Кластеризация коллекции текстов*

А. А. Романенко
angriff07@gmail.com
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В работе предлагается метод кластеризации текстовой коллекции с помощью стандартных метрических алгоритмов, например, K-means. Для этого вводится функция расстояния между текстами, учитывающая «схожесть» лексики используемой в тексте. В работе также исследуется соответствие между введенным расстоянием на множестве реальных текстов и близостью тематик этих текстов. Возможность кластеризации и соответствие ее результатов с заранее известным распределением текстов по тематике исследована в вычислительном эксперименте на синтетической коллекции текстов.

Ключевые слова: *информационный поиск, метрические алгоритмы кластеризации, кластеризация текстов, K-means.*

Feature selection and stepwise logistic regression for credit scoring*

A. A. Romanenko
Moscow Institute of Physics and Technology

The article suggests a method of clustering text collection based on classical algorithms of clustering, for example, K-mean. The authors consider metric between texts taking into account similarity of their vocabularies. Also the authors investigate applicability of this metric to measure distance between real texts. The computational experiment compares results of clustering with given distribution of texts over the set of topics.

Keywords: *text clustering, algorithms of clustering, K-means, information retrieval.*

Введение

Кластеризация текстов может применяться для выделения из текстовой коллекции групп текстов одинаковой тематики. Эта задача относится к задачам поиска скрытой неструктурированной информации. Из-за больших размеров текстовых коллекций и из-за субъективности восприятия читателя темы текста оценить качество кластеризации сложно. Поэтому пока нет общепринятого функционала качества кластеризации текстовых коллекций и алгоритма, являющегося абсолютно лучшим.

Кластеризацию текстов можно провести с помощью вероятностных методов [1], например с помощью вероятностного латентного семантического анализа (англ. *PLSA — probabilistic latent semantic analysis*) [2] или латентного размещения Дирихле (англ. *LDA — latent Dirichlet allocation*) [3]. В данной работе ставиться задача кластеризации текстовой коллекции с помощью стандартных метрических алгоритмов кластеризации, например K-means [4], FOREL [5], C-means [6]. Для этого на множестве документов предлагается ввести функцию расстояния.

Пусть есть некоторое множество слов русского языка, каждое из которых хотя бы раз встретилось в одном из документов текстовой коллекции. Назовем это множество словарем. В данной работе под документом будем понимать неупорядоченное множество слов из словаря. Слова в документе могут повторяться. Тогда каждому документу можно поставить в соответствие вектор, содержащий информацию о словарном составе документа. Размерность этого вектора равна количеству слов в словаре. Тогда расстояние между

Научный руководитель В. В. Стрижов

документами можно ввести как расстояние между векторами, соответствующими этим документам.

Для улучшения работы алгоритма предлагается сделать предобработку текстов. Во-первых, предлагается привести все слова к своей начальной лексической форме и удалить все знаки препинания. Во-вторых, предлагается убрать из текста слова, встречающиеся в нем малое количество раз, а также слова, встречающиеся в большинстве текстов (стоп-слова). В-третьих, предлагается воспользоваться методикой TFIDF (от англ. *TF — term frequency, IDF — inverse document frequency*), описанной в [7].

Для тестирования предложенного метода кластеризации было проведены эксперименты на синтетических данных и на реальных текстах. Цель эксперимента на синтетических данных — проверить возможность кластеризации и то, насколько эта кластеризация соответствовала заранее известному распределения текстов по тематике. Словарь, используемый для генерации текстов, был разбит на множества слов, относящихся к определенной теме. Текст относился к той теме, к которой относилось большинство слов текста.

Цель эксперимента на реальных данных — изучить, отражают ли метрики, рассматриваемые в работе, действительные расстояния между текстами, т. е. сравнить расстояния между реальными текстами на введенных метриках с экспертными расстояниями между ними. Тексты, используемые в эксперименте, — это работы студентов Восточной экономико-юридической гуманитарной академии.

Далее будет представлена математическая постановка задачи и предлагаемое решение. Затем будет представлены результаты вычислительного эксперимента с использованием различных функций расстояний на множестве документов на синтетической коллекции документов.

Постановка задачи и предлагаемое решение задачи

Пусть $W = \{w_1, \ldots, w_{|W|}\}$ — заданное множество слов, словарь. Документом d назовем множество слов из W, порядок которых не важен:

$$d = \{w_j\}, \text{ где } w_j \in W \ \text{ — } j\text{-ое слово в документе } d, \ j = 1, \cdots, |d|.$$

Таким образом, документ имеет модель «мешка слов» [8].

Пусть $D = \{d_1, \ldots, d_{|D|}\}$ — множество всех текстовых документов, k — заданное число кластеров, на которое требуется разбить множество D.

Требуется задать функцию расстояния на множестве документов:

$$\rho(d_i, d_j) : D \times D \longrightarrow \mathbb{R}_+,$$

и провести кластеризацию текстовой коллекции.

Предлагается удалить из текстов стоп-слова и слова, встречающиеся не более одного раза в тексте, как шумовую составляющую. Стоп-слово формально определим как слово из некоторого заранее заданного списка S.

Представим каждый преобразованный документ в виде вектора:

$$\mathbf{d}_i = \begin{pmatrix} n(d_i, w_1) \\ \vdots \\ n(d_i, w_j) \\ \vdots \\ n(d_i, w_{|W|}) \end{pmatrix}, \tag{1}$$

где $n(d_i, w_j)$ — число вхождений слова $w_j \in W$ в текст d_i.

Далее используя это представление документа, ввести расстояние между документами как расстояние между векторами:

$$\rho(d_i, d_j) = \rho(\mathbf{d}_i, \mathbf{d}_j).$$

В качестве функции расстояния $\rho(\mathbf{x}, \mathbf{y})$ между векторами $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ можно взять метрику Минковского для различных значений p, в частности расстояние городских кварталов при $p = 1$

$$\rho_1(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k|,$$

Евклидово расстояние при $p = 2$

$$\rho_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

или расстояние Чебышева при $p = \infty$

$$\rho_\infty(\mathbf{x}, \mathbf{y}) = \max_{k=1,\ldots,n} |x_k - y_k|.$$

Предлагается, используя функцию расстояния $\rho(\mathbf{d}_i, \mathbf{d}_j)$, провести кластеризацию текстовой коллекции одним из метрических алгоритмов кластеризации (K-means, C-means, и т.п.).

Каждый метод кластеризации можно рассматривать как точный или приближённый алгоритм поиска оптимума некоторого функционала. Если y_j — номер кластера, к которому отнесет j-ый документ алгоритм кластеризации, то можно ввести следующие функционалы качества:

$$F_0 = \frac{\sum_{i<j} [y_i = y_j] \rho(\mathbf{d}_i, \mathbf{d}_j)}{\sum_{i<j} [y_i = y_j]} \to \min,$$

т.е. нужно минимизировать среднее внутрикластерное расстояние,

$$F_1 = \frac{\sum_{i<j} [y_i \neq y_j] \rho(\mathbf{d}_i, \mathbf{d}_j)}{\sum_{i<j} [y_i \neq y_j]} \to \max,$$

т.е. нужно максимизировать среднее межкластерное расстояние.

Чтобы учесть и внутрикластерное и межкластерное расстояние можно ввести функционал F :

$$F = \frac{F_0}{F_1} \to \min.$$

Вычислительный эксперимент на синтетических данных

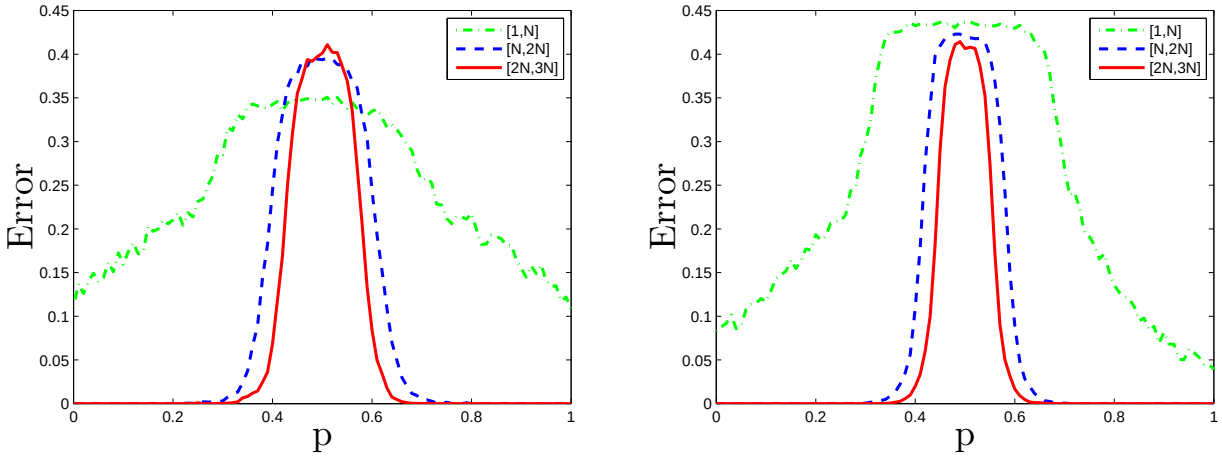
Формирование текстовой коллекции. Для иллюстрации работы алгоритма проведен эксперимент кластеризации на два кластера на синтетической коллекции документов. В качестве словаря был взят список из 50 слов. Предполагалось, что каждое слово из словаря относилось либо к теме 1, либо к теме 2. Каждый документ порождался следующим образом:

1. За длину документа принималось некоторое целое произвольное число из отрезка $[A, B]$.
2. Каждое слово документа, относящегося к теме 1, с вероятностью p выбиралось произвольным образом из списка слов, относящихся к теме 1, и с вероятностью $(1 - p)$ — из списка слов, относящихся к теме 2.
3. Каждое слово документа, относящегося к теме 2, с вероятностью p выбиралось произвольным образом из списка слов, относящихся к теме 2, и с вероятностью $(1 - p)$ — из списка слов, относящихся к теме 1.

Таким образом была получена текстовая коллекция состоящая из 100 текстов: 50 текстов, относящихся к теме 1, и 50 текстов, относящихся к теме 2.

Результаты кластеризации. Сопоставив каждому документу из текстовой коллекции вектор описанным выше способом (1), используем алгоритм кластеризации K-means для кластеризации на два кластера. Ниже представлены зависимости ошибки кластеризации $Error$ от параметра p для метрик ρ_1 и ρ_2 при разных A и B, влияющих на длину документов. Под ошибкой кластеризации здесь понимается доля документов, ошибочно попавших в кластер с документами другой тематики.

Стоит отметить, что так как в алгоритме используется рандомизация, то графики зависимостей получаются не гладкими. Чтобы этого избежать, проводилось 500 экспериментов, а затем ошибка кластеризации усреднялась.



(a) Расстояние городских кварталов

(b) Евклидово расстояние

Рис. 1. Зависимость ошибки кластеризации от параметра p, если длина текста — произвольное число из отрезка $[1, N]$, $[N, 2N]$ или $[2N, 3N]$.

Из рис. 1 видно, что наибольшая ошибка кластеризации наблюдается при p близком к 0.5, что соответствует наибольшему количеству шума в документе. Если же $p \approx 1$ или $p \approx 0$, то ошибка кластеризации мала. Как видно на рис. 1, для метрики ρ_2 с увеличением длины документов при фиксированном p ошибка кластеризации уменьшается. Это же справедливо и для метрики ρ_1, если p не лежит в окрестности 0.5. Если же $p \approx 0.5$, то чем больше размер документа, тем больше ошибка кластеризации. При этом видно, что при использовании метрики ρ_2 ошибка кластеризации меньше, чем при использовании метрики ρ_1.

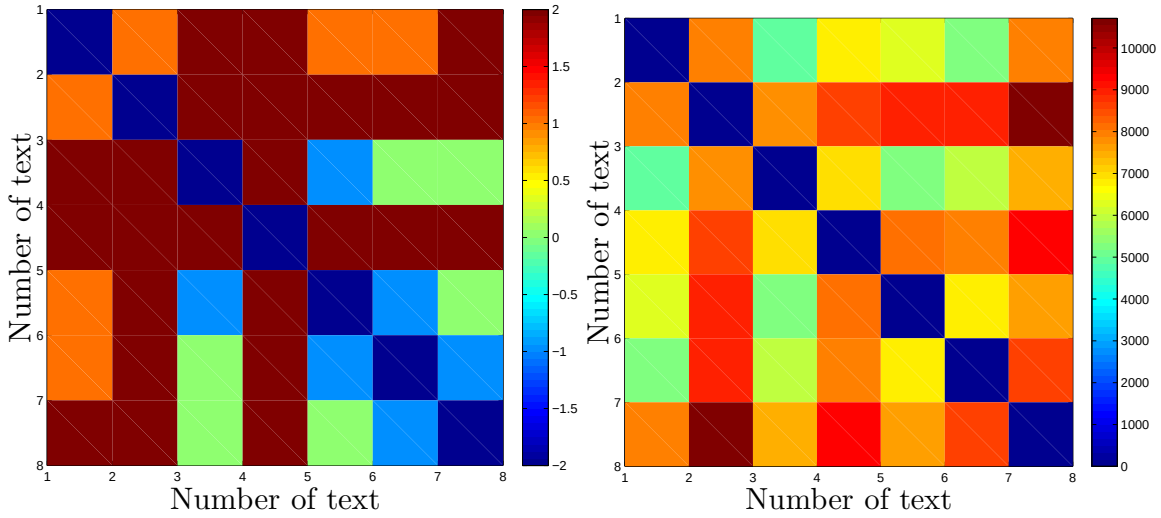
Код для проведения эксперимента можно взять здесь [9].

Вычислительный эксперимент на реальных данных

Описание эксперимента. Для того чтобы убедиться, что рассматриваемые метрики адекватно описывают расстояние между реальными документами, был проведен следующий эксперимент. Из коллекции работ студентов Восточной Экономико-юридической Гуманитарной Академии было взято восемь произвольных текстов. Оказалась, что тексты имели следующие темы:

1. «Система защиты трудового права»
2. «Происхождение государства и права»
3. «Предпринимательство. Сущность, формы и современные особенности»
4. «Оценка состояния новорожденного по его поведению и мимике»

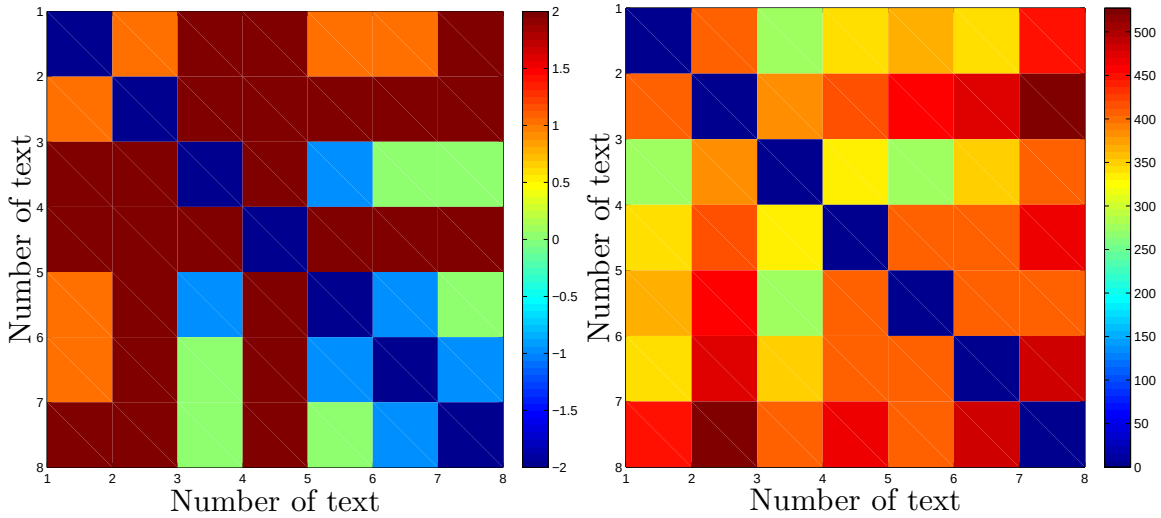
5. «Экономика организации предприятия»
6. «Государственное и муниципальное правление объектами здравоохранения»
7. «Анализ качества продукции ООО «Оренбургский хлебозавод»»
8. «Правоотношения»



(a) Экспертная матрица расстояний

(b) Матрица расстояний между текстами с метрикой ρ_1

Рис. 2. Матрицы расстояний между текстами



(a) Экспертная матрица расстояний

(b) Матрица расстояний между текстами с метрикой ρ_2

Рис. 3. Матрицы расстояний между текстами

Данные восемь текстов были предложены эксперту для определения степени отличия тематики содержимого текстов. Эксперт давал оценку в следующей лингвистической шкале:

-2 — «содержимое текстов очень похоже по тематике»
-1 — «содержимое текстов скорее похоже по тематике, чем отлично»

0 — «трудно определить, похоже содержимое или отлично»
1 — «содержимое текстов скорее отлично по тематике, чем похоже»
2 — «содержимое текстов сильно отличается по тематике»

Также эти тексты были представлены как элементы векторного пространства описанным выше способом (1), и было посчитано расстояние между ними на основании метрик ρ_1 и ρ_2.

Результаты эксперимента и вывод. На рис. 2 и 3 изображены для для сравнения экспертная матрица расстояний и матрицы расстояний между документами, подсчитанных с метриками ρ_1 и ρ_2.

Из рисунков 2, 3 видно, что резкие различия в тематике рассматриваемые метрики выявить могут. Так, например, по мнению эксперта, текст №2 сильно отличается от всех текстов. Действительно, он находится на большом расстоянии от оставшихся документов. Но с определением тонких различий в тематике они справляются хуже.

Код для проведения эксперимента и данные можно взять здесь [9].

Заключение

В работе описан способ представления документа как элемента векторного пространста. Это дает возможность ввести функцию расстояния между документами и кластеризовать коллекцию документов на основе метрических алгоритмов кластеризации. В эксперименте на синтетической коллекции документов исследуется соответствие результатов кластеризации коллекции с заранее известным распределением текстов по тематике.

Также в работе исследуется соответствие между введенным расстоянием на множестве реальных документов и близостью тематик реальных документов. Для этого проведен эксперимент с реальными текстами.

Литература

[1] A. Daud, J. Li, L. Zhou, F. Muhammad. *Knowledge discovery through directed probabilistic topic models: a survey.* Frontiers of Computer Science in China. 2010.

[2] Hofmann T. *Probabilistic latent semantic indexing.* SIGIR '99. New York, NY, USA: ACM, 1999.

[3] Blei D. M., Ng A. Y., Jordan M. I. *Latent dirichlet allocation.* 2003.

[4] Hartigan J. A., Wong M. A. *Algorithm as 136: A k-means clustering algorithm.* 1978.

[5] Н.Г.Загоруйко, В.Н.Ёлкина, Г.С.Лбов. *Алгоритмы обнаружения эмпирических закономерностей.* Новосибирск: Наука, 1985.

[6] Pal N. R., Bezdek J. C. *On cluster validity for the fuzzy c-means model.* 1995.

[7] Manning C. D., Raghavan P., Schütze H. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press, 2008.

[8] Lewis D. D. *Naive (bayes) at forty: The independence assumption in information retrieval.* Springer Verlag, 1998.

[9] А.А.Романенко *Кластеризация коллекции текстов: вычислительный эксперимент.* 2012. `http://bit.ly/IT2OXW`.