Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Анализ методов кластеризации данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Основина Андрея Николаевича

Научный руководитель
д.э.н., проф. _____ Л.В. Кальянов
подпись, дата

Зав. кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
подпись, дата

Саратов 2016

ВВЕДЕНИЕ

В современном мире информация имеет огромную ценность, поэтому важно уметь грамотно ее структурировать, обобщать и представлять для последующего анализа. С ростом объемов информации становится важным поиск оптимальных методов ее обработки.

Кластеризация – объединение в группы схожих объектов – является одной из фундаментальных задач в области анализа данных и Data Mining. Список прикладных областей, где она применяется, широк: сегментация изображений, маркетинг, борьба с мошенничеством, прогнозирование, анализ текстов и многие другие.

Целью данный работы является построение модели кластеризации текстовых документов и на основе построенной модели проведение анализа применения к данной задачи иерархических методов, в частности агломеративного метода и неиерархических методов, в частности итеративного метода k- Means, кластеризации данных.

Актуальность выбранной темы обусловлена широким применением кластеризации, в различных областях. Она часто выступает первым шагом при анализе данных: выделение групп похожих объектов помогает понять структуру данных и использовать свой подход к обработке каждой группы. Кластеризация позволяет сократить объем хранимых данных (оставив по одному наиболее типичному представителю каждого кластера) и обнаружить нетипичные объекты, которые не удается причислить ни к одному из кластеров (задача обнаружения новизны). На сегодняшний момент число методов разбиения групп объектов на кластеры довольно велико – несколько десятков алгоритмов и еще больше их модификаций.

Структура выпускной квалификационной работы состоит из введения, трех глав, заключения, списка использованных источников и приложения.

В первой главе приведены определения основных понятий связанных с задачей кластеризации данных в Data Mining, рассмотрена классификация методов кластеризации данных, описаны принципы работы метода кластеризации k-Means и алгоритм Ланса-Уильямса.

Во второй главе проведен анализ информационных технологий кластеризации KNIME и RapidMiner.

В третьей главе описана реализации модели кластеризации текстовых документов и проведен анализ применения к данной задачи метода k-Means и агломеративного метода кластеризации данных.

1 Кластерный анализ в Data Mining. Термин “кластерный анализ” впервые был предложен профессором Калифорнийского университета Робертом Трионом в 1939 году [4]. Кластерный анализ включает в себя целый ряд алгоритмов и методов для группировки элементов некоторого множества по сравнительно однородным группам. Кластеризация относится к методам Data Mining с обучением без учителя.

Дадим формальное определение термина кластерный анализ. Объект - элементарная группа данных, с которой оперируют алгоритмы кластеризации. Каждому объекту ставится в соответствие вектор характеристик $x = (x_1, ..., x_k)$. Компоненты x_j являются отдельными характеристиками объекта. Количество характеристик k определяет размерность пространства характеристик. Через X будем обозначать пространство объектов с заданными характеристиками.

На заданном пространстве объектов X должна быть определена функция расстояния между объектами $\rho(x, x')$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.

Обозначим через $Y = (y_1, \ldots, y_n)$ множество кластеров. Алгоритм кластеризации - это функция $a: X \to Y$, которая каждому объекту $x \in X$ ставит в соответствие метку кластера $y \in Y$. Зачастую число кластеров неизвестно заранее, поэтому по мимо задачи разбиения объектов по кластерам необходимо решить задачу определения оптимального числа кластеров, с точки зрения того или иного критерия качества кластеризации.

Методы кластерного анализа делятся на иерархические и неиерархические методы. В результате работы иерархического метода создается иерархия, которую можно визуализировать в виде дендрограммы. Дендрограмма - древовидная структура, которая описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения или разделения, в зависимости от типа иерархического метода, кластеров. Иерархические методы бывают двух типов: агломеративные и дивизимные методы.

Агломеративные методы выполняют последовательное объединение исходных объектов и соответствующие уменьшение числа кластеров. В начале работы алгоритма все объекты $x_i \in X$ принадлежат к отдельным кластерам $y_j \in Y$. На каждом шаге работы алгоритма два наиболее похожих кластера, для которых функция расстояния $p(x_i, x_j)$ между произвольными объектами этих кластеров $x_i \in y_i$ и $x_i \in y_i$ имеет минимальное значение объединяются в кластер $y_{ij} \in Y$. Алгоритм выполняется до тех пор, пока все кластеры не будут объединены в один общий кластер.

Дивизимные методы выполняют последовательное разделение исходного кластера, состоящего из всех объектов, и соответствующие увеличением числа кластеров. В начале работы алгоритма все объекты $x_i \in X$ принадлежат одному кластеру $y \in Y$. На каждом шаге работы алгоритма происходит последовательное разделение одного из кластеров $y_j \in Y$ на пару кластеров y_{j1} и y_{j2} таким образом, что бы расстояние $p(x_i, x_j)$ между всеми объектами этих кластеров $x_i \in y_{j1}$ и $x_j \in y_{j2}$ было максимально. Алгоритм выполняется до тех пор, пока все объекты не будут принадлежать к отдельным кластерам.

Неиерархические методы кластерного анализа бывают одного вида, итеративные методы. Процесс кластеризации в итеративных методах начинается с задания некоторых начальных условий $P = \{p_1, \ldots, p_n\}$, которые определяют количество образуемых кластеров, порог завершения процесса классификации и т. д. Изменение начальных условий существенно меняет и результаты кластеризации, поэтому применение этих методов требует предварительного изучения генеральной совокупности, в частности, с помощью иерархических методов кластерного анализа. Процесс кластеризации выполняется пока не будут выполнены эти условия.

Метод k-Means (k-средних) - это метод кластеризации стремящийся минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Он применяется к объектам, которые представляются точками в n - мерном векторном пространстве $X = R^n$. Он разбивает множество X на k кластеров $y \in Y$. При этом каждый объект $x \in X$ относится только к одному кластеру $y \in Y$, то есть данный метод по способу анализа данных является чет-

ким. В качестве функции расстояния p(x, y) используется Евклидово расстояние.

Кластеризации в иерархических агломеративных методах начинается с отнесения каждого объекта в свой отдельный кластер. Затем наиболее близкие в заданном смысле кластеры попарно объединяются в более крупные, образуя вложенные друг в друга кластеры. Существует множество методов объединения похожих кластеров в один, но их отличие заключается в способе определения расстояния между кластерами или между кластерами и объектами. И в 1967 Ланс и Уильямс предложили формулу, которая обобщала ряд существующих способов объединения кластеров.

Формула предоставляет ответ на вопрос как определить расстояние R(W, S) между кластерами $W = V \cup U$ и S , зная расстояния R(U, S), R(V, S), R(U, V).

2 Сравнительный анализ информационных технологий кластеризации. RapidMiner - это программная платформа, разработанная одноименной компанией, которая обеспечивает интегрированную среду для машинного обучения, интеллектуальный анализ данных, анализ текста, прогнозного анализа и бизнес-аналитики [6]. Для решения практической задачи в RapidMiner строятся процессы. Процесс - это совокупность операторов соединенных между собой в заданном порядке для выполнения требуемой задачи анализа или обработки данных. Логическими единицами процесса является операторы. Оператор производит какие-либо манипуляции с данными.

KNIME - это модульная платформа с открытым исходным кодом, предназначенная для анализа данных и составления отчетов. KNIME интегрирует различные компоненты для машинного обучения и интеллектуального анализа данных с помощью модульной концепции конвейера данных.

Среда KNIME позволяет пользователю визуально создавать поток данных, выборочно выполнять анализ шагов, а затем исследовать результаты посредством интерактивного просмотра данных и моделей [5].

Потоки данных могут быть запущенны через интерактивный интерфейс пользователя и обрабатываться в пакетном режиме, упрощая интеграцию про-

цесса анализа данных для менеджмента и выполнения на периодической основе.

3 Практическая часть. Кластеризация текстовых документов может применяться для выделения из текстовой коллекции групп текстов одинаковой тематики. Эта задача относится к задачам поиска скрытой неструктурированной информации. Из-за больших размеров текстовых коллекций и из-за субъективности восприятия читателя темы текста оценить качество кластеризации сложно. Поэтому пока нет общепринятого функционала качества кластеризации текстовых коллекций и алгоритма, являющегося абсолютно лучшим.

Пусть есть некоторое множество слов английского языка, каждое из которых хотя бы раз встретилось в одном из документов текстовой коллекции. Назовем это множество словарем. В данной работе под документом будем понимать неупорядоченное множество слов из словаря. Слова в документе могут повторяться. Тогда каждому документу можно поставить в соответствие вектор, содержащий информацию о словарном составе документа. Размерность этого вектора равна количеству слов в словаре. Тогда расстояние между документами можно ввести как расстояние между векторами, соответствующими этим документам.

Построенная в платформе KNIME модель осуществляющая кластеризацию документов показана на рисунке 1.

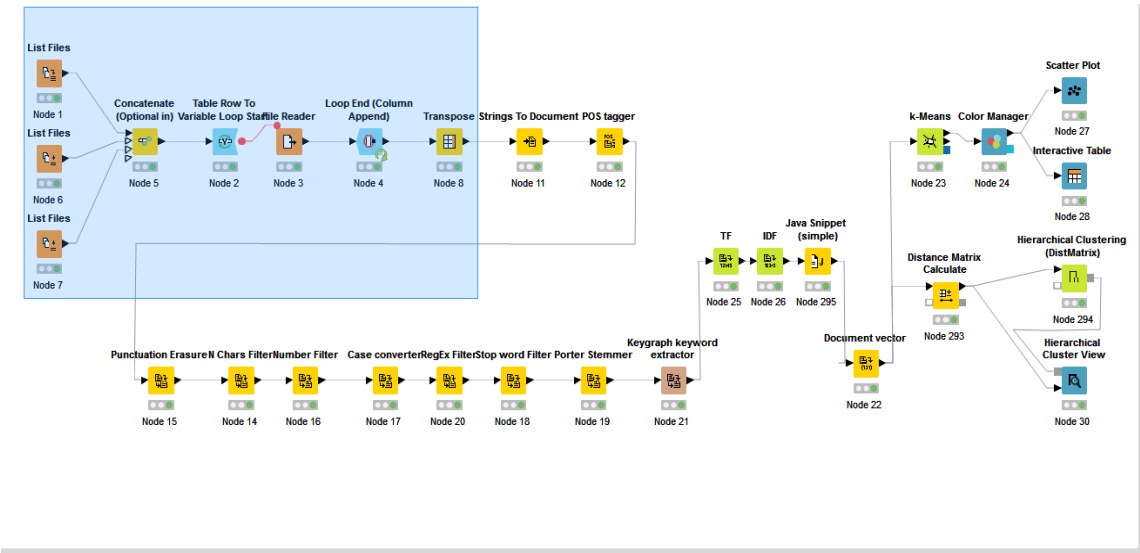

Рисунок 1 — Модель кластеризации текстовых документов в KNIME

Для иллюстрации работы модели был взят архив электронных писем трех разных тематик: спорт (хоккей), религия (христианство) и криптография. Из каждой категории писем было выбрано по 50 произвольных писем, таким образом полученная текстовая коллекция содержит 150 текстовых документов.

Метод k-Means разбил исходное множество документов практически на верное количество кластеров, была допущена всего одна ошибка. Дендрограмма построенная агломеративным методом кластеризации не подходит для решения данной задачи. Так как из полученной дендрограммы можно выделить взаимосвязь отдельных документов, но общую структуру множества практически невозможно определить.

Исходя из полученных результатов можно сделать выводы о том, что для решения задачи кластеризации текстовых документов, где число кластеров значительно меньше числа исследуемых объектов лучше использовать итеративные методы кластеризации, так как результат их работы лучше отображает группы текстов одинаковой тематики.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были получены следующие результаты:

1. Приведены определения основных понятий связанных с задачей кластеризации данных в Data Mining.
2. Рассмотрена классификация методов кластеризации данных.
3. Описаны принципы работы метода кластеризации k-Means и алгоритм Ланса-Уильямса.
4. Проведен анализ информационных технологий кластеризации KNIME и RapidMiner.
5. Построена модель, осуществляющая кластеризацию текстовых документов в среде KNIME.
6. На основе построенной модели проведен эксперимент, результаты которого послужили обоснованием для выбора итеративного метода кластеризации данных для решения задачи кластеризации текстовых документов.

Полученную модель кластеризации текстовых документов можно применять для любого рода документов в задачах, где требуется разбить исходное множество текстов по их содержимому.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 J.J. Shen, Using Cluster Analysis, Cluster Validation, and Consensus Clustering to Identify Subtypes of Pervasive Developmental Disorders - Queen's University Kingston, Ontario, Canada , 2007

2 А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод: Методы и модели анализа данных OLAP и Data Mining – СПб.: БХВ-Петербург, 2004

3 А.А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP - СПб.: БХВ-Петербург, 2007

4 Кластерный анализ. [Электронный ресурс]: URL: https://ru.wikipedia.org/wiki/Кластерный_анализ (дата обращения 18.05.16)

5 Официальный сайт платформы KNIME. [Электронный ресурс]: URL: https://www.knime.org/ (дата обращения 02.05.16)

6 Официальный сайт платформы RapidMiner. [Электронный ресурс]: URL : https://rapidminer.com/ (дата обращения 10.05.16)

7 К.В. Воронцов, Лекции по алгоритмам кластеризации и многомерного шкалирования, 2004