ПРИМЕНЕНИЕ ЭНТРОПИЙНОГО ПОДХОДА К ПРОБЛЕМЕ ВЫБОРА ЧИСЛА ТЕМ В ТЕМАТИЧЕСКИХ МОДЕЛЯХ

Кольцов С.Н.
(*Высшая школа экономики*
Национальный исследовательский университет

Статистическая физика все чаще используется для описания объектов и процессов, выходящих за пределы физических явлений. Так, большие массивы текстовых данных, стремительно накапливающиеся в интернете в последнее десятилетие, требуют все более сложных методов их автоматической обработки и моделирования. Одним из таких методом является тематическое моделирование [1].

Коллекция текстов рассматривается в виде сложной неравновесной системы. Соответственно, для таких систем можно сформулировать понятия свободной энергии и энтропии. Эволюция такой системы, происходящая при изменении числа тем, может быть описана в терминах деформированной энтропии Реньи и Тсаллиса.

Неравновесная свободная энергия тематической модели выражается следующей формулой [2]:

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T = -\ln\left(\frac{\sum_{t=1}^{T}\sum_{n=1}^{N} P_{nt}}{T}\right) - T \cdot \ln\left(\frac{N_{k1}}{N \cdot T}\right),$$

где N_{k1} – число состояний, в которых $P_{nt} > 1/N$, $(N \cdot T)$ – общее число всех состояний, T – число тем (варьируемый параметр), N – размер словаря уникальных слов, E_0 S_0,– энергия и энтропия системы при начальном распределении, которые соответствуют максимуму энтропии. В свою очередь, свободная энергия может быть выражена в терминах энтропии Реньи за счет использования эскорт распределения $S_{q=1/T}^{R} = \frac{F}{T-1}$, где параметр деформации q = 1/T является обратной величиной числа тем. Энтропия Тсаллиса также выражается через энтропию Реньи.

Исходя из того, что вариация числа тем в тематическом моделировании существенно влияет на величину неравновесной энтропии [2], а также, что информации представляется как энтропия, взятая с обратным знаком, то есть максимум энтропии соответствует минимуму информации, то поиск оптимального числа тем в сложных системах можно свести к поиску минимума неравновесной энтропии.

В данной работе анализировалось поведение русскоязычной и англоязычной коллекций документов в терминах деформированной неравновесной статистики. Исследование проводилось для 4 тематических моделей: pLSA, VLDA, LDA, GLDA. В каждом вычислительном эксперименте для каждой модели измерялось величины энтропии Реньи, Тсаллиса и определялся минимум энтропий. Следует отметит, что для англоязычной коллекции были известны результаты кластерного анализа, которые проводились другими авторам [3]. Сравнение результатов тематического моделирования и кластерного анализа показывает, что минимум деформированной энтропии соответствует оптимальным кластерным решениям.

В данной работе сформулирован энтропийный подход к анализу поведения сложных текстовых систем, который позволяет определять оптимальное число тем в тематических моделях на основе поиска экстремума деформированной энтропии.

Литература

1. Griffiths T, M. Steyvers. Finding Scientific Topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101 (Suppl. 1). P. 5228–5335.
2. Koltcov S. N., A thermodynamic approach to selecting a number of clusters based on topic modeling, Technical Physics Letters, 43(6), 584-586.
3. Basu S, I. Davidson, Wagstaff K. Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall, 2008.