Physica A 512 (2018) 1192–1204

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Application of Rényi and Tsallis entropies to topic modeling optimization

Sergei Koltcov

Laboratory for Internet Studies (LINIS), National Research University Higher School of Economics, ul. Soyuza Pechatnikov, d. 16, 190008 St. Petersburg, Russia

HIGHLIGHTS

- Rényi and Tsallis entropies are found to detect the optimal number of topics.
- Entropy approach reveals a meaningful difference in algorithm performance.
- The content of all topics is quasi-periodically related to the number of topics.

ARTICLE INFO

Article history:
Received 16 March 2018
Available online xxxx

Keywords:
Topic modeling
Renyi
Entropy
Free energy
Complex system

ABSTRACT

This study proposes to minimize Rényi and Tsallis entropies for finding the optimal number of topics T in topic modeling (TM). A promising tool to obtain knowledge about large text collections, TM is a method whose properties are underresearched; in particular, parameter optimization in such models has been hindered by the use of monotonous quality functions with no clear thresholds. In this research, topic models obtained from large text collections are viewed as nonequilibrium complex systems where the number of topics is regarded as an equivalent of temperature. This allows calculating free energy of such systems—a value through which both Rényi and Tsallis entropies are easily expressed. Numerical experiments with four TM algorithms and two text collections show that both entropies as functions of the number of topics yield clear minima in the middle area of the range of T. On the marked-up dataset the minima of three algorithms correspond to the value of T detected by humans. It is concluded that Tsallis and especially Rényi entropy can be used for T optimization instead of Shannon entropy that decreases even when T becomes obviously excessive. Additionally, some algorithms are found to be better suited for revealing local entropy minima. Finally, we test whether the overall content of all topics taken together is resistant to the change of T and find out that this dependence has a quasi-periodic structure which demands further research.

1. Introduction

Statistical physics is increasingly being used to describe objects and processes that go beyond physical phenomena. Thus, large arrays of textual data, which have been rapidly accumulating on the Internet in the last decade, require ever more complex methods for their automatic processing and modeling. A wide range of mathematical tools, including topic modeling, is used for this [1], but their properties and behavior remain underresearched. This makes parameter optimization for such models a difficult task. However, if the results of topic modeling are considered equivalent to nonequilibrium complex systems (since the former, as it will be shown below, possess some properties of such systems), this would make it possible to apply a whole range of approaches from statistical physics. First of all, these are models for analyzing the

E-mail address: skoltsov@hse.ru.

https://doi.org/10.1016/j.physa.2018.08.050
0378-4371/

processes of self-organization of large ensembles. The basis for such an analysis may be an approach in which behavior of a topic model of a textual collection as an ensemble would be determined by thermodynamic functions, such as entropy or free energy. It is known that complex systems can be characterized by exponential and power law distributions, which is especially true for social [2,3], biological [4,5] and economic systems [6,7]. However, for topic models of textual collections, where the units are documents, words and latent semantic variables (topics), Pareto-like distributions are more typical [8,9]. Proceeding from this, when applying the maximum entropy principle for such systems, we propose to use an approach based on deformed statistic with the underlying Rényi or Tsallis entropies [10,11]. In this case, the deformed statistic of complex systems, like its non-deformed equivalent in other cases, will describe the probabilistic features that characterize the topic model of a textual collection as a system that has a large number of "particles" and that can remain in thermodynamically equilibrium and nonequilibrium states. If the deformation parameter q is accounted for while modeling thermodynamically atypical systems with long-range interactions, we expect that behavior of such systems could be explained much better than with any standard statistic. Moreover, the search for optimal parameters describing the state of these systems can be achieved on the basis of an entropy maximization procedure [12].

Our attention in this work is focused on topic modeling [1], since it is the most effective and sometimes the only available method of obtaining knowledge about the topic structure of large textual collections of which nothing is known in advance. This task is often encountered in the studies of Internet content, including news, consumer reviews, and social network messages. At the same time, topic modeling (TM) as a mathematical approach is applicable not only to textual data [13], but also to mass spectra [14], images [15], and other objects. In essence, topic modeling is an expanded version of cluster analysis that allows simultaneous estimation of distributions of both words and documents over topics/clusters. Moreover, topic models also provide the opportunity to rank words and documents according to the probability within each topic/cluster, which is not typical for traditional cluster analysis. The major problem of this group of methods is the lack of ground truth, that is, of knowledge about the correct number and composition of clusters. This hinders investigation of the properties of these models and makes us seek solutions based on theories from other areas of science.

Thus, in this paper we use the concept of deformed entropy and a range of thermodynamic concepts to investigate behavior of topic models under conditions of the changing number of topics. The purpose of such a study is to find the optimal number of topics/clusters, first, based on the maximum information approach, and second, on the basis of the T-invariance principle introduced further in the work.

The rest of the paper proceeds as follows. Section 2 first briefly explains the logic of topic modeling, which is necessary to further describe the proposed solutions. Next, this section provides an overview of the available approaches for determining the optimal number of clusters in cluster analysis and topics in topic modeling, and their limitations are indicated. In Section 3, we propose our entropy approach to the analysis of topic models as complex nonequilibrium systems, an approach that allows finding the optimal number of topics. Sections 4 and 5 describe the data used and the results of numerical experiments performed to verify our approach. Section 4 shows that the minimum q-deformed entropy is reached with the 'correct' number of topics taken from the marked up textual data, and therefore can be used as a criterion for selecting the right number of topics. Section 5 shows experimentally that the overall lexical composition of all the topics taken together is nearly invariant, that is, it is resistant to changing the number of topics in the greater part of the range of variation, but this invariance is intermittent and is described by a quasiperiodic function. We conclude that the T-invariance parameter must be accounted for while choosing the number of topics and that it must be included in the general theory of parameter optimization for topic models in the future research.

2. Problems of topic modeling and cluster analysis

2.1. Introduction to topic modeling

Topic modeling as a version of cluster analysis is based on the following propositions [16]:

1. Let D be a collection of text documents, and W—a set (dictionary) of all unique words. Each document $d \in D$ is a set of terms $w_1, \ldots, w_n$ from dictionary W.
2. It is assumed that there is a finite number of topics T, and every entry of word w in document d is associated with some topic $t \in T$. A topic is taken to mean a set of words that are often found together in a large number of documents.
3. A collection of documents is considered a random and independent selection of triads (w_i, d_i, t_i), $i = 1, \ldots, n$ from the discrete distribution p(w,d,t) over the finite probabilistic space $W \times D \times T$. Words w and documents d are observable variables, topic $t \in T$ is a latent (hidden) variable.
4. It is assumed that the order of terms in documents is not important for identifying topics (the 'bag of words' approach), and neither is the order of documents in the collection.

In TM, it is also assumed that probability p(w|d) of the occurrence of terms w in documents d can be expressed as a product of distributions p(w|t) and p(t|d). According to the formula of total probability and the hypothesis of conditional independence, we have the following expression [17]:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td} \tag{1}$$