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

 # 03, март 2016

УДК 81'322.2

Анализ методов кластеризации текстов применительно к работе с корпусом научных статей

***Столяренко А.В.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»

Научный руководитель: Волкова Л. Л., ассистент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»
irudakov@bmstu.ru

Постановка задачи

Основной целью работы является анализ связей между научными текстами с учётом их распределения по научным областям для сбора статистики о взаимосвязях (например, соотношении между статьями биологов, ссылающихся на математиков и математиков, ссылающихся на биологов). В связи с большим количеством текстов, необходимо разбить тексты в автоматическом режиме по выделенным признакам, в частности, функциональному стилю, научной области. Эту задачу решают алгоритмы кластеризации. Необходимо определить, какие алгоритмы подойдут лучше всего для данной предметной области.

Классификация алгоритмов

На данный момент разработано множество алгоритмов кластеризации, которые делятся на несколько групп [1]:

- *иерархические* (выстраивают дерево кластеров, где корнем является весь набор текстов, а кластеры-листья содержат лишь один текст) и *плоские*;
- *чёткие* (каждый документ или принадлежит, или не принадлежит кластеру) и *нечёткие* (у каждого документа есть степень принадлежности данному кластеру);
- допускающие и не допускающие добавление новых документов после проведения кластеризации.

Классификация алгоритмов согласно [2] приведена на рис. 1.

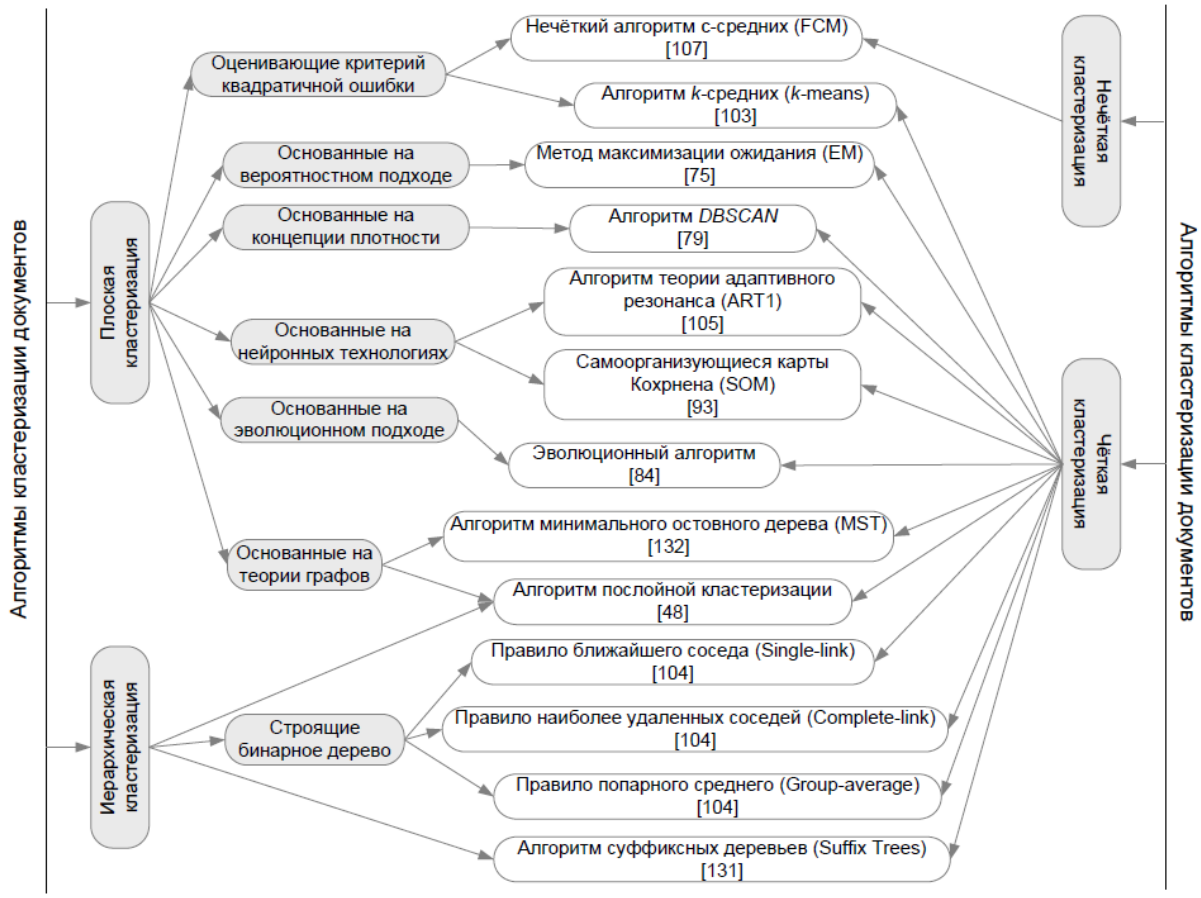

Рис. 1. Классы алгоритмов кластеризации

Определение расстояния между текстами

Определение 1. Термин — словоформа, нормальная форма (представленная начальной формой) или несколько слов, характеризующие документ по смыслу (т. е не являющиеся общеиспользуемыми).

Документы поступают на вход алгоритмов в виде векторов в пространстве терминов $d_i = (d_{i1}, d_{i2}, \dots, d_{i})^T$, где каждое действительное число является координатой вектора, соответствующего термину, и равняется *весу термина* в данном документе.

Вес термина обычно вычисляется по метрике TF-IDF:

$$d_{i,j} = \frac{w_{i,j}}{\|w_L\|}, \qquad (1)$$

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{D}{df_j}\right),$$

где $tf_{i,j}$ — количество раз, которое j-й термин встретился в i-m документе,
df_j — количество документов, в которых встретился j-й термин,
$||w_L||$ — евклидова норма w_L .

При использовании данной метрики общеупотребимые слова получают небольшой вес, что позволяет легко отличить их от слов, представляющих интерес для анализа научных текстов.

Для определения расстояния между документами обычно пользуются формулой:

$$dist(d_i, d_j) = \left(\sum_{k=1}^{n} |d_{i,k} - d_{j,k}|^r\right)^{\frac{1}{r}}, \quad (2)$$

где r задаётся пользователем: $r \in R, r > 0$

При различных значениях r получим:

1. r = 1 — Манхэттенское расстояние
2. r = 2 — Евклидово расстояние
3. $r \to \infty$ — расстояние Чебышёва

Также на практике используется косинусная мера:

$$similarity(d_i, d_j) = \cos(d_i, d_j) = \sum_{k=1}^{n} d_{i,k} \frac{d_{j,k}}{\sqrt{\sum_{k=1}^{n} d_{i,k}^2} \times \sqrt{\sum_{k=1}^{n} d_{j,k}^2}} \quad (3)$$

При совпадающих векторах значение меры близости будет равняться 1, при ортогональных — 0. Значение Евклидова расстояния совпадает с значением косинусной меры при нормализованных векторах.

Рассмотрим алгоритмы с приведением общего принципа их работы.

Алгоритм агломеративной иерархической кластеризации

1. Составить матрицу сходства между кластерами
2. Для всех k от 1 до N-1, где N – количество текстов в корпусе
1. Сохранить информацию о текущем наборе кластеров
2. Выбрать два кластера с максимальным сходством и объединить их
3. Пересчитать матрицу сходства

Алгоритм может использовать несколько разных мер связи для определения меры сходства кластеров:

а) правило одиночной связи — расстоянием между кластерами считается расстояние между самыми близкими их документами;

б) правило полной связи — расстоянием между кластерами считается расстояние между самыми дальними их документами;

в) правило групповой связи — расстоянием между кластерами считается среднее арифметическое расстояния между всеми документами в обоих кластерах, включая документы из одного кластера, но исключая сходство документа с самим собой.

Сложность алгоритма — $O(N^2)$ при использовании правила одиночной связи или правила групповой связи в качестве меры сходства, $O(N^2 logN)$ для правила полной связи.

Алгоритм k-средних

1. Выбрать k случайных текстов в качестве «центров» кластеров, где k – указанное пользователем значение
2. Разместить все остальные тексты в тех кластерах, с центрами которых они наиболее схожи
3. Пока не выполнено пороговое условие:
1. Для каждого кластера вычислить новый центр, которым становится текст с минимальным среднеквадратичным отклонением от остальных
2. Тексты перераспределяются по кластерам по признаку схожести с центром

В качестве порогового условия может использоваться:

1. Достижение максимального количества итераций
2. Отсутствие изменений в центрах кластеров
3. Достигнуто пороговое значение ошибки кластеризации

На практике используется какая-либо комбинация всех трёх признаков.

Оценка сложности: сложность алгоритма линейно зависит от количества документов, кластеров, терминов и итераций. На практике для достижения сложности O(D), где D — количество документов, применяют агломеративный иерархический алгоритм к случайной выборке документов размером $\sqrt{D}$ для получения начальных центров кластеров.

Модификация: нечёткий алгоритм c-средних

Алгоритм допускает принадлежность одного документа нескольким кластерам. Результатом работы алгоритма является матрица степеней принадлежности текста данному кластеру (от 0 до 1). Сам алгоритм идентичен алгоритму k-средних, но для нахождения центра кластера ищется документ с минимальным значением следующей функции:

$$e_m(D,C)=\sum_{k=1}^{|D|}\sum_{j=1}^{|C|}u_{i,j}^m\|d_i-v_j\|^2, \quad (4)$$

где $u_{i,j}$ — степень принадлежности документа кластеру, $0<u_{i,j}<1$

$$\sum_{j=1}^{|C|}u_{i,j}=1,\forall d_i\in D \quad (5)$$

где d_i - документ, D – корпус текстов
m — степень нечёткости, задаваемая пользователем, $1\le m\le\infty$,
v_j — центроид кластера Cj, вычисляющийся по формуле (7)

$$u_{i,j}=\frac{1}{\sum_{k=1}^{|C|}\left(\frac{\|d_i-c_j\|}{\|d_i-c_k\|}\right)^{\frac{2}{m-1}}} \quad (6)$$

$$v_j=\frac{\sum_{i=1}^{|D|}u_{mi,j}\times d_i}{\sum_{i=1}^{|D|}u^m{}_{i,j}} \quad (7)$$

Плотностный алгоритм DBSCAN

Входные данные: параметры Eps и MinPt.

Определение 1: Eps-соседство точки p – это множество всех точек, находящихся на расстоянии не более Eps от p:

$N_{eps}(P)=\{q\in D|dist(p,q)<Eps\}$

Определение 2. Точка p *непосредственно плотно-достижима* из q, если $q\in N_{eps}(P)\wedge|N_{eps}(P)>MinPt|$

Определение 3. Точка p *плотно-достижима* из q, если существует последовательность точек $q,q_1,q_2,\dots,q_n,p$, где q_{i+1} непосредственно плотно-достижима из q_i .

Определение 4. Точка p плотно-связана с точкой q, если существует точка o, из которой плотно-достижима p и q.

Определение 5: Кластером C_j называется непустое подмножество документов, удовлетворяющих следующим условиям:

1. $\forall p,q: p\in C_j\wedge q\ плотно-достижима\ из\ p, то\ q\in C_j$
2. $\forall p,q\in C_j: p\ плотно-связана\ с\ q$

Определение 6: Шум — подмножество документов, которые не принадлежат ни одному кластеру.

Порядок действий:

1. Пометить все точки флагом «не посещён»
2. Для каждой точки с флагом «не посещён»:
    1. Снять флаг «не посещён»
    2. $N_i = \{q \in D \mid dist(d_i, q) \leq Eps\}$
    3. Если $N_i < MinPt$

Отметить точку как «шум»

Иначе

1. Создать новый кластер $j = j + 1$
2. Для всех $d_k \in N_i$
    1. Если d_k помечен как «не посещён»
        1. Снять флаг «не посещён»
        2. $N_{ik} = N_{eps}(d_k)$
        3. Если $|N_{ik}| \geq MinPt$, то $N_i = N_i + N_{ik}$
        4. Если $\neg \exists p : d_k \in C_p, p = \overline{1, |C|}$, то $C_j = C_j + d_k$

Выход: набор кластеров C.

Оценка сложности: $O(n^2)$ в общем случае, при использовании специальной структуры данных (R*-дерево) для хранения информации о точках — $O(n \log n)$.

Применимость к прикладной задаче

Для анализа научных текстов представляют интерес алгоритм агломеративной иерархической кластеризации и нечёткий алгоритм c-средних.

Алгоритм агломеративной иерархической кластеризации позволит проанализировать разбиение научных областей текстов на подобласти (например, из биологии — анатомия, микробиология, генетика).

Нечёткий алгоритм c-средних позволит обнаружить тексты «на стыке» научных областей — например, относящиеся к биоинформатике, если документ будет иметь высокую степень принадлежности и к «биологическому» кластеру, и к кластеру информатики. Так как нечёткий алгоритм c-средних подразумевает и реализацию алгоритма k-средних, стоит воспользоваться и им для сравнения результатов.

Алгоритм DBSCAN кажется менее полезным при разборе корпуса научных статей, так как возникновение кластеров не сферической формы маловероятно, и расстояние между документами в кластере не должно сильно отличаться. Однако, алгоритм может оказаться полезен при тех же условиях, что алгоритм с-средних — при анализе текстов из областей на стыке научных областей. Кроме того, подобные конструкции документов могут образоваться при добавлении нового термина к уже существующим.

Ещё одним полезным алгоритмом в случае анализа научных текстов может оказаться нечёткий плоский алгоритм кластеризации FLAME, который анализирует любые формы кластеров, не только сферические, и также может быть полезен при анализе текстов, находящихся на стыке нескольких научных областей.

Оценка качества кластеризации

Для оценки качества кластеризации (числовых характеристик правильности распределения документов по кластерам) могут использоваться как *внутренние*, так и *внешние* меры[3]. Внутренняя мера основывается на определении кластера — документы внутри кластера должны быть расположены значительно ближе друг к другу, чем документы вне кластера. *Внешним* критерием является степень совпадения распределения части документов по кластерам группой экспертов и алгоритмом кластеризации.

Для оценки *чистоты* кластеризации каждому кластеру присваивается тот класс документов, которых в нём оказалось больше всего (в случае научных текстов — соответствующая предметная область). *Чистота* — это отношение правильно распределённых по кластерам документов к общему количеству документов:

$$purity = \frac{1}{N}\sum_{k} \max_{j} \left| w_k \cap c_j \right|, \qquad (8)$$

где w_k — кластер, а cj — множество документов.

Недостатком чистоты как способа оценки является то, что она стремится к единице при увеличении количества кластеров и при большом количестве кластеров может иметь значения, близкие к единице, при не самом корректном определении.

Индекс Рэнд и *F-мера* используют оценку пар документов. *Истинно-положительным* (ИП) результатом называется попадание двух схожих документов в один кластер, *истинно-отрицательным* (ИО) — попадание двух несхожих документов в разные кластеры, *ложно-положительным* (ЛП) — попадание двух несхожих документов в один кластер, *ложно-отрицательным* (ЛО) — попадание двух схожих документов в разные кластеры.

$$ИР = \frac{ИП + ИО}{ИП + ИО + ЛП + ЛО} \qquad (9)$$

Недостатком индекса Рэнд является присваивание одного и того же веса как ложно-положительным, так и ложно-отрицательным ошибкам, так как в ряде случаев ложно-отрицательные ошибки являются более важными. Эту проблему решает F-мера, основанная на промежуточных показателях P и R.

$$P = \frac{ИП}{ИП + ЛП},$$
$$R = \frac{ИП}{ИП + ЛО}, \qquad (10)$$
$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

где β — коэффициент значимости ложно-отрицательных ошибок.

Данные меры оценки качества позволяют оценить качество разбиения статей по тематикам с минимальными затратами времени. При этом полезными окажутся все три характеристики, так как количество документов будет превосходить количество кластеров как минимум на один порядок и количество кластеров окажется недостаточным для того, чтобы серьёзно повлиять на значения чистоты. Индекс Рэнд и F-мера также дадут данные для оценки ошибок кластеризации.

Заключение

В результате работы были описаны способы определения расстояния между документами и проанализированы алгоритмы кластеризации с точки зрения их полезности при анализе набора научных текстов. Даны рекомендации по выбору необходимых алгоритмов для решения задачи. Также были рассмотрены способы оценки качества кластеризации.

Список литературы

[1]. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ, 2011. 272 с.

[2]. Пескова О.В. Разработка метода автоматического формирования рубрикатора полнотекстовых документов: дис. кандидат технических наук, М., 2008. 151 с.

[3]. Маннинг К., Рагхаван П., Шютце Х.. Введение в информационный поиск: пер. с англ. Москва, 2014. 528 с. [Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.].