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

Кафедра информатики и программирования

# КЛАСТЕРИЗАЦИЯ СОЦИАЛЬНЫХ СЕТЕЙ: РЕАЛИЗАЦИЯ, ОЦЕНКА И СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем (профиль Параллельное программирование)
факультета компьютерных наук и информационных технологий
Еремушкина Даниила Константиновича

Научный руководитель
к.ф.-м.н. _____ М.В. Огнёва

Зав. кафедрой
к.ф.-м.н. _____ М.В. Огнёва

Саратов 2018

# ВВЕДЕНИЕ

За последние несколько веков, общество претерпело кардинальные изменения. На сегодняшний момент интернет проник во все сферы жизни человека, что в совокупности с развитием технологий привело к тому, что в интернете каждый день появляется миллионы гигабайт информации. Из-за таких тенденций остро встает вопрос обработки и анализа больших данных. *Большие данные* (англ. *Big Data*) – обозначение структурированных или неструктурированных данных огромных объемов [1].

Для решения возникших проблем естественным образом появилась *наука о данных. Наука о данных* – раздел информатики, изучающий проблемы обработки, анализа и представления данных в цифровом виде. Она включает в себя методы обработки данных больших объемов, статистические методы и методы интеллектуальной обработки данных с возможностью приложения искусственного интеллекта для работы с ними [2].

Одним из методов обработки *больших данных* является кластерный анализ, представляющий собой процедуру, которая упорядочивает данные из предоставленной выборки в сравнительно однородные группы [3].

Одним из важных этапов при анализе данных является извлечение информации – это задача автоматического сбора (построения) и структурирования данных из неструктурированных или мало структурированных машиночитаемых документов или цифровых источников информации.

В рамках современного мира главные источники информации находятся в глобальной сети Интернет, например, информация о поисковых запросах в популярных поисковых сервисах, информация о покупках в интернет-магазинах или открытые данные из социальных сетей. Социальная сеть в Интернете представляет собой веб-сайт с возможностью регистрации пользователей и

размещения информации о себе, а также с возможностью коммуникации между пользователями, посредством установления социальных связей.

За последние несколько лет социальные сети получили большую популярность, например, около 60% Российских пользователей Интернета, состоят в какой-либо социальной сети [4]. Поэтому на их базе представляется возможность проводить множество интересных исследований с помощью *науки о данных*.

**Цель бакалаврской работы** – реализация и сравнительный анализ алгоритмов кластеризации k-means, k-means++, FOREL, SCAN, DBSCAN на базе графов построенных из данных о пользователях социальных сетей «ВКонтакте» и «Twitter».

Поставленная цель определила следующие **задачи**:

- рассмотреть основные понятия, связанные задачей кластеризации;
- разобрать алгоритмы: k-means, k-means++, FOREL, SCAN, DBSCAN;
- реализовать вышеуказанные алгоритмы кластеризации;
- подобрать и произвести разбор метрик качества;
- реализовать алгоритмы оценки качества результатов кластеризации;
- сгенерировать искусственный граф, с заданным количеством кластеров;
- построить граф на основе данных из социальной сети «Twitter»;
- собрать данные из социальной сети «ВКонтакте»;
- на основе полученных данных из социальной сети «ВКонтакте», построить граф;
- произвести проверку результатов работы алгоритмов кластеризации на искусственно сгенерированном графе и на графах сформированных из данных социальных сетей «ВКонтакте» и «Twitter»;

- произвести сравнительный анализ результатов работы различных алгоритмов, с использованием метрик оценки качества работы алгоритмов.

**Практическая значимость бакалаврской работы.** В ходе выполнения практической части бакалаврской работы были реализованы следующие алгоритмы кластеризации: k-means, k-means++, FOREL, SCAN, DBSCAN, а также две метрики качества: метрика эффективности кластеризации и метрика качества кластеров. Для тестирования алгоритмов созданы следующие графы:

- Искусственно сгенерированный граф с заданным количеством кластеров;
- Граф на основе социальной сети «Twitter», для создания которого был взят готовый набор данных из открытых источников [5];
- Граф на основе данных из социальной сети «ВКонтакте», данные для которого были предварительно собраны.

После создания графов, был проведен сравнительный анализ работы вышеупомянутых алгоритмов кластеризации, на созданных графах, при помощи реализованных метрик качества.

**Структура и объем работы.** Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и пяти приложений. Общий объем работы – 54 страницы, из них 41 страница – основное содержание, включая 39 формул и 12 таблиц, список использованных источников информации – 20 наименований.

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Первый раздел «Алгоритмы кластеризации»** посвящён:

- описанию задач и целей кластеризации,
- теоретическому описанию алгоритмов k-means, k-means++, FOREL, SCAN, DBSCAN,
- описанию метрик качества кластеризации.

Подраздел «Задачи кластеризации» включает в себя описание задач кластеризации в интуитивном и формальном виде.

Основная задача кластеризации звучит следующем образом: необходимо разбить весь массив данных, на котором уже задана «степень похожести», метрика, на отдельные кластеры, в которых будут сгруппированы «похожие» элементы.

Группировка информации в кластеры, дает возможность работать с типичными представителями группы, что сокращает объемы информации и соответственно время работы с ней.

В подразделе «Цели кластеризации» сформулированы основные цели кластеризации, которые звучат следующем образом:

- Упрощение обработки данных. Понять структуру объектов $X^l$, разбив на кластеры и работая с каждым кластером в отдельности, тем самым упростив обработку данных.
- Сокращение объемов хранимой информации. После разбиения выборки $X^l$ на кластеры оставить только по одному «типичному» представителю каждого кластера, тем самым сократив количество хранимой информации.
- Обнаружение не типичности или новизны. Выделение не типичных объектов, которые не подходят ни к одному кластеру.

Подраздел «Графовые алгоритмы кластеризации» включает в себя основные понятия [6] и теоретическое описание алгоритмов кластеризации, которые требуют для работы представление выборки данных в виде неориентированного взвешенного графа (за исключением алгоритма SCAN).

Алгоритмы k-means и k-means++ разбивают граф на заданное число кластеров. Исходя из этого количества, алгоритмы в начале своей работы выбирают необходимое число центров для будущих кластеров, и на последующих итерациях этих алгоритмов все вершины графа распределяются в кластеры, в соответствии с тем, к какому центру кластера ближе вершина.

Алгоритм FOREL разбивает граф на заранее неизвестное количество кластеров. В основе этого алгоритма лежит следующая базовая процедура. Пусть задана некоторая точка $x_0 \in X^l$ и параметр $R$. Выделяются все точки выборки $x_i \in X_l$, попадающие внутрь сферы $\rho(x_i, x_0) < R$, и точка $x_0$ переносится в центр графа построенного из выделенных точек. Эта процедура повторяется до тех пор, пока состав выделенных точек, а значит и положение центра, не перестанет меняться [7].

Алгоритм SCAN предназначен для анализа структуры графа. Преимущество данного алгоритма заключается в том что, он сам определяет количество кластеров, а так же он умеет определять вершины – перемычки. *Перемычка* это вершина, которая является переходной от одного кластера к другому. Кроме того, алгоритм SCAN выделяет внешние вершины, то есть такие вершины, которые находиться на краях графа и не являются частью, какого либо кластера [8].

Алгоритм DBSCAN представляет собой измененный алгоритм SCAN, основное отличие данного алгоритм в том, что он строит кластеры исходя не только из структуры графа, но также исходя из расстояний между вершинами.

В разделе «Оценка качества работы алгоритмов кластеризации» описываются две метрики оценки качества работы алгоритмов кластеризации.

Метрика эффективности, применима, когда известны истинные составы кластеров. Такой способ определения качества показывают насколько кластеры определенные алгоритмом, близки к истинным кластерам.

Метрика качества кластеров, подходит для оценки качества непосредственно кластеров предсказанных алгоритмом, без знания истинного состава кластеров в графе.

Второй раздел «Сравнительный анализ алгоритмов кластеризации» посвящён:

- реализации вышеописанных алгоритмов кластеризации и метрик оценки качества,
- генерации искусственного графа с явно выделенными кластерами, для проверки работы алгоритмов,
- построению графа на основе данных из социальной сети «Twitter», которые были взяты из открытого ресурса [5],
- сбору доступной информации о пользователях социальной сети «ВКонтакте» [9],
- построению графа на основе информации о пользователях социальной сети «ВКонтакте»,
- анализу работы алгоритмов кластеризации на выше представленных графах
- проведению сравнительного анализа вышеописанных алгоритмов кластеризации.

# ЗАКЛЮЧЕНИЕ

В ходе данной работы были рассмотрены и реализованы следующие алгоритмы кластеризации: k-means, k-means++, FOREL, SCAN, DBSCAN. Помимо алгоритмов были реализованы два алгоритма оценки качества результатов работы кластеризации.

Для проведения сравнительного анализа алгоритмов кластеризации, при помощи рассмотренных метрик качества, были построены три графа. Один из них был искусственно сгенерирован таким образом, что бы в нем были четко выделены четыре кластера, а два других графа были построены на основе данных из социальных сетей, для чего был проведен сбор информации из социальных сетей «Twitter» и «ВКонтакте».

На каждом наборе данных были алгоритмы, которые показали себя лучше чем остальные. Для искусственно сгенерированного графа, хорошо показал себя алгоритм *forel,* так как он в среднем отрабатывал быстрее остальных алгоритмов. Для графа, основанного на данных из социальной сети «Twitter», фаворитом оказался алгоритм SCAN, так как ему хорошо удается анализировать структуры и связи в граф. Для графа, полученного из данных социальной сети «ВКонтакте», оказался алгоритм DBSCAN так как он умеет работать и со структурой вершин, так и с метрикой графа.

Протестировав работу всех представленных алгоритмов кластеризации и проведя их сравнительный анализ, можно сделать вывод, что каждый алгоритм заслуживает внимания и  для каждого алгоритма есть свой класс задач, на которых использование этого алгоритма будет оправданно.

## Список литературы


1. Preimesberger. Hadoop, Yahoo, 'Big Data' Brighten BI Future. EWeek, 2011.
2. Naur P. Concise Survey of Computer Methods. Studentlitteratur AB, 1974.
3. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999.
4. Дужникова А.С. СОЦИАЛЬНЫЕ СЕТИ: СОВРЕМЕННЫЕ ТЕНДЕНЦИИ И ТИПЫ ПОЛЬЗОВАНИЯ // Мониторинг общественного мнения. сентябь - октябрь 2010. No. 5 (99).
5. Higgs Twitter Dataset [Электронный ресурс] // snap: [сайт]. [2006]. URL: http://snap.stanford.edu/data/higgs-twitter.html (дата обращения: 15.Май.2018).
6. EM алгоритм (пример) [Электронный ресурс] // Профессиональный информационно-аналитический ресурс, посвященный машинному обучению: [сайт]. URL: http://www.machinelearning.ru/wiki/index.php?title=EM (дата обращения: 30.апреля.2018).
7. Воронцов К.В. Лекции по алгоритмам кластеризации. 2007.
8. Xiaowei X., Yuruk N., Feng. SCAN: A Structural Clustering Algorithm for Networks.
9. Описание методов API [Электронный ресурс] // ВКонтакте для разработсиков: [сайт]. URL: https://vk.com/dev/methods (дата обращения: 10.май.2018).
10. Arthur D., Vassilvitskii S. How Slow is the k-means Method? 2006.
11. Arthur D., Vassilvitskii S. k-means++: the advantages of careful seeding. 2007.
12. Бураго Д.Ю. Курс метрической геометрии. Рхд, 2004.
13. Python 3.6.5 documentation [Электронный ресурс] // python: [сайт]. [2001]. URL: https://docs.python.org/3/index.html (дата обращения: 14.май.2018).

14. NetworkX documentation [Электронный ресурс] // NetworkX: [сайт]. URL: https://networkx.github.io/documentation/networkx-1.9.1/ (дата обращения: 20.май.2018).
15. Welcome to the MongoDB Docs [Электронный ресурс] // mongodb: [сайт]. URL: https://docs.mongodb.com/ (дата обращения: 13.май.2018).
16. Стилин И., Панов М. Обзор и эксперементальное сравнение алгоритмов кластеризации графов. Долгопрудный: Институт Проблем Передачи Информации. Россия pp.
17. Yang J., Leskovec J. Defining and Evaluating Network Communities based on Ground-truth. 2012.
18. Интеллектуальный анализ данных социальных сетей: эмоциональная направленность и географическая привязка [Электронный ресурс] // Университет ИТМО: [сайт]. URL: http://escience.ifmo.ru/research/view/26 (дата обращения: 28.март.2018).
19. Stanford Large Network Dataset Collection [Электронный ресурс] // SNAP: [сайт]. URL: http://snap.stanford.edu/data/index.html#communities (дата обращения: 29.март.2018).
20. Social Network Analysis: Spark GraphX [Электронный ресурс] // habr: [сайт]. (дата обращения: 24.апрель.2018).