электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 81'322.2

Исследование влияния прореживания терминов на точность кластеризации алгоритмом meanshift

Гречищев К.М., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
кафедра «Компьютерные системы и сети»

Научный руководитель: Самарев Р.С., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
bauman@bmstu.ru

Введение

Сегодня многие задачи, связанные с автоматической обработкой текстов на естественном языке, решаются с использованием модели термин-документ (англ. vector space model)[1]. В данной модели каждый документ представляется вектором в n-мерном пространстве термов. Недостаток модели, особенно заметный при обработке текстовых коллекций большого объема, состоит в том, что число уникальных термов быстро растет при увеличении размера коллекции. Это приводит к тому, что, зачастую, размерность векторов достигает десятков тысяч измерений. Обработка векторов такой длины требует большого расхода памяти и высоких затрат машинного времени.

В настоящее время существует ряд методов снижения размерности модели представления документов, например, методы, основанные на латентно-семантическом анализе и сингулярном разложении[2]. Недостатки данных методов состоят в потере части информации о документах, а также в потере четкой привязки размерности к конкретному уникальному терму, что затрудняет анализ результатов работы применяемого алгоритма, например, алгоритма кластеризации.

В данной работе предлагается метод снижения размерности, который заключается в исключении из модели уникальных терминов с низкой частотой появления в коллекции. Наиболее распространенными алгоритмами кластеризации в настоящее время являются алгоритм К-средних[4] и Meanshift[5], однако алгоритму Meanshift обычно уделяется меньшее внимание, поэтому в этой работе оценка будет производиться именно для этого алгоритма.

Модель термин-документ

Рассмотрим множество (коллекцию) документов D. Функцию $M: D \to V \subset R^m$, где m – число уникальных терминов коллекции D, называют моделью термин-документ. Тогда:

$$\forall d_i \in D: \; M(d_i) = v_i = \begin{pmatrix} v_{i,1} \\ v_{i,2} \\ \dots \\ v_{i,m} \end{pmatrix},$$ где величина $v_{i,j}$ может быть вычислена различными способами, например, по формуле $v_{i,j} = TF_{i,j} \cdot IDF_j$, где $TF_{i,j}$ - частота термина в документе, IDF_j - инверсная частота термина в коллекции.

Величину v_i называют представлением документа d_i. Введем функцию расстояния (метрику) $\rho: V \times V \to R$. Расстоянием между документами будем называть значение этой функции для представлений этих документов.

В качестве функции расстояния обычно используют меру косинуса угла (англ. Cosine Distance) [1].

Особенности текстовых коллекций

Интуитивно понятно, что не все слова в естественном языке употребляются с одинаковой частотой. Проведенные исследования[3] показывают, что распределение слов естественного языка по частоте подчиняется закону Ципфа (Zipf law), который является частным случаем распределения Парето:

$$p(x, Q) = \begin{cases} \frac{\alpha}{Q} \cdot \left(\frac{Q}{x}\right)^{\alpha+1}, & x \geq Q \\ 0, & x < Q \end{cases}$$, где α и Q - параметры распределение Парето.

На практике отдельно взятая коллекция текстовых документов зачастую имеет специфические особенности, из-за которых распределение слов по частоте будет иметь другой вид. Кроме того, в модели термин-документ часто используются не только однословные термы, но и термы, состоящие из двух или трех слов.

В распоряжении авторов была коллекция из 12830 запросов в службу технической поддержки. Документы этой коллекции содержали большое количество серийных номеров, названий устройств и версий программного продукта, жаргонных выражений, а

также опечаток. Кроме того, при анализе коллекции не использовался морфологический анализатор.

В таблице приведено число уникальных термов, которые встречаются в текстовой коллекции. N – размер термина в словах. Видно, что от 72 до 90% термов встречаются лишь в одном документе из всей коллекции. Таким образом, если удалить эти редкие термы можно достичь значительного снижения размерности модели. Однако при этом необходимо оценить влияние удаляемых размерностей на меру расстояния между документами.

	N = 1		N = 2		N=3		N=1,2,3	
Всего термов	33 199	100.00%	140 464	100.00%	162 689	100.00%	336 352	100.00%
Число термов, встречающихся лишь в одном документе	23 908	72.01%	124 736	88.80%	156 480	96.18%	305 124	90.72%
Лишь в 2-х документах	3 017	9.09%	9 637	6.86%	4 478	2.75%	17 132	5.09%
Лишь в 3-х документах	1 397	4.21%	2 849	2.03%	989	0.61%	5 235	1.56%
Лишь в 4-х документах	775	2.33%	1 175	0.84%	270	0.17%	2 220	0.66%

Оценка метода

Сущность предлагаемого подхода заключается в удалении из модели всех размерностей, соответствующих терминами, частота появления которых в коллекции ниже заданной. Формально удаление из модели представления одного из терминов (l-го термина t_l) эквивалентно переходу от модели представления M к модели $M_1 : D \to \Lambda \subset R^{m-1}$. Тогда:

$M_1(d_i) = \lambda_i = \begin{pmatrix} \lambda_{i,1} \\ \dots \\ \lambda_{i,m-1} \end{pmatrix}$, где вектор λ_i может быть получен из вектора v_i вычеркиванием l-го элемента v_l.

Рассмотрим эффект, который оказывает удаление термина на величину расстояния между документами в общем случае:

$$\cos(<\vec{v_1},\vec{v_2}) = \frac{\vec{v_1}\cdot\vec{v_2}}{\left|\vec{v_1}\right|\cdot\left|\vec{v_2}\right|};$$

$$\cos(<\vec{\lambda_1},\vec{\lambda_2}) = \frac{\vec{v_1}\cdot\vec{v_2} - v_{1,l}v_{2,l}}{\sqrt{\left|\vec{v_1}\right|^2 - v_{1,l}^2}\cdot\sqrt{\left|\vec{v_2}\right|^2 - v_{2,1}^2}}.$$

Обозначим буквой Δ абсолютную погрешность расстояния. Тогда в общем случае:

$$\Delta = \rho(\vec{v_1},\vec{v_2}) - \rho(\vec{\lambda_1},\vec{\lambda_2}) = 1 - \cos(<\vec{v_1},\vec{v_2}) - 1 + \cos(<\vec{\lambda_1},\vec{\lambda_2}) =$$

$$= \frac{\vec{v_1}\cdot\vec{v_2} - v_{1,l}v_{2,l}}{\left|\vec{v_1}\right|\left|\vec{v_2}\right|\sqrt{1-\frac{v_{1,l}^2}{\left|\vec{v_1}\right|^2}}\cdot\sqrt{1-\frac{v_{2,1}^2}{\left|\vec{v_2}\right|^2}}} - \frac{\vec{v_1}\cdot\vec{v_2}}{\left|\vec{v_1}\right|\cdot\left|\vec{v_2}\right|} = \frac{\vec{v_1}\cdot\vec{v_2}\cdot\left(1-\sqrt{1-\frac{v_{1,l}^2}{\left|\vec{v_1}\right|^2}}\cdot\sqrt{1-\frac{v_{2,1}^2}{\left|\vec{v_2}\right|^2}}\right) - v_{1,l}v_{2,l}}{\left|\vec{v_1}\right|\left|\vec{v_2}\right|\sqrt{1-\frac{v_{1,l}^2}{\left|\vec{v_1}\right|^2}}\cdot\sqrt{1-\frac{v_{2,1}^2}{\left|\vec{v_2}\right|^2}}} \quad (1)$$

Если ни один из документов не содержал терма t_l , то

$$\forall\ d_1,d_2 \in D : t_l \notin d_1, t_l \notin d_2 \Rightarrow v_{1,l} = 0, v_{2,l} = 0 \Rightarrow \Delta = 0$$

Следовательно, при удалении измерения l расстояние между документами, не содержащими терма t_l, не изменяется. В промежуточном случае, когда термин t_l содержится лишь в одном документе, имеем из (1):

$$\forall\ d_1,d_2 \in D : t_l \in d_1, t_l \notin d_2 \Rightarrow v_{1,l} \neq 0, v_{2,l} = 0 \Rightarrow$$

$$\Delta = \frac{\vec{v_1}\cdot\vec{v_2}\cdot\left(1-\sqrt{1-\frac{v_{1,l}^2}{\left|\vec{v_1}\right|^2}}\right)}{\left|\vec{v_1}\right|\left|\vec{v_2}\right|\sqrt{1-\frac{v_{1,l}^2}{\left|\vec{v_1}\right|^2}}} = \frac{\vec{v_1}\cdot\vec{v_2}}{\left|\vec{v_1}\right|\left|\vec{v_2}\right|}\cdot\left(\sqrt{\frac{\left|\vec{v_1}\right|^2}{\left|\vec{v_1}\right|^2 - v_{1,l}^2}} - 1\right) \quad (2)$$

На практике при формировании векторов $\vec{v_1}$ и $\vec{v_2}$ применяют процедуру нормализации, следовательно $\left|\vec{v_1}\right| = \left|\vec{v_2}\right| = 1$, тогда формула (1) существенно упроститься:

$$\Delta = \frac{\vec{v_1}\cdot\vec{v_2}\cdot\left(1-\sqrt{1-v_{1,l}^2}\cdot\sqrt{1-v_{2,1}^2}\right) - v_{1,l}v_{2,l}}{\sqrt{1-v_{1,l}^2}\cdot\sqrt{1-v_{2,1}^2}}.$$

Аналогично для (2):

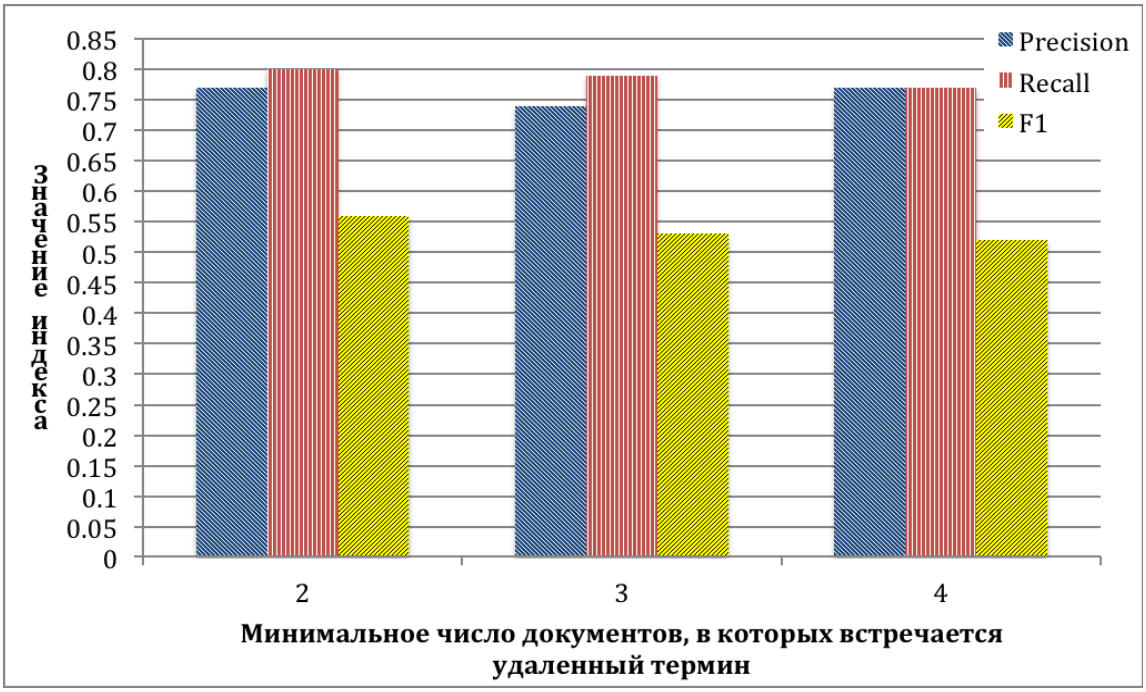
$$\Delta = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2} \cdot \left(1 - \sqrt{1 - v_{1,l}^2}\right)}{\sqrt{1 - v_{1,l}^2}}.$$

Эксперименты и результаты

Эксперимент заключался в оценке влияния удаления компонентов на качество работы алгоритма Meanshift[5] на исходной и прореженных выборках модели термин-документ.

Результаты кластеризации на полной модели, считались эталонными. С ними сравнивались результаты работы алгоритма на модели с удаленными терминами. Для сравнения двух кластерных структур использовались метрики precision, recall и F1[6]. Результаты эксперимента приведены на рисунке.

Результаты, полученные на измененной модели термин документы, отличаются от исходных на 15 – 20 %. Недостатком приведенного подхода является удаление части терминов из словаря. Эти термины не оказывают значительного влияния на меру расстояния, но могут быть важны для последующего анализа результатов работы алгоритма. Однако это недостаток может быть компенсирован использованием словаря термов, удаление которых запрещено, или повторной индексацией документов перед анализом результатов.



Выводы

Рассмотрен метод снижения размерности модели термин-документ. Получены теоретические оценки абсолютной погрешности расстояний при удалении из коллекции одного терма.

Подход проверен на тестовой коллекции из 12 тысяч запросов в техническую поддержку. Проведена оценка влияния на алгоритм кластеризации MeanShift на исходной и измененной модели термин документ. Установлено, что различие между полученными результатами составило порядка 15-20%. Замечено, что использование метода снижения размерности модели позволяет более точно адаптировать ее под решаемую задачу.

Предложенный подход может быть использован для снижения размерности модели термин-документ при использовании косинусной меры расстояния в задачах классификации, кластеризации текстовых документов. Направлением будущей работы является оценка влияния снижения размерности исходной модели на результаты кластеризации другими алгоритмами кластеризации.

Список литературы

1. Feldman R., Sanger J. The Text Mining Handbook. New York, Cambridge University Press, 2007. 410 p.
2. Aggarwal C., Zhai C. A survey of text clustering algorithms. Available at: http://charuaggarwal.net/text-cluster.pdf, accessed 07.10.2014.
3. Zipf G.K. Human Behavior and the Principle of least Effort. New York, Addison-Wesley Press, 1949. 537 p.
4. Ту Дж., Гонсалес Р. Алгоритм К внутригрупповых средних // Принципы распознавания образов. М.: Мир, 1978. Гл. 3.3.5. С. 109-112.
5. Comaniciu D., Meer. P. Mean shift: A robust approach toward feature space analysis // Pattern Analysis and Machine Intelligence, IEEE Transactions. № 24, 2002. P. 603-619. DOI: 10.1109/34.1000236.
6. Сивоголовко Е. Оценка качества четкой кластеризации. Режим доступа: http://synthesis.ipi.ac.ru/sigmod/seminar/s20111124 (дата обращения: 07.10.2014).