А.А. Дубаков, А.М. Воробьев

Тюменский государственный университет, г. Тюменнь

УДК 004.912

РАЗРАБОТКА АЛГОРИТМА ИЕРАРХИЧЕСКОЙ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ ДЛЯ АНАЛИЗА ТЕКСТОВЫХ ДОКУМЕНТОВ

Аннотация. В данной статье рассматриваются различные подходы к кластеризации данных. Более подробно рассматриваются подходы к интерпретации результатов иерархических алгоритмов кластеризации данных. Рассмотрена векторизация текстовых документов с помощью метрики TF-IDF.

Ключевые слова: интеллектуальный анализ данных, анализ текстовых данных, иерархические алгоритмы кластеризации, дендрограмма.

Введение

Кластеризация является одной из фундаментальных задач в области анализа данных и Data Mining. Список областей применения широк: сегментация изображений, маркетинг, борьба с мошенничеством, прогнозирование, анализ текстов и многие другие. На современном этапе кластеризация часто выступает первым шагом при анализе данных.

Кластеризация данных позволяет:

- упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа;
- производить сжатие данных, так как, если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера;
- выделять нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров;
- производить последующую классификацию других объектов того же типа.

Кластеризация в Data Mining приобретает ценность тогда, когда выступает одним из этапов анализа данных. На практике выгоднее выделить

группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель на всех данных. Например, таким приемом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой их них отдельную стратегию.

Цель исследовательской работы - разработать приложение «Clustering Analysis» для проведения кластерного анализа данных текстового формата и визуализации результатов.

Алгоритмы кластерного анализа

Алгоритмы кластерного анализа предназначены для выделения групп схожих по свойствам объектов в массивах данных [1]. Подразделяются на два типа: иерархические и неиерархические.

При иерархической кластеризации выполняется последовательное объединение меньших кластеров в большие (Агломеративные алгоритмы) или разделение больших кластеров на более мелкие (Дивизивные алгоритмы). К данному типу алгоритмов относят AGNES, CURE, BIRCH, MST и другие (см. Табл.1).

Группа аггломеративных алгоритмов (AGNES[1]) характеризуется последовательным объединением исходных элементов, для которых используются три типа метрик расстояния: между ближайшими элементами кластера (метод одиночной связи), между самыми дальними элементами (метод полной связи), среднее расстояние между объектами двух кластеров (метод средней связи).

Дивизивные алгоритмы характеризуются последовательным разделением исходного кластера на меньшие кластеры. Алгоритм CURE[2], применяемый для кластеризации очень больших наборов числовых данных, выполняет иерархическую кластеризацию с использованием набора определяющих точек. Устойчив к информационным помехам, выделяет кластеры сложной формы,

[1] AGNES – Agglomerative Nesting
[2] CURE – Clustering Using REpresentatives

оптимизирует выходные данные для уменьшения их размера. В локальном алгоритме BIRCH[3] предусмотрен двухэтапный процесс кластеризации и предназначен для работы с очень большим набором числовых данных, при чем низкие требования к объему выделяемой памяти, для работы не требуется постоянного доступа к исходным данным. Алгоритм MST[4] предназначен для кластеризации больших наборов произвольных данных, выделяет кластеры произвольной формы, производит оптимизацию найденного решения и не требует числового представления данных.

Таблица 1. Сравнение иерархических алгоритмов

Алгоритм	Достоинства	Недостатки	Особенности
AGNES	Нет необходимости задавать количество кластеров	Время работы	Позволяют детально проанализировать кластерную структуру данных
CURE	Стабильный результат для данных любой сложности	Необходимость задания пороговых значений	Способен выделять кластеры сложной формы
BIRCH	Не требует больших объемов памяти	Необходимость задания пороговых значений	Учитывает характер распределения данных
MST	Выделяет кластеры произвольной формы	Время работы	Не требует числового представления данных

В неиерархической кластеризации [2] используют алгоритмы такие, как: алгоритм k-средних, PAM, CLOPE и другие (см. Табл.2).

[3] BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies
[4] MST – algorithm based on Minimum Spanning Trees

Алгоритм k-средних строит k кластеров, которые находятся на максимальном расстоянии друг от друга. Выбор числа k может базироваться на результатах предшествующих исследований и подбирается в ходе экспериментов. Алгоритм PAM[5], эффективный на небольших объемах данных, аналогичен алгоритму k-средних, но с перераспределением объектов относительно медианы кластера. Алгоритм CLOPE. предназначенный для кластеризации больших наборов категорийных данных, обладает высокой скоростью работы и производит качественную кластеризацию учитывая особенности исходных данных, при этом требует минимальное число обращений к исходному набору данных.

Таблица 2. Сравнение неиерархических алгоритмов

Алгоритм	Достоинства	Недостатки	Особенности
k-средних	Простота использования, высокая скорость работы	Необходимость задания количества кластеров	Прост в использовании, наилучший результат определяется экспериментальным путем
PAM	Простота использования, высокая скорость работы, устойчивость к помехам в данных	Необходимость задания количества кластеров, низкая скорость работы	Оптимизированный вариант алгоритма k-средних
CLOPE	Высокие масштабируемость и скорость работы	Требователен к памяти	Эффективен на больших объемах данных

[5] PAM – partitioning around medoids

Приложение «Clustering Analysis»

С целью подробного изучения кластерного анализа текстовых данных было разработано пользовательское приложение «Clustering Analysis» на языке программирования C# с использованием технологии WPF.

Программа обладает следующим функционалом:

- векторизация выбранного корпуса документов;
- разбиение векторизованных документов на кластеры согласно заданным параметрам;
- загрузка файла, содержащего результаты кластеризации;
- построение дендрограммы;
- выбор кластера конкретного документа;
- отображение документов (и их векторов), входящих в кластер;
- отображение дополнительной информации.

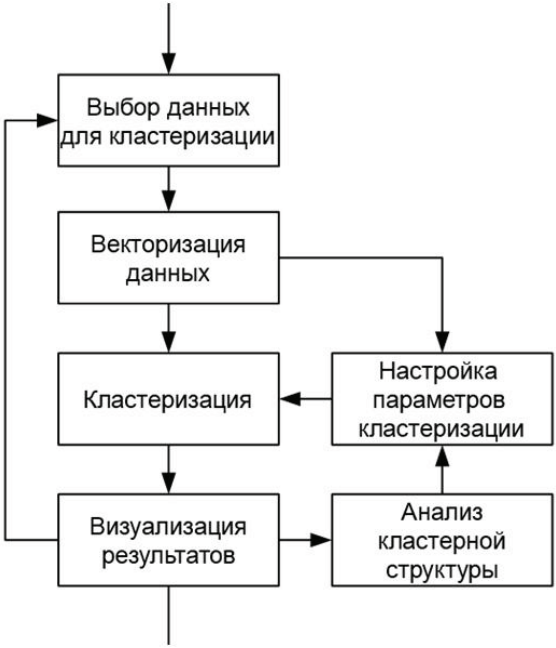

Рис.1. Схема работы приложения

Для проведения кластерного анализа текстовых данных выбран агломеративный иерархический алгоритм [3], так как:

- не требует предварительного задания количества кластеров;
- предоставляет более широкие возможности для экспериментов благодаря большому количеству параметров, которые можно изменять;
- результаты выполнения могут быть представлены в виде дендрограммы, что позволяет детально проанализировать процесс кластеризации.

Работа приложения «Clustering Analysis» включает в себя несколько этапов (см. рис.1).

1 этап – векторизация

Для выбранных текстовых документов происходит векторизация с помощью метрики TF-IDF [4]. В результате каждое слово получает свой вес, который зависит от частоты появления слова в корпусе документов.

Для всего набора документов строится общий словарь, вектора образуются согласно порядку следования записей в словаре. Благодаря этому все документы оказываются в одном информационном пространстве данных, что делает возможным вычисление метрик расстояния между файлами.

2 этап - кластеризация

При помощи агломеративного иерархического алгоритма производится кластеризация, при этом есть возможность задавать условия объединения объектов, изменять метрику для расчета расстояния между объектами в кластере и расстояния между кластерами, устанавливать максимально возможное расстояние между объектами кластера.

3 этап - результаты кластеризации

Результаты кластерного анализа представляются в виде дендрограммы. Данный способ визуализации удобен тем, что имеется возможность рассмотреть, как объединялись кластеры в процессе работы алгоритма. Помимо этого, отображается информация о документах каждого кластера.

Для апробации разработанного алгоритма кластеризации были выбраны 125 текстовых документов, в которых содержалась информация по биологии,

биофизике, экологии, нейросетям, алгоритмическим методам, анализу текстов и другие.

Рассмотрим работу приложения «Clustering Analysis» на примере выбранных 8 файлов (см. Табл. 3).

Таблица 3. Информация о текстовых документах

Обозначение	Файл	Описание	Кол-во слов
D1	Biology	Назначение и основные понятия науки биологии	2100
D2	Biophysics	Основные положения биофизики — одного из ответвлений биологии	1850
D3	Desktop	Системный файл операционной системы Windows	15
D4	Ecology	Общее описание науки экологии	1630
D5	Neuralnet	Нейронные сети и способы их применения	1900
D6	Neuralrecur	Статья об особенностях рекуррентных нейронных сетей	1400
D7	Textmining	Интеллектуальный анализ текстовых данных: определение и основные подходы	900
D8	Tfidf	Определение метрики TF-IDF и способы ее применения	650

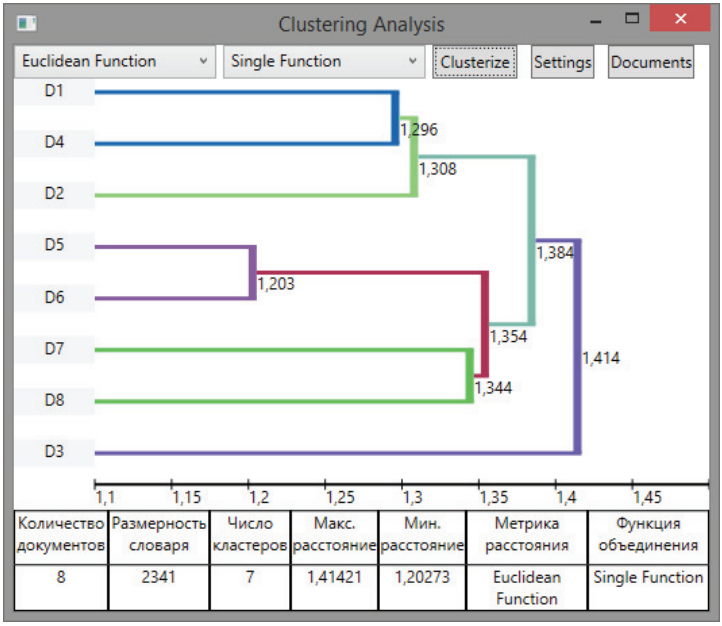

Рис. 2. Построение дендрограммы

В результате на выбранных тестовых данных проведен анализ, построена дендрограмма (см. рис.2), для набора корпуса документов были определены кластеры, выделены 3 основных кластера, посчитаны метрики TF-IDF.

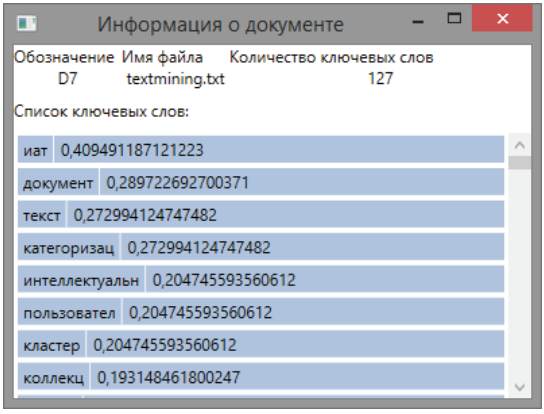

Рис. 3. Просмотр информации о документе D3

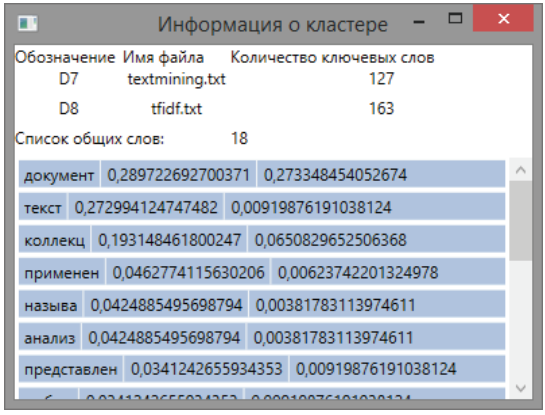

Рис. 4. Просмотр информации о кластере с документами D7, D8

Используя дендрограмму, список ключевых слов каждого документа (см. рис.3) и функцию просмотра параметров кластера (см. рис.4), для данного корпуса выбранных документов можно сделать следующие выводы:

1) группу документов D5- D6 и D7 - D8, которые имеют соответственно 187 и 85 общих ключевых слов, можно объединить в один кластер, так как эти тексты относятся к одной тематике - сфера IT технологий;

2) группа документов D1 - D4 (155 общих ключевых слов) и документ D2 относятся к общему кластеру, поскольку описывают науки, изучающие живую и неживую природу;

3) документ D3 является системным файлом и почти не имеет общих ключевых слов с остальными документами, поэтому был добавлен в дендрограмму на последнем этапе работы алгоритма, в следствии чего D3 вынесен в отдельный кластер.

Таким образом, разработанное приложение «Clustering Analysis» предоставляет возможность строить кластеры, выводить полученную кластеризацию в виде дендрограммы, что позволяет определить структуру корпуса документов.

В дальнейшем планируется использовать результаты кластерного анализа для поиска похожих документов по выбранной тематике, выводить дополнительную информацию о причинах агломерации кластеров.

СПИСОК ЛИТЕРАТУРЫ

1. Нейский, И.М. Классификация и сравнение методов кластеризации / И.М. Нейский // URL: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf «Clustering Analysis»Егоров, А.В. Особенности методов кластеризации данных / А.В. Егоров, Н.И. Куприянова // Известия Южного федерального университета. Технические науки – 2011 – №4 – С. 14-19.
2. Ломакина, Л.С. Иерархическая кластеризация текстовых документов / Л.С. Ломакина, В.Б. Родионов, А.С. Суркова // Системы управления и информационные технологии. – 2012 – №2 – С. 39-44.
3. Романенко, А.А. Категоризация текстов на основе монотонного классификатора ближайшего соседа / А.А. Романенко – Математические и информационные технологии – 2011.