the number of states with preset energy during variation of parameter T and, thus, to construct the dependence of the free energy of the information system on the number of topics. The minimum of free energy and the point of phase transition will correspond to the optimum number of topics [3].

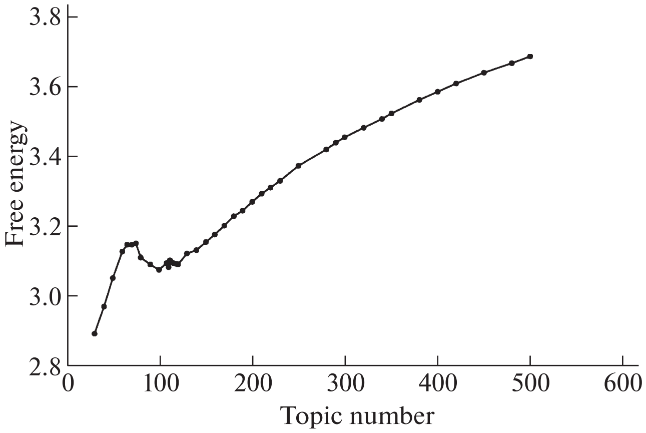

Fig. 1. Plot of normalized free energy vs. number of topics.

In the present study, we are interested in knowing the behavior of the free energy normalized to the number of topics:

$$F/T = E(T)/T - S(T).$$

The TM result is matrix Φ that contains distributions of the probabilities for all unique words to belong to topics. The matrix dimension is NT, where N is the number of unique words (rows) and T is the number of topics (columns) in the matrix. The distribution of probabilities in the TM is such that their sum over all words and topics is T (i.e., the total energy of the given information thermodynamic system is equal to the number of topics). It should be noted that the initial distribution for matrix Φ in the TM represents a uniform distribution in which all words have the same probability $1/N$ (corresponding to the maximum entropy). The TM process consists in the transition to a strongly nonequilibrium state in which one part of states has a high probability $P_{nt} > 1/N$, while the other part has a low probability $P_{nt} < 1/N$ (close to zero). Therefore, the number of states with probabilities above (or below) $1/N$ is a function of the number of topics. Since the main contribution to the total free energy is due to the highly probable states [8], calculations of the number of these states allow the behavior of the nonequilibrium free energy to be studied as dependent on the number of topics.

In the framework of the present investigation, the following series of computer experiments have been carried out. A TM based on the latent Dirichlet allocation (LDA) with Gibbs sampling at various numbers of topics was performed using a fixed set of 101 481 documents from the Livejournal social network with N = 172 939 unique words. The number of topics was varied within T = [30; 500]. In a region featuring strong fluctuations in the nonequilibrium free energy, the model calculations were performed at a step of one topic. As is known, the Gibbs sampling procedure has some instability [9]. For this reason, free energy $F(t)$ of a nonequilibrium state was averaged over three calculations for each T. In addition, three TM runs were used to determine the level of TM fluctuations. The energy and entropy of the nonequilibrium state for each T were calculated as follows:

$$E(T) = E(T) - E(T)_0 = \ln\left(\frac{\sum_{t=1}^{T}\sum_{n=1}^{N} P_{nt}}{T}\right),$$

$$S(T) = S(T) - S_0 = \ln\left(\frac{N_t}{NT}\right),$$

where N_t is the number of states with $P_{nt} > 1$, (NT) is the total number of states, T is the number of topics (variable), N is the size of a vocabulary of unique words, and E_0 and S_0 are the energy and entropy, respectively, of the system with initial uniform probability distribution. Figure 1 shows a plot of normalized free energy $F(T)/T$ constructed based on the obtained dependences of $E(T)$ and $S(T)$.

Analysis of the obtained results showed that it was possible to separate two regions with nearly linear dependence of the free energy on the number of topics. The interval of T = 100–118 corresponded to a transition region where a change in the number of topics does not significantly influence the free energy and a minimum of the nonequilibrium free energy is observed. The two regions can be more precisely separated based on analysis of the second derivative of free energy with respect to temperature, which determines the heat capacity of the system. In the present work, we introduce the notion of information topic capacity C_{inf} at a fixed number of words, which is defined as

$$C_{\text{inf}} = \left(\frac{d^2F}{dT^2}\right),$$

where F is the free energy of the information system and T is the number of topics. This topic capacity characterizes variation in the system energy upon a change in the number of topics by unity. Figure 2 shows a plot of the topic capacity versus number of topics. Strong jumps of the topic capacity in the interval of T = 100–118 corresponds to the so-called "entropy phase transition" [10].

Therefore, numerical experiments on the TM revealed two phase states of the given information thermodynamic system. The first phase corresponds to a region in which small changes in the number of topics lead to rather significant variations of the free energy. The average fluctuation of free energy in this phase amounts to $\Delta F(T) \cong 0.8\%$ of the mean free energy. Inside the phase transition interval, the average fluctuation is about $\Delta F(T) \cong 0.7\%$. The second phase of the given information thermodynamic system

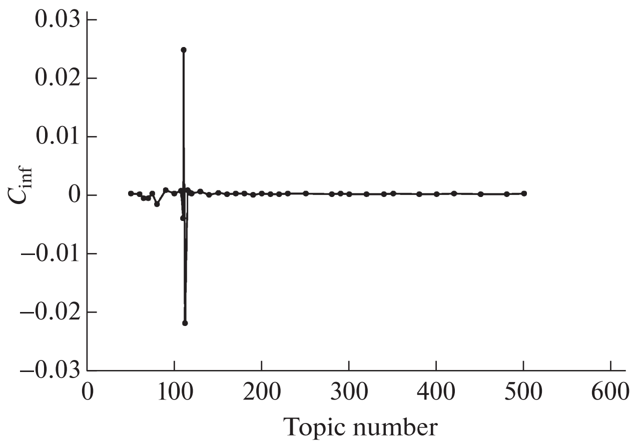

Fig. 2. Plot of topic capacity vs. number of topics.

corresponds to a region (beginning with $T = 120$) in which the average fluctuation of free energy is almost three times as small ($\Delta F(T) \cong 0.2\%$) as compared to the first phase. From this, it can be preliminarily concluded that the TM stability may be one of the criteria for selecting the optimum number of topics. Thus, a set of criteria for selecting the optimum number of topics in the TM may be as follows. First, the optimum region should correspond to a phase state with minimum level of fluctuations. Second, the selected region should contain a minimum of free energy. Therefore, a candidate for the optimum region is the second phase state of the given information thermodynamic system. However, since an increase in the number of topics is accompanied by the system passage toward the uniform distribution (corresponding to maximum entropy), the optimum number of topics for the given collection probably corresponds to the beginning of the second-phase region ($\cong$120 topics).

SPELL:OK

In concluding, the use of a thermodynamic formalism provides a deeper insight into the behavior of big-data arrays with variation of parameters such as the number of distributions in the mixture. The choice of the optimum size of a mixture of distributions for the TM can be justified based on an analysis of the behavior of a given information thermodynamic system in various phase states. Further development of this approach can be provided by the application of a multifractal formalism.

Acknowledgments. This study was performed in the framework of the Program of Fundamental Investigation of National Research University Higher School of Economics (St. Petersburg) for 2016.

REFERENCES

1. K. Friston, M. Levin, B. Sengupta, and G. Pezzulo, J. R. Soc. Interface **12**, 20141383 (2015). doi doi 10.1098/rsif.2014.1383
2. G. Tkacik, T. Mora, O. Marre, et al., arXiv:1407.5946 [q-bio.NC] (2014).
3. K. Rose, E. Gurewitz, and G. Fox, Phys. Rev. Lett. **65**, 945 (1990).
4. T. Griffiths and M. Steyvers, Proc. Nat. Acad. Sci. **101**, 5228 (2004).
5. D. M. Blei, A. Y. Ng, and M. I. Jordan, J. Machine Learn. Res. **3**, 993 (2003).
6. S. Fortunato, Phys. Rep. **486**, 75 (2010).
7. V. A. Ivanskoi, Tech. Phys. **53**, 455 (2008).
8. S. G. Abaimov, *Statistical Physics of Complex Systems* (URSS, Moscow, 2011) [in Russian].
9. S. N. Koltcov, S. I. Nikolenko, and E. Yu. Koltsova, Tech. Phys. Lett. **42**, 837 (2016).
10. A. G. Bashkirov, Theor. Math. Phys. **149**, 1559 (2006).

Translated by P. Pozdeev