УДК 004.912:004.632

В.В. Литвинов, д-р техн. наук
О.П. Мойсеенко, ассистент
Черниговский национальный технологический университет, г. Чернигов, Украина

ФОРМИРОВАНИЕ КЛАСТЕРОВ ПРИ РАБОТЕ С НЕИЕРАРХИЧЕСКИМИ МЕТОДАМИ КЛАСТЕРНОГО АНАЛИЗА

В.В. Литвинов, д-р техн. наук
О.П. Мойсеєнко, асистент
Чернігівський національний технологічний університет, м. Чернігів, Україна

ФОРМУВАННЯ КЛАСТЕРІВ ПРИ РОБОТІ З НЕІЄРАРХІЧНИМИ МЕТОДАМИ КЛАСТЕРНОГО АНАЛІЗУ

Vitaliy Litvinov, Doctor of Technical Sciences
Oleg Moyseenko, assistant
Chernigov National Technological University, Chernigov, Ukraine

THE FORMATION OF CLUSTERS AT WORK WITH NON-HIERARCHICAL METHODS OF CLUSTER ANALYSIS

Рассматриваются принципы объединения похожих текстовых документов в тематические кластеры, механизмы формирования центров кластеров и правила остановки процесса автоматической кластеризации. Подробно описаны основные характеристики кластера и виды кластеризации. Сделан отступ в сторону итеративных методов как таких, что пригодны для работы с большими коллекциями документов и потому, что именно на них основана разрабатываемая система автоматизированной обработки больших объемов динамической текстовой информации. Разрабатываемая система нацелена на выполнение функций поиска, классификации и кластеризации текстовых документов согласно пользовательским запросам.

***Ключевые слова**: текстовая коллекция, центроид, неиерархическая кластеризация, обработка текстовых документов.*

Розглянуто принципи об'єднання схожих текстових документів у тематичні кластери, механізми формування центрів кластерів та правила зупинки процесу автоматичної кластеризації. Детально описані основні характеристики кластера та види кластеризації. Зроблений відступ у бік ітеративних методів як таких, що придатні для роботи з великими колекціями документів і тому, що саме на них ґрунтується розроблювана система автоматизованої обробки великих об'ємів динамічної текстової інформації. Розроблювана система націлена на виконання функцій пошуку, класифікації та кластеризації текстових документів за запитами користувача.

***Ключові слова:** колекція документів, центр кластера, неієрархічна кластеризація, обробка текстових документів.*

Discusses the principles of association or similar text documents into thematic clusters, mechanisms of formation of the cluster centers and the stopping rule of the automatic clustering. Described in detail the main characteristics of the cluster and the kinds of clustering. Indented toward iterative methods such as that are suitable to work with large collections of documents s because it is based on their developed system of automated processing of large volumes of dynamic textual information. The developed system aims to perform search functions, classification and clustering text documents according to user requests. The system itself is described in more detail in other works of the author.

***Key words**: text collection, centroid, non-hierarchical clustering, the processing of text documents.*

Введение. Кластеризация текстовых данных является многоэтапной процедурой, на каждом шаге которой должна решаться отдельная задача выбора наиболее адекватного способа реализации, влияющего на последующие этапы.

Термин кластерный анализ, впервые введенный Робертом Трионом (Robert Tryon) в 1939 году, является обобщенным термином для целого ряда методов, используемых для группировки объектов, событий или индивидов в классы (кластеры) на основе сходства их характерных признаков [1].

Существует большой ряд методов автоматической кластеризации, применимых для работы с текстовыми коллекциями (локальными и динамическими), например: SVM, k-means, Concept Indexing, самоорганизующиеся карты Кохонена (SOM) и др. Всех их объединяет идея выделения малых и максимально «похожих» групп документов или текстовых фрагментов из большого массива входных данных. Отличия существующих методов заключаются в принципах формирования кластеров и механизмах остановки этого процесса.

Свойства кластеров. Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ – вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y:

$$D_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \quad (1)$$

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.

Основной характеристикой кластера является его ценроид (центр кластера), и все существующие алгоритмы кластеризации в процессе образования центроидов тем или иным образом стремятся к стабилизации или полному прекращению изменений их положения в пространстве.

Центр кластера – это вектор, который вычисляется как среднее арифметическое векторов всех документов кластера:

$$\vec{C} = \frac{1}{|S|}\sum_{d \in S}\vec{d}, \quad (2)$$

где S – группа документов, или кластер. Такое определение центроида кластера еще называют центром масс кластера, подмножества документов.

Координаты центра масс в декартовых координатах:

$$c_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \qquad c_y = \frac{\sum_i m_i y_i}{\sum_i m_i}. \quad (3)$$

Если все точки равнозначны, то $m = 1$ для всех точек.

Дисперсия кластера – это мера рассеяния точек в пространстве относительно центра кластера:

$$D_k = \frac{\sum_{i=1}^{I_1}\sum_{j=1}^{n} w_j (x_{ij} - \overline{x}_{kj})^2}{I_k - 1}. \quad (4)$$

Среднеквадратичное отклонение (СКО) объектов относительно центра кластера:

$$S_k = \sqrt{D_k}. \quad (5)$$

Радиус кластера – максимальное расстояние точек от центра кластера, он же радиус наименьшей сферы, содержащей все объекты кластера [2]:

$$R_k = max\sqrt{\sum_{j=1}^{n} w_j (x_{ij} - \overline{x}_{kj})^2}. \quad (6)$$

Принципы иерархической кластеризации. Свыше 100 существующих методов кластерного анализа можно разделить на две группы:

- иерархические;
- неиерархические.

Суть иерархической кластеризации состоит в последовательном объединении (агломеративные методы) меньших кластеров в большие или разделении (дивизимные методы) больших кластеров на меньшие.

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т. е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов [3].

В отличии от иерархических методов кластеризации, итеративные (повторяющиеся) методы до недавнего времени не имели широкого применения. Их актуальность возникла в виду быстрого роста объемов, получаемых и хранимых данных, требующих анализа и обработки. Почти все существующие итеративные методы работают по единому принципу:

1. Исходное разбиение множества данных на указанное число кластеров и вычисление центров тяжести этих кластеров.

2. Помещение точек векторного представления данных в кластер с ближайшим центром тяжести.

3. Вычисление новых центров до тех пор, пока не будут обработаны все входные данные.

4. Повторение двух последних шагов пока не прекратится перестройка кластеров или не будет выполнено правило остановки процесса кластеризации [4].

Согласно первому пункту, пользователь должен указать предполагаемое количество схожих, по его мнению, тематических групп, которые присутствуют в обрабатываемых данных, еще до создания кластеров. То есть для заданного пользователем числа кластеров k создается аналогичное количество k-центров, для каждого из которых сумма расстояний между объектами в n-мерном пространстве и точкой, характеризирующей центр кластера, была минимальной.

$$R = \sum_{i=1}^{k} \sum_{x \in K_i} dist(\mathrm{x}^i, \omega) \to \min, \qquad (7)$$

где ω – объект кластеризации (например, документ); x^i – центр *i*-го кластера; *dist* – дистанция между ними; R – сумма расстояний внутри кластера.

Правила прекращения процесса кластеризации. Касательно правила остановки кластеризации, существует большое множество подходов, среди которых стоит выделить: критерий сферической разделимости, когда сумма радиусов двух кластеров меньше расстояния между их центрами (происходит объединение кластеров); среднюю меру близости, когда среднее значение расстояний объектов кластера от его ядра больше половины расстояния между центрами соседних кластеров (происходит формирование нового ядра и разделение кластера). Как только объединение или разбиение кластеров прекращается, процесс кластеризации останавливается [5].

Следует отметить, что выбор правила прекращения кластеризации существенным образом влияет на показатели оценки работы того или иного метода, а реализовать все возможные правила для достижения наилучших показателей в одной программной системе вряд ли удастся. Следовательно, две программные системы, построенные на базе одного и того же итеративного метода кластеризации, но с разными заложенными правилами остановки процесса обработки, могут формировать одинаковое количество, но

при этом не похожих кластеров с разной степенью распределения классифицируемых объектов (рис. 1 и рис. 2).

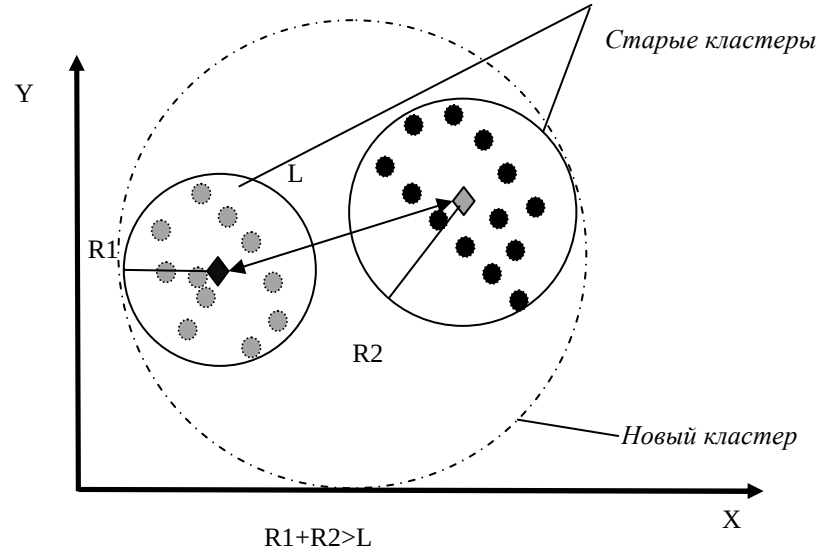

Рис. 1. Формирование нового кластера путем объединения двух существующих кластеров (правило сферической неразделимости)

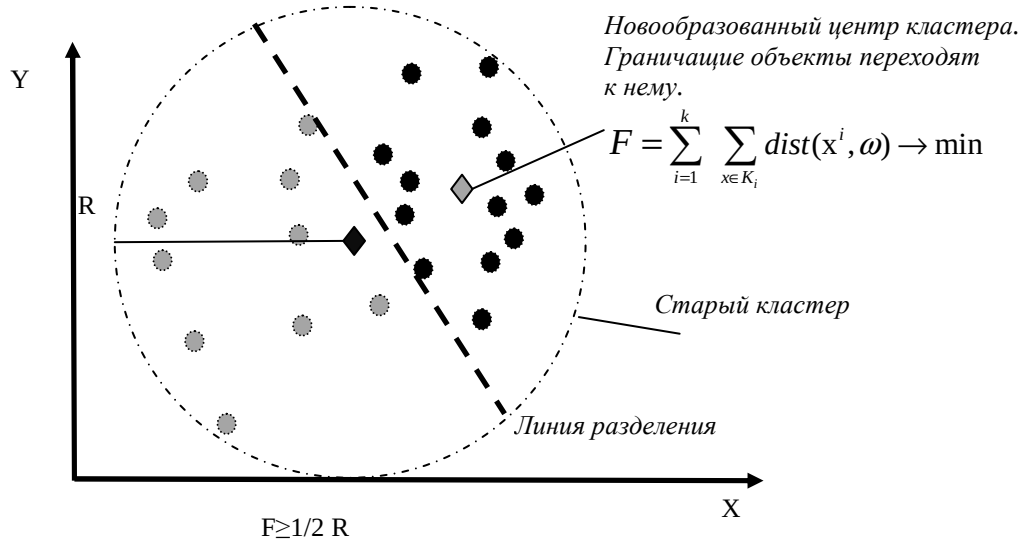

Рис. 2. Образование двух новых кластеров из одного путем его разделения, учитывая среднюю меру близости точек класса от ядра

Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным. Ошибки в результатах кластеризации могут быть устранены экспертом или аналитиком.

Выводы. Кластеризация делается для того, чтобы сократить объем входных данных и в случае возникновения спорных элементов может вызывать потерю соответствующих им документов, что критично при работе с малыми коллекциями. Первостепенный этап выборки начальных центроидов может быть реализован как функцией рандомного распределения, так и по определенному алгоритму, в зависимости от поставленных целей. Конечный результат кластеризации (количество итераций, количество создаваемых кластеров и правильность отнесения к ним обрабатываемых данных), в равной степени зависит от правил начального выбора центроидов и правил остановки процесса кластеризации.

Список использованных источников

1. *Словари* и энциклопедии. Кластерный анализ [Электронный ресурс]. – Режим доступа : http://dic.academic.ru/dic.nsf/enc_psychology/349/Кластерный.

2. *Задачи* кластерного анализа [Электронный ресурс]. – Режим доступа : http://ru.science.wikia.com/wiki/Кластерный_анализ.

3. *Методы* кластерного анализа [Электронный ресурс]. – Режим доступа : http://bug.kpi.ua/stud/work/RGR/DATAMINING/clusteranalisismethods.html.

4. *Факторный*, дискриминантный и кластерный анализ : пер. с англ. / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др. ; под ред. И. С. Енюкова. – М. : Финансы и статистика, 1989. – 215 с.

5. *SVM-Light* Support Vector Machine [Электронный ресурс]. – Режим доступа : http://www.svmlight.joachims.org.

6. *Литвинов В. В.* Автоматизованная система обработки динамических коллекций разноязычных текстовых документов по морскому и речному делу / В. В. Литвинов, О. П. Мойсеенко // Математические машины и системы. – 2014. – № 2. – С. 59–64.