АНАЛИЗ МЕТОДОВ КЛАСТЕРИЗАЦИИ ПРОИЗВОЛЬНЫХ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

Ивахненко Максим
Зверева Дарья
Гавенко Ирина
Жарков Андрей
Кривчанский Николай

Московский физико-технический институт (государственный университет);
кафедра Компьютерной лингвистики Abbyy.

В работе представлен алгоритм, позволяющий проводить кластеризацию в один клик, который в среднем достигает более высоких показателей, чем любой из популярных сейчас алгоритмов кластеризации. В ходе работы был проведён обзор существующих алгоритмов кластеризации и анализ результатов их работы на различных датасетах. На основе результатов нашего исследования мы выбрали алгоритм и набор параметров и оптимизаций, который в среднем получает лучший результат без дополнительной информации по задаче. Лучшим результатом мы называем максимальную компактность похожих документов в кластерах. Также к данной статье прилагается ссылка на web-сервис, который позволяет читателю протестировать предложенную технологию на своём датасете.

ANALYSIS OF METHODS OF CLUSTERING OF ARBITRARY TEST COLLECTIONS

Krivchanskiy Nikolay
Ivakhnenko Maxim
Zvereva Daria
Gavenko Irina
Zharkov Andrey

This paper suggests an algorithm, which allows to perform text clusterization in a click of a button and also achieves better results in average, compared to most of the well-known clustering algorithms. As a part of the research, a number of clustering algorithms were tested on a variety of datasets on various settings. As a result we chose an algorithm and a number of features, that achieves best result in average. As an index of “wellness” we treat level of compactness of documents in cluster. Furthermore, in bibliography can be found a link to a web-service, which allows reader to try out suggested algorithm on dataset of preference.

Введение

Первоначальной задачей было представить алгоритм кластеризации, который будет в среднем, без какой-либо информации по задаче, предоставлять хороший результат. Так как результат работы вполне может быть востребован для малого и среднего бизнеса, было принято решение создать небольшой ресурс, на котором пользователь может после регистрации предъявить свой датасет и получить его в виде кластеров. В будущем пользователь сможет помечать документы, как сильно различные, для улучшения качества, однако для данной работы это не является предметом рассмотрения.

Исследование

На первом этапе была проведена работа по поиску и анализу существующих алгоритмов кластеризации и датасетов. Самыми популярными (и эффективными) были признаны BIRCH, Affinity Propagation, Hierarchical algorithm, Spectral algorithm, K-Means. Задача поиска датасета оказалась достаточно непростой. Авторы большей части статей, в которых описаны алгоритмы (включая перечисленные выше), приводили в пример цифры, полученные на новостном датасете Reuters. Этот датасет представляет большой интерес в задачах классификации и размечен под них. Тот факт, что в основном используется именно этот датасет, связан с тем, что задача кластеризации в мире преимущественно используется для подготовки данных под классификацию, поэтому в некотором смысле использование этого датасета оправдано. Однако стоит понимать, что цифры, полученные в задачах кластеризации, будут хуже, чем в алгоритмах классификации на тех же датасетах, просто исходя из специфики решаемой задачи. В ходе нашего исследования компания ABBYY предоставила нам ряд корпусов. В результате исследование проводилось на следующих корпусах: Reuters-1251, корпус патентов из Google Patents, корпус новостей BBC news, бразильский корпус новостей, корпус Российской государственной библиотеки и небольшой размеченный нами корпус шаблонов корпоративных документов ABBYY.

Алгоритм

Так как целью проекта было построение как можно более универсальной и простой для конечного пользователя системы, на первом шаге алгоритм в автоматическом режиме определяет язык, на котором написан текст. Следующим очевидным шагом является нормализация слов языка. Экспериментально было выявлено, что для русского языка лучшие результаты показывает лемматизация, для остальных языков применяется стеммизация. Извлечение признаков делается на основе обычного tf-idf алгоритма. Были проведены эксперименты с различными эмбеддингами, однако сильного улучшения качества они не дают. Более того, применение эмбеддингов сильно связывает руки, потому что найти в открытом доступе

качественные эмбеддинги для всех языков, для которых поддерживается стеммизация, тяжело. На следующем шаге мы уменьшаем размерность пространства признаков с помощью SVD. Без этого шага время выполнения возрастало бы экспоненциально; к тому же, с уменьшенной размерностью получается лучше отобразить точки на плоскость при визуализации. Финальным шагом предложенного алгоритма является запуск иерархической кластеризации, на основе всех полученных выше признаков. При использовании предложенного веб-сервиса в будущем пользователь сможет задать не точную, а приблизительную оценку предполагаемого количества кластеров (оценку сверху или снизу), а алгоритм сам будет выбирать оптимальный момент для остановки. Кроме того, планируется реализовать для пользователя возможность работы на частично размеченных данных. То есть, при желании пользователь сможет отметить документы, которые гарантированно должны лежать в одном кластере или же, наоборот, в разных.

Метрики качества

Метрики качества при кластеризации ─ довольно тонкий момент в анализе алгоритмов кластеризации. Очевидно, что в разных случаях можно анализировать и кластеризовать данные по разному, в зависимости от желания пользователя, но в общем случае важны два показателя: точность кластеризации относительно некоторого эталона и близость друг к другу похожих документов. Первую меру качества отражает **V-мера**, которая является адаптацией F-меры под анализ кластеров, она несёт похожий смысл и вычисляется аналогично. V-мера оперирует понятиями однородности и полноты кластера (аналогично с понятиями точность и полнота, применяемые в F-мере). Кластер называется однородным (homogenity = 1), если все элементы, которые были к нему отнесены, лежат в нём. Кластер называется полным (completness = 1), если все элементы, которые к нему должны быть отнесены, лежат в нём. Таким образом формула для вычисления V-меры выглядит следующим образом:

$$V_\beta = (1+\beta) \cdot \frac{h \cdot c}{\beta \cdot h + c}$$

Где β в нашем исследовании мы приняли за единицу. Подробнее про V-меру можно прочитать в [13].

Вторую меру качества отражает индекс Рэнда, представленный в таблице как **adjust_rand**. Индекс Рэнда оценивает насколько много из тех пар элементов, которые в эталоне находились в одном кластере, и тех пар элементов, которые находились в разных кластерах (не привязываясь к меткам кластеров), сохранили это состояние после кластеризации алгоритмом. Индекс нормализован таким образом, что 1 означает точное совпадение, -1 абсолютное несовпадение. Исправленный индекс Рэнда (Adjusted Rand Index) при сравнении кластеризаций X и Y вводится через таблицу случайностей:

$X \setminus Y$	Y_1	Y_2	...	Y_s	Сумма
X_1	n_{11}	n_{12}	...	n_{1s}	a_1
X_2	n_{21}	n_{11}	...	n_{2s}	a_2
...	...	...	...	...	...
X_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Сумма	b_1	b_2	...	b_s	

А сам индекс вычисляется по формуле:

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

Или, в соответствии с таблицей:

$$ARI = \frac{\sum_{ij} C_{n_{ij}}^2 - \left[\sum_i C_{a_i}^2 \sum_j C_{b_j}^2\right]/C_n^2}{\frac{1}{2}\left[\sum_i C_{a_i}^2 + \sum_j C_{b_j}^2\right] - \left[\sum_i C_{a_i}^2 \sum_j C_{b_j}^2\right]/C_n^2}$$

Результаты и оценка результатов.

Результаты работы алгоритмов представлены в таблице ниже

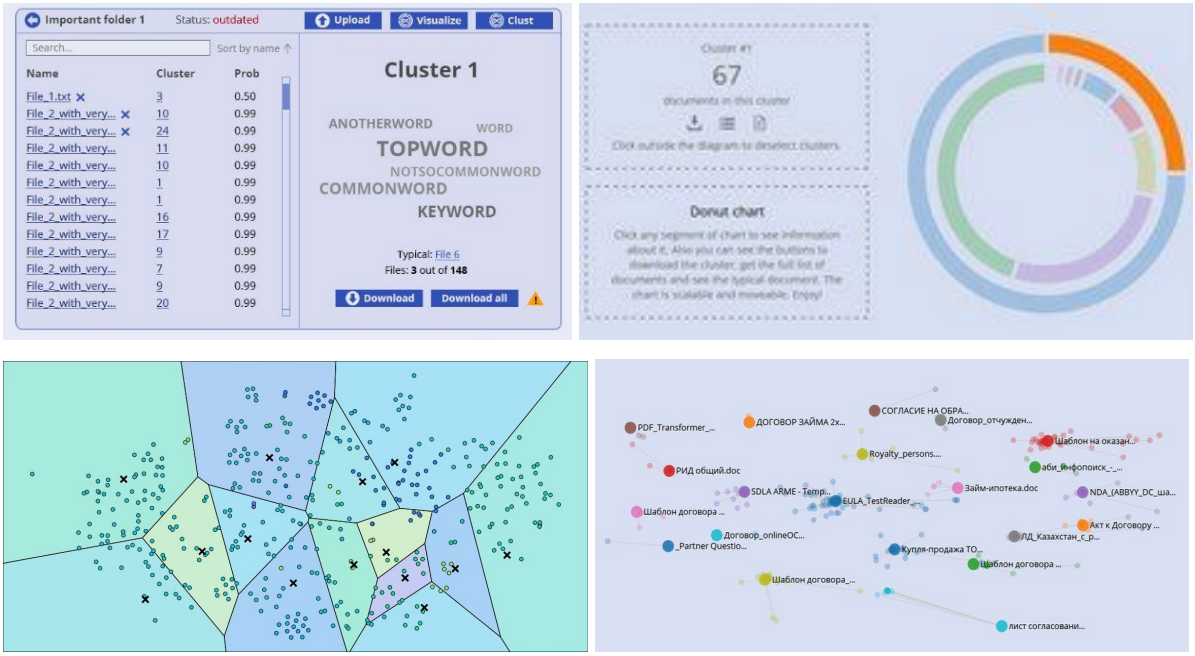
Корпус	Алгоритм	Эмбеддинг	homogeneity	completeness	v_measure	adjusted_rand	Время обучения	Кол-во кластеров	Кол-во классов	Число документов
Корпоративные документы ABBYY	Birch	-	0,814	0,625	0,707	0,502	0,049	27	23	470
	Affinity+Hierarchy		0,770	0,645	0,702	0,499	0,295	18		
	Hierarchy		0,797	0,625	0,701	0,469	0,233	23		
	Hierarchy		0,773	0,632	0,695	0,475	0,055	23		
	Birch		0,776	0,613	0,685	0,497	0,050	29		
	Hierarchy	ABBYY	0,700	0,690	0,695	0,616	0,162	26		
	Hierarchy		0,681	0,690	0,685	0,614	0,165	23		
	Affinity+Hierarchy		0,722	0,605	0,658	0,494	0,280	28		
	Birch		0,868	0,519	0,650	0,391	0,011	61		
	SpectralClustering		0,727	0,586	0,649	0,414	0,094	15		
	Affinity+Hierarchy		0,630	0,594	0,611	0,510	0,291	23		
	AffinityPropagation	Mail.ru	0,767	0,559	0,647	0,373	0,192	23		
	Birch		0,796	0,542	0,645	0,382	0,012	43		
	Hierarchy		0,774	0,549	0,642	0,367	0,052	37		
	KMeans		0,783	0,537	0,637	0,339	0,123	30		
	Affinity+Hierarchy		0,752	0,538	0,627	0,347	0,258	31		
BBC news	KMeans	-	0,827	0,830	0,828	0,857	29,680	5	5	2225
	SpectralClustering		0,816	0,817	0,816	0,852	0,799	5		
	Hierarchy		0,788	0,787	0,788	0,812	0,382	5		
	Affinity+Hierarchy		0,698	0,732	0,715	0,669	9,757	5		
Российская гос. библиотека	Hierarchy	-	0,749	0,507	0,605	0,255	0,096	23	9	100
	SpectralClustering		0,614	0,570	0,591	0,328	0,059	10		
	Affinity+Hierarchy		0,693	0,484	0,570	0,231	0,085	20		
	Hierarchy		0,582	0,551	0,566	0,323	0,094	10		
	AffinityPropagation		0,518	0,500	0,509	0,234	0,027	9		

	KMeans	Mail.ru	0,551	0,541	0,546	0,361	0,045	10		
	SpectralClustering		0,553	0,509	0,530	0,302	0,060	10		
	Affinity+Hierarchy		0,592	0,502	0,544	0,280	0,031	16		
	Hierarchy		0,523	0,472	0,496	0,232	0,012	11		
Brazillian News	SpectralClustering	-	0,829	0,952	0,886	0,772	0,023	5	6	103
	MiniBatchKMeans		0,788	0,916	0,847	0,728	0,030	5		
	Hierarchy		0,869	0,776	0,820	0,801	0,027	11		
	Birch		0,721	0,844	0,778	0,635	0,012	5		
	Hierarchy		0,732	0,795	0,762	0,698	0,017	6		
Reuters	AffinityPropagation	-	0.7897	0.2641	0.3958	0.0355	0.7897	209	10	8861
	Agglomerative		0.5255	0.5706	0.5471	0.5449	0.5255	10		
	Birch		0.5818	0.4775	0.5245	0.4487	0.5818	10		
	KMeans		0.5627	0.4461	0.4977	0.2666	0.5627	10		
	AffinityPropagation		0.8337	0.2206	0.3489	0.0319	0.8337	167	5	7745
	Agglomerative		0.4100	0.5998	0.4871	0.3764	0.4100	5		
	Birch		0.4741	0.4186	0.4446	0.3496	0.4741	5		
	KMeans		0.5164	0.4607	0.4869	0.2979	0.5164	5		

Ссылки на используемые корпуса доступны в статьях, на которые мы опирались при создании алгоритма. Исключением является корпус корпоративных документов Abbyy, в котором могла содержаться конфиденциальная информация о компании. Хочется отметить сразу, что на каждой из тестовых коллекций заметно лидирует выбранный нами алгоритм. Кроме того, временами появляется связка AffinityPropagation + Hierarchy. Именно эту комбинацию мы изначально предполагали использовать в сервисе, однако после проведения большего количества тестов стало заметно, что эта связка в среднем проигрывает иерархической кластеризации как по качеству, так и по скорости работы. Как можно легко увидеть в таблице алгоритмов, мы пробовали не только

использовать разные алгоритмы и эмбеддинги и метрики, но и задавать алгоритмам разные параметры. Например, некоторые алгоритмы выдавали неожиданно неплохие результаты при несоответствующем размеру задачи количеству кластеров. В частности, этим слегка страдает и иерархическая кластеризация, однако для достижения целей это не является проблемой. Очевидно, что задача объединения двух кластеров вручную пользователем намного проще, чем задача разделения кластера на два и более новых, поэтому такое поведение можно назвать особенностью алгоритма.

Текущая версия [веб-сервиса] выполняет кластеризацию документов по выбранному алгоритму, дает возможность просмотреть облако ключевых слов для каждого кластера; соотношения размеров кластеров, количество документов в них, типовой документ для каждого кластера; различные визуализации для оценки вероятности попадания документа в конкретный кластер.



* [s2.voropz.ru:8080]

Заключение

В этой статье был представлен краткий обзор существующих популярных алгоритмов кластеризации, а также оценка качества их работы на разных датасетах. Как результат исследованией было отмечено, что алгоритм

иерархической кластеризации лучше всего подходит для решения как можно большего количества задач с наименьшими затратами времени и сил для пользователя и достижением наилучшего результата, а также представлен прототип веб-сервиса, который позволит в одно нажатие осуществлять кластеризацию любого датасета с достойным качеством.

Благодарности

Хочется отметить тех людей, без которых весь этот проект не был бы возможен. Для начала разработчики сервиса и дизайнеры:
Бояркина Елизавета, Сопильняк Ольга, Воропаев Павел, Врутчель Серафима, Холопов Игорь. Также отметим и менторов и менеджеров проекта: Даниэлян Татьяна, Очеретный Андрей, Володин Денис, Карацапова Ирина, Зацепин Михаил. А также людей, которые оказывали посильную помощь в создании этого проекта и при написании статьи: Селегей Владимир, Маринина Дарья.

Список литературы

1. Isha Sharma TIEIT (Bhopal) Department of CSE, Mahak Motwani TIEIT (Bhopal) Department of CSE *“An Efficient Text Clustering Approach using Biased Affinity Propagation”*, - International Journal of Computer Applications (0975 – 8887) Volume 96– No.1, June 2014
2. Ruksana Akter, Yoojin Chung *“An Evolutionary Approach for Document Clustering”*, - 2013 International Conference on Electronic Engineering and Computer Science
3. Ilya Karpov, Alexandr Goroslavskiy *“Application of BIRCH to text clustering”*
4. Campello R.J.G.B., Moulavi D., Sander J. (2013) *“Density-Based Clustering Based on Hierarchical Density Estimates”*. In: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg

5. Michael Steinbach, George Karypis, Vipin Kumar *"Comparison of Document Clustering Techniques"*, - KDD Workshop on Text Mining
6. R. J. G. B. Campello, D. Moulavi, A. Zimek, J. Sander (2013) *"A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies"*, - Data Mining and Knowledge Discovery
7. Derek Greene, Pʼadraig Cunningham (2006) *"Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering*",* - International Conference on Machine Learning
8. Christopher Issal, Magnus Ebbesson (2010) *"Document Clustering"*, - Master of Science Thesis
9. Marcus Lonnberg, Love Yregard (2013) *"Large scale news article clustering"*, - Master of Science Thesis
10. Donghui Yan, Ling Huang, Michael I. Jordan (2009) *"Fast Approximate Spectral Clustering"*, - Conference on Knowledge Discovery and Data Mining (SIGKDD), Paris
11. Yazhou Ren, Carlotta Domeniconi, Guoji Zhang, Guoxian Yu *"A Weighted Adaptive Mean Shift Clustering Algorithm"*
12. Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu (1996) *"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"*, - KDD-96: Proceedings
13. Andrew Rosenberg and Julia Hirschberg *"V-Measure: A conditional entropy-based external cluster evaluation measure"*, - Department of Computer Science Columbia University, New York
14. W. M. Rand (1971). *"Objective criteria for the evaluation of clustering methods"*. *Journal of the American Statistical Association*. American Statistical Association. **66** (336): 846–850. doi:10.2307/2284239. JSTOR 2284239