01

Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования

© *С.Н. Кольцов*

НИУ Высшая школа экономики, Санкт-Петербург
E-mail: skoltsov@hse.ru

Поступило в Редакцию 31 января 2017 г.

Применен термодинамический подход к решению проблемы выбора числа кластеров/тем в тематическом моделировании. Сформулированы основные положения подхода, и исследуется поведение тематических моделей при вариации температуры. При помощи термодинамического формализма показано существование энтропийного фазового перехода в тематических моделях и сформулированы критерии выбора оптимального числа тем/кластеров.

DOI: 10.21883/PJTF.2017.12.44713.16725

Термодинамический формализм, реализованный на основе минимизации свободной энергии, успешно применяется в различных областях, таких как обработка изображений [1], нейронные сети [2], кластерный анализ [3]. Существенное развитие методов кластеризации произошло в рамках тематического моделирования (ТМ) [4,5]. В ТМ решается задача восстановления исходного многомерного распределения в виде смеси мультиномиальных распределений со скрытыми параметрами. Одной из нерешенных проблем в ТМ является выбор числа распределений в смеси. Причем эта проблема возникает как в кластерном анализе, сетевом анализе [6], так и при исследовании фазовых переходов веществ с различной пространственной структурой [7].

Поскольку ТМ ориентировано на работу с большими данными, совокупность документов и слов можно рассматривать как мезоскопическую систему из большого числа частиц (миллионы документов и слов в них), характеризующуюся термодинамическими величинами, такими как энергия, энтропия и свободная энергия. Исходя из этого, термодинамический подход к проблеме выбора числа в тематическом моделировании можно сформулировать в виде следующих положений:

1. В рассматриваемой информационной термодинамической системе общее количество слов и документов является константой, т. е. изменение объема отсутствует. 2. Под темой понимается состояние (аналог направления спина), которое может принимать каждое слово и документ в коллекции. 3. Информационная термодинамическая система является открытой и обменивается только энергией с внешней средой за счет изменения температуры. В данном подходе под температурой системы понимается число тем (или кластеров), которое задается извне. Изменение числа тем приводит к изменению числа состояний $N(T)$, энергии $E(T)$ и энтропии $S(T)$. 4. Энтропию системы можно выразить в виде логарифма от числа состояний с энергией E: $S = \ln\big(N(E)\big)$ [2]. 5. Энергию каждого элемента системы можно определить через вероятность элемента в системе [2], в контексте данной работы, через вероятность принадлежности слова к теме: $E_{nt} = -\ln(P_{nt})$, где n — номер слова в словаре, t — номер темы. Общепринятым подходом при исследовании термодинамических свойств систем является расчет статистической суммы, так как знание данной суммы позволяет вычислить различные термодинамические величины как функции от температуры. Для микроканонического ансамбля предполагается, что система в неравновесном состоянии будет находиться на части состояний с высокой вероятностью [8]. Статистическую сумму такой системы можно выразить следующим образом [2]: $Z = \int de^{E-TS(E)} = \int de^{F}$, где F — свободная энергия неравновесной системы. В вычислительных экспериментах по тематическому моделированию можно напрямую подсчитать число состояний с заданной энергией при вариации параметра T и тем самым построить зависимость свободной энергии информационной системы от числа тем. Минимум функции свободной энергии, и точка фазового перехода будет соответствовать оптимальному числу тем [3]. В данной работе нас интересует поведение свободной энергии, которая нормирована на число тем, т. е.

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

Результатом тематического моделирования является матрица Ф, содержащая распределения вероятностей принадлежности всех уникальных слоев к темам. Размер этой матрицы — NT, где N — число уникальных слов в матрице (число строк), T — число тем (число столбцов). При этом распределение вероятностей в ТМ таково, что сумма всех вероятностей по всем словам и всем темам равна T, т. е. полная энергия

данной информационной термодинамической системы равна числу тем. Необходимо отметить, что в ТМ в качестве начального распределения для матрицы Ф используется равномерное распределение, в котором вероятность всех слов одинакова и равна величине $1/N$ (что соответствует максимуму энтропии). В ходе тематического моделирования происходит переход к сильно неравновесному состоянию, которое характеризуется тем, что одна часть состояний имеет высокую вероятность $P_{nt} > 1/N$, а другая — низкую $P_{nt} < 1/N$ вероятность, близкую к нулю. Соответственно число состояний с вероятностью больше (или меньше) величины $1/N$ является функцией от числа тем. Основной вклад в общую величину свободной энергии дают именно состояния с высокой вероятностью [8], поэтому расчет числа таких состояний позволяет оценить поведение неравновесной свободной энергии как функции от числа тем.

В рамках данного исследования был проведен следующий набор компьютерных экспериментов. На наборе данных фиксированного размера: 101481 постов из социальной сети „Живой журнал“ и $N = 172939$ уникальных слоев — проводилось тематическое моделирование на основе LDA с сэмплированием Гиббса, при разном количестве тем. Число тем варьировалось в диапазоне $T = [30; 500]$. В области, где была обнаружена сисльная флуктуация неравновесной свободной энергии, модель просчитывалась с шагом в одну тему. Известно что процедура сэмплирования Гиббса обладает определенной нестабильностью [9], поэтому величина свободной энергии $F(t)$ неравновесного состояния усреднялась по трем расчетам для каждого числа T. Кроме того, по трем запускам ТМ определялся уровень флуктуации тематического моделирования. Энергия и энтропия неравновесного состояния при каждой величине T рассчитывались следующим образом:

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

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

где N_t — число состояний, в которых $P_{nt} > 1$; (NT) — общее число всех состояний; T — число тем (варьируемый параметр); N — размер словаря уникальных слов; E_0, S_0 — энергия и энтропия системы

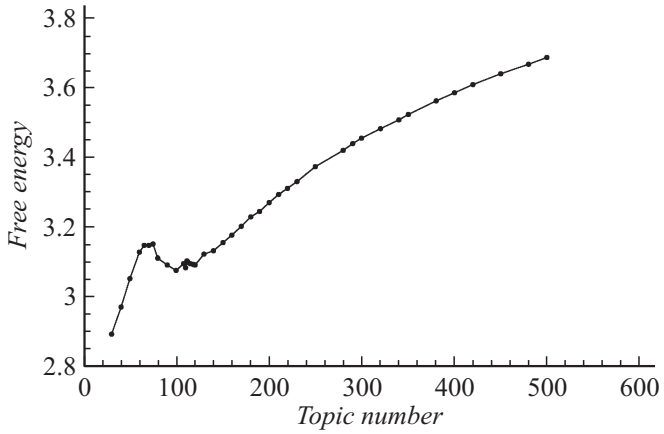

Рис. 1. Зависимость свободной энергии от числа тем.

при начальном равномерном распределении. На основании полученных функциональных зависимостей $E(T)$ и $S(T)$ была построена нормированная функция свободной энергии $F(T)/T$ (рис. 1).

Анализ полученных результатов показывает, что можно выделить две области, в которых наблюдается линейно-образная зависимость свободной энергии от числа тем. Область диапазона тем $T = [100; 118]$ соответствует переходной области, в которой изменение числа тем не сильно влияет на величину свободной энергии, и в этой области наблюдается минимум неравновесной свободной энергии. Более точное разделение областей можно получить за счет анализа второй производной свободной энергии по температуре, которая определяет теплоемкость системы. В данной работе мы определяем понятие информационной тематической емкости при фиксированном числе слов следующим образом: $C_{inf} = \left(\frac{d^2F}{dT^2}\right)$, где C_{inf} — информационная тематическая емкость, F — свободная энергия информационной системы, T — число тем. Тематическая емкость характеризует изменение энергии системы при изменении количества тем на единицу. График тематической емкости в зависимости от числа тем приведен на рис. 2. Сильнейшие скачки тематической емкости в диапазоне тем [100–118] соответствуют так называемому энтропийному фазовому переходу [10]. Таким образом,

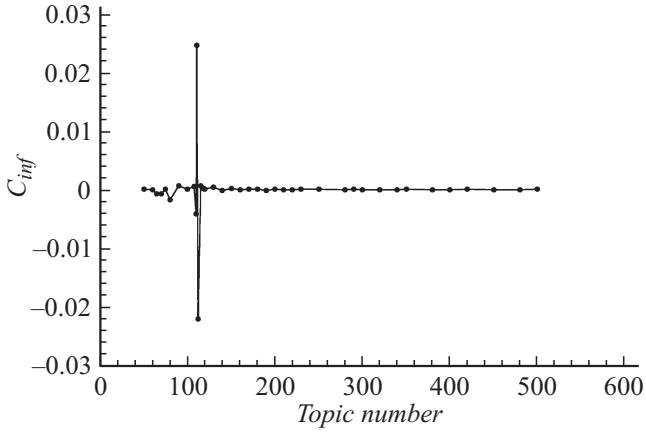

Рис. 2. Зависимость тематической емкости от числа тем.

в численных экспериментах по ТМ получены два фазовых состояния информационной системы. Первая фаза соответствует области, в которой небольшие изменения числа тем приводят к достаточно сильному изменению величины свободной энергии. При этом средняя величина флуктуации свободной энергии в первой фазе $\Delta F(T) \cong 0.8\%$ от средней величины свободной энергии. Внутри области фазового перехода флуктуация составляет в среднем $\Delta F(T) \cong 0.7\%$. Вторая фаза информационной системы соответствует области (начиная со 120 тем), в которой флуктуация свободной энергии почти в три раза меньше $\Delta F(T) \cong 0.2\%$. Исходя из этого, можно сделать предварительный вывод о том, что стабильност ТМ может служить одним из критериев для выбора числа тем. Таким образом, критериями выбора оптимального числа тем в ТМ могут служить следующие положения. Во-первых, в качестве оптимальной области следует выбирать фазовое состояние с наименьшей величиной флуктуаций. Во-вторых, следует выбирать область, в которой присутсвует минимум свободной энергии. Следовательно, на роль оптимальной области претендует второе фазовое состояние информационной системы. Однако, так как с увеличением числа тем система приходит к равномерному распределению, что соответствует максимуму энтропии, оптимальным числом тем для данной коллекции можно считать начало второй фазы $\cong 120$ тем.

Таким образом, применение термодинамического формализма позволяет глубже продвинуться в понимании поведения массивов больших данных при изменении таких параметров, как число распределений в смеси. При этом выбор оптимального размера смеси распределений в тематическом моделировании можно обосновать с помощью анализа поведения информационной термодинамической системы в разных фазовых состояниях. Дальнейшее развитие данного подхода возможно за счет применения мультифрактального формализма.

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2016 году.

Список литературы

[1] *Friston K., Levin M., Sengupta B., Pezzulo G.* // J. Royal. Soc. Interface. 2015. V. 12. P. 20141383. doi:10.1098/rsif.2014.1383
[2] *Tkacik G., Mora T., Marre O.* et al. Thermodynamics for a network of neurons: Signatures of criticality. 2014.
[3] *Rose K., Gurewitz E., Fox G.* // Phys. Rev. Lett. 1990. V. 65 (8). P. 945–948.
[4] *Griffiths T., Steyvers M.* // Proc. Nation. Acad. Sci. 2004. V. 101 (Suppl. 1). P. 5228–5335.
[5] *Blei D.M., Ng A.Y., Jordan M.I.* // J. Mach. Learn. Res. 2003. V. 3 (4-5). P. 993–1022.
[6] *Fortunato S.* // Phys. Reports. 2010. V. 486. Iss. 3-5. P. 75–174.
[7] *Иванской В.А.* // ЖТФ. 2008. Т. 78. В. 4. С. 65.
[8] *Абаимов С.Г.* Статистическая физика сложных систем. М.: УРСС, 2011.
[9] *Кольцов С.Н., Николенко С.И., Кольцова Е.Ю.* // ЖТФ. 2016. Т. 42. В. 16. С. 21–25.
[10] *Башкиров А.Г.* // ТМФ. 2006. Т. 149. № 2. С. 299–317.