

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тольяттинский государственный университет»

_____ Институт математики, физики и информационных технологий _____

(наименование института полностью)

_____ Кафедра «Прикладная математика и информатика» _____

(наименование кафедры)

_____ 02.03.03 Математическое обеспечение и администрирование
информационных систем _____

(код и наименование направления подготовки, специальности)

_____ Технология программирования _____

(направленность (профиль)/специализация)

БАКАЛАВРСКАЯ РАБОТА

_____ на тему «Сетевое программное обеспечение для выполнения
кластерного анализа» _____

Студент

_____ К.С. Малышев _____

(И.О. Фамилия)

_____ (личная подпись)

Руководитель

_____ В.С. Климов _____

(И.О. Фамилия)

_____ (личная подпись)

Допустить к защите

Заведующий кафедрой к.тех.н, доцент, А.В. Очеповский _____

(ученая степень, звание, И.О. Фамилия)

_____ (личная подпись)

« _____ » _____ 2018 г.

АННОТАЦИЯ

СЕТЕВОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ, КЛАСТЕРНЫЙ АНАЛИЗ,
PHP, MYSQL

Тема: Сетевое программное обеспечение для выполнения кластерного анализа

Целью ВКР является разработка сетевого программного обеспечения для выполнения кластерного анализа.

Объектом исследования является кластерный анализ.

Предмет исследования – сетевое программное обеспечение для выполнения кластерного анализа.

Методы исследования: метод кластерного анализа, Web-технологии, объектно-ориентированное программирование.

Произведен обзор методов и выбран алгоритм кластерного анализа для реализации в сетевом режиме.

На основе объектно-ориентированного подхода разработана логическая модель и модель данных сетевой информационной системы (ИС). В качестве среды разработки выбраны PHP и MySQL.

Разработано программное обеспечение сетевой ИС и даны рекомендации по его аппаратно-программному обеспечению.

Апробация программного обеспечения подтвердило его работоспособность.

Работа включает 46 страниц, 14 рисунков, 6 таблиц, 22 источника.

ABSTRACT

NETWORK, SOFTWARE, CLUSTER ANALYSIS, PHP, MYSQL

The title of the graduation work is “ Network Software for Cluster Analysis”.

The aim of the graduation work is to develop of network software for cluster analysis.

The object of the graduation work is cluster analysis.

The subject of the graduation work is network software for cluster analysis.

Research methods: cluster analysis, Web-technology, object-oriented programming.

The review of clustering methods is conducted. The algorithm for cluster analyzing is chosen.

On the basis of the object-oriented approach a logical model and data model of the network information system (IS) are developed. As a development tools PHP and MySQL are chosen.

The network IS software is developed. The recommendations on its hardware and software requirements are given.

Testing of the software has confirmed its working capacity.

The graduation work consists of an explanatory note on 46 pages including 14 figures, 6 tables, the list of 22 references.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	6
Глава 1 АНАЛИЗ МЕТОДОВ И ВЫБОР АЛГОРИТМА КЛАСТЕРНОГО АНАЛИЗА	8
1.1 Методы кластерного анализа.....	8
1.1.1 Иерархические методы кластерного анализа.....	9
1.1.2 Неиерархические методы кластерного анализа	10
1.2 Описание алгоритма кластерного анализа.....	11
1.3 Постановка задачи на разработку сетевого программного обеспечения для выполнения кластерного анализа	13
1.4 Обзор и анализ аналогов программного обеспечения для выполнения кластерного анализа	14
Глава 2 РАЗРАБОТКА И РЕАЛИЗАЦИЯ ПРОЕКТНЫХ РЕШЕНИЙ СЕТЕВОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ВЫПОЛНЕНИЯ КЛАСТЕРНОГО АНАЛИЗА	20
2.1 Разработка программной архитектуры сетевой информационной системы для выполнения кластерного анализа.....	20
2.2 Логическое моделирование сетевой информационной системы для выполнения кластерного анализа	22
2.2.1 Разработка диаграммы вариантов использования процесса выполнения кластерного анализа.....	23
2.2.2 Разработка диаграммы классов сетевой информационной системы для выполнения кластерного анализа	24
2.2.3 Разработка диаграммы последовательности выполнения кластерного анализа	25
2.2 Разработка логической модели данных сетевой информационной системы для выполнения кластерного анализа.....	26
2.3 Выбор архитектуры и средств разработки сетевой информационной системы для выполнения кластерного анализа	28

2.3.1 Выбор архитектуры сетевой информационной системы для выполнения кластерного анализа	28
2.3.2 Выбор среды разработки программного обеспечения сетевой информационной для выполнения кластерного анализа	29
2.3.3 Физическая модель данных сетевой информационной системы для выполнения кластерного анализа	31
2.3.4 Диаграмм развертывания сетевой информационной системы для выполнения кластерного анализа	33
2.4 Описание работы сетевой информационной системы для выполнения кластерного анализа	35
ЗАКЛЮЧЕНИЕ	38
СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ	39
ПРИЛОЖЕНИЕ А Фрагменты программного кода приложения	41

ВВЕДЕНИЕ

Кластерный анализ – одна из основных задач поискового анализа данных и общий метод анализа статистических данных, используемый во многих областях, включая машинное обучение, распознавание образов, анализ изображений, поиск информации, биоинформатику, сжатие данных и компьютерную графику.

Выполнение кластерного анализа без применения вычислительной техники, оснащенной необходимым программным обеспечением, представляет собой сложный и трудоемкий процесс, не гарантирующий высокой точности полученных результатов.

Кроме того, зачастую кластерный анализ проводится группой пользователей в рамках общего исследовательского проекта. Вполне понятно, что применение сетевых информационных систем (ИС) для кластерного анализа могло бы существенно увеличить производительность таких исследований.

Таким образом, **актуальность темы** выпускной квалификационной (ВКР) работы обусловлена необходимостью разработки сетевого программного обеспечения для выполнения кластерного анализа.

Объектом исследования ВКР является кластерный анализ.

Предмет исследования – сетевое программное обеспечение для выполнения кластерного анализа.

Целью ВКР является разработка сетевого программного обеспечения для кластерного анализа.

Для достижения поставленной цели необходимо решить следующие задачи:

- дать обзор существующих алгоритмов и выбрать алгоритм для кластерного анализа;
- разработать логическую модель сетевой информационной системы (ИС) для выполнения кластерного анализа;
- разработать логическую модель данных сетевой ИС;

- выбрать средства разработки программного обеспечения (ПО) сетевой ИС;
- разработать ПО сетевой ИС;
- проверить работоспособность ПО сетевой ИС и подтвердить его работоспособность.

Методы исследования: алгоритмы кластерного анализа, Web-технология, объектно-ориентированное программирование.

Практическая значимость работы заключается в разработке сетевого программного обеспечения для выполнения кластерного анализа.

Бакалаврская работа состоит из введения, двух глав, заключения, приложения и списка используемой литературы.

Во введение обозначается тема работы и ее актуальность, описывается объект и предмет исследования, цели и задачи, которые необходимо решить в данной работе.

Первая глава включает в себя описание объекта автоматизации – кластерного анализа, обзор и анализ известных алгоритмов, выбор алгоритма для сетевой ИС.

Вторая глава посвящена разработке логической модели и модели данных сетевой ИС и ее модели данных.

Выбрано средство разработки ПО ИС. Описаны процесс разработки и функциональность ПО ИС.

В заключении описываются результаты выполнения ВКР.

В приложении приведены фрагменты программного кода приложения сетевой ИС для выполнения кластерного анализа.

Глава 1 АНАЛИЗ МЕТОДОВ И ВЫБОР АЛГОРИТМА КЛАСТЕРНОГО АНАЛИЗА

1.1 Методы кластерного анализа

Кластерный анализ или кластеризация – это задача группировки набора объектов таким образом, что объекты в одной и той же группе (называемой кластером) были более схожи (в том или ином смысле) друг с другом, чем с другими группами (кластерами) [4].

Кластерный анализ - это не один конкретный алгоритм, а общая задача, которую нужно решить. Это может быть достигнуто с помощью различных алгоритмов, которые значительно различаются и по своему представлению о том, что представляет собой кластер, и по подходу, как их эффективно его найти.

Популярными понятиями кластеров являются группы с небольшими расстояниями между членами кластера, плотные области пространства данных, интервалы или отдельные статистические распределения. Поэтому кластеризация может быть сформулирована как задача многоцелевой оптимизации. Соответствующий алгоритм кластеризации и настройки параметров (включая такие значения, как используемая функция расстояния, порог плотности или количество ожидаемых кластеров) зависят от индивидуального набора данных и предполагаемого использования результатов.

Кластерный анализ как таковой является итеративным процессом обнаружения знаний или интерактивной многоцелевой оптимизации, которая включает в себя метод проб и ошибок.

Часто необходимо модифицировать предварительную обработку данных и параметры модели до тех пор, пока не будет достигнут желаемый результат.

Математическая постановка задачи кластерного анализа представлена ниже.

Пусть $m \leq n$. Необходимо на основании данных о Π -пространстве разбить множество из n операционных таксономических единиц (ОТЕ) на m подмножеств (кластеров) K_1, K_2, \dots, K_m так, чтобы ни одна ОТЕ не принадлежала 2-м или более кластерам [7].

Существует несколько различных методов, которые можно использовать для проведения кластерного анализа.

Рассмотрим подходы к кластерному анализу.

1.1.1 Иерархические методы кластерного анализа

Иерархические методы кластерного анализа подразделяются на:

– агломерационные методы, в которых каждый объект первоначально находится в отдельном кластере. Затем объединяются два «ближайших» (наиболее похожих) кластера, процесс повторяется до тех пор, пока все объекты не будут в одном кластере. В конечном итоге оптимальное количество кластеров выбирается из всех кластерных решений;

– дивизивные методы, в которых все объекты первоначально находятся в одном кластере, и приведенная выше стратегия применяется в обратном порядке, пока каждый объект не перейдет в отдельный кластер.

В рамках этого подхода к кластерному анализу существует ряд различных методов, используемых для определения того, какие кластеры должны быть объединены на каждом этапе.

Рассмотрим основные методы данного подхода:

– метод ближайшего соседа (метод одиночной связи).

В этом методе расстояние между двумя кластерами определяется как расстояние между двумя ближайшими членами или соседями. Этот метод относительно прост, но его часто критикуют, поскольку он не учитывает структуру кластера и может привести к проблеме, называемой цепочкой, в результате которой кластеры становятся длинными и страстными. Тем не менее, это лучше, чем другие методы, когда естественные кластеры не являются сферическими или эллиптическими по форме;

– метод самого дальнего соседа (полный метод сцепления).

В этом случае расстояние между двумя кластерами определяется как максимальное расстояние между членами - то есть расстояние между двумя субъектами, которые расположены дальше всего. Этот метод имеет тенденцию создавать компактные кластеры одинакового размера, но как и в методе ближайшего соседа, не учитывает структуру кластера. Он также весьма чувствителен к выбросам;

– средний (между группами) метод связи.

Расстояние между двумя кластерами рассчитывается как среднее расстояние между всеми парами предметов в двух кластерах. Это считается довольно надежным методом;

– метод кластерного центроида.

Здесь вычисляется центроид (среднее значение для каждой переменной) каждого кластера и используется расстояние между центроидами. Кластеры, чьи центроиды расположены ближе всего, сливаются. Этот метод достаточно надежен;

– метод Уорда.

В этом методе объединяются все возможные пары кластеров и вычисляется сумма квадратов расстояний внутри каждого кластера. Затем выполняется суммирование по всем кластерам. Выбирается комбинация, которая дает наименьшую сумму квадратов. Этот метод имеет тенденцию создавать кластеры примерно одинакового размера, что не всегда желательно. Он также весьма чувствителен к выбросам. Несмотря на это, это один из самых популярных методов, как и средний метод сцепления.

Обычно рекомендуется использовать два или три из вышеперечисленных методов. Если методы согласуются достаточно хорошо, то результаты будут гораздо более правдоподобными.

1.1.2 Неиерархические методы кластерного анализа

Неиерархические методы.

В этих методах заранее задано желаемое количество кластеров, и выбирается «лучшее» решение.

Примером такого метода является k-means (k-средние), который обычно состоит из следующих этапов:

- выбираются начальные центры кластеров (по существу это набор наблюдений, которые находятся далеко друг от друга - каждый объект формирует кластер из одного, а его центр - значение переменных для этого объекта;
- для каждого объекта устанавливается его ближайший кластер, определенный в терминах расстояния до центроида;
- находятся центроиды кластеров, которые были сформированы;
- пересчитывается расстояние от каждого объекта до каждого центроида и обращаем внимание на объекты, которые не находятся в том кластере, к которому они ближе всего;
- процесс продолжается до тех пор, пока центроиды не станут относительно стабильными.

Неиерархический кластерный анализ, как правило, используется, когда задействованы большие наборы данных, что иногда предпочтительнее, поскольку позволяет объектам перемещаться из одного кластера в другой (это невозможно в иерархическом кластерном анализе, когда объект после его назначения не может перемещаться в другой кластер) [19].

1.2 Описание алгоритма кластерного анализа

Как показывает практика, метод k-means является самым популярным методом кластерного анализа, так как реализуется с помощью быстрого и относительно простого алгоритма [20].

Поэтому выбираем в качестве алгоритма кластеризации для сетевой ИС алгоритм k-means.

Алгоритм k-means описывается с помощью следующего выражения:

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2,$$

где k -число кластеров, S_i – полученные кластеры, $i = 1, 2, \dots, k$ и μ_i – центры масс (центроиды) векторов $x_j \in S_i$.

Пример графического представления алгоритма приведен на рисунке 1.1 [11].

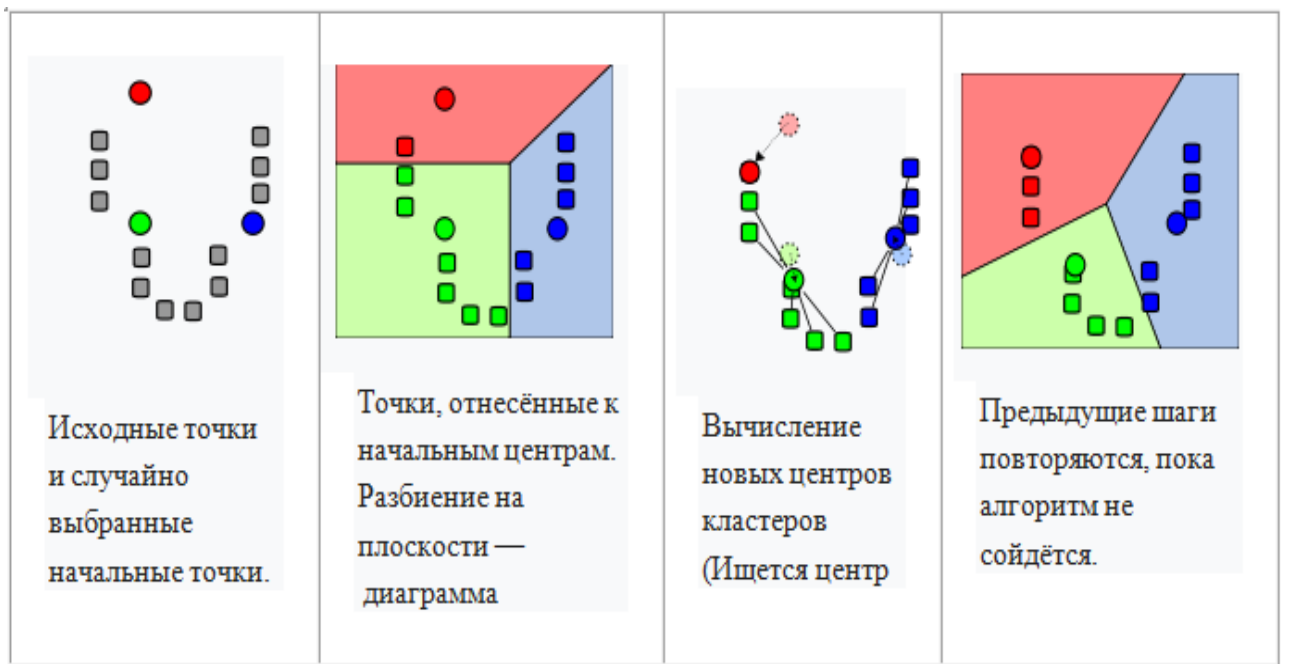


Рисунок 1.1- Представление алгоритма k-means

Действие алгоритма - в двумерном случае. Начальные точки выбраны случайно.

Данный алгоритм также называется алгоритмом Ллойда.

Алгоритм Ллойда довольно просто реализуется и применяется даже для больших наборов данных.

Он успешно используется для исследований в различных областях, включая сегментацию рынка, компьютерное зрение, геостатистику, астрономию и сельское хозяйство.

1.3 Постановка задачи на разработку сетевого программного обеспечения для выполнения кластерного анализа

Для выполнения кластерного анализа в сетевом режиме принято решения использовать сетевую информационную систему [3].

На основании проведенного анализа и пожеланий Заказчика определены следующие требования к архитектуре и функциональности сетевой ИС:

- выполнение кластерного анализа в режиме онлайн;
- реализация алгоритма k-means;
- использование для разработки ПО сетевой ИС современных Web-технологий [2];
- низкая стоимость владения ИС.
- для поддержки и эксплуатации ПО ИС от персонала не должно требоваться специальных технических навыков, знаний технологий или программных продуктов, за исключением общих навыков работы с персональным компьютером и стандартным Web-браузером.

Следует отметить, что ключевым моментом алгоритма k-means является вычисление на каждой итерации расстояния между записями и центрами кластеров, что необходимо для определения, к какому из кластеров принадлежит данная запись. Правило, по которому производится вычисление расстояния в многомерном пространстве признаков, называется метрикой.

Наиболее часто в практических задачах кластеризации используются следующие метрики:

- евклидово расстояние;
- расстояние Манхэттена.

В системе рекомендуется использовать в качестве метрики «евклидово расстояние», как более простое в реализации.

В математике евклидово расстояние или евклидова метрика является расстоянием по прямой между двумя точками в евклидовом пространстве, вычисляемой по теореме Пифагора. На этом расстоянии евклидово пространство становится метрическим пространством.

С учетом вышеперечисленных требований были выявлены аналоги сетевой ИС и произведен их сравнительный анализ.

1.4 Обзор и анализ аналогов программного обеспечения для выполнения кластерного анализа

Для сравнения на предмет соответствия выработанным требованиям к ПО сетевой ИС были выбраны ИТ-решения, относящиеся к категории информационных систем поддержки задач финансового рынка.

1.4.1 Платформа ClusterDELTA

ClusterDelta.com - биржевой аналитический портал по торговле объемами на фьючерсах и акциях [12].

Благодаря современным технологиям данные по указанным объемам стали более доступны рядовым трейдерам.

До этого времени основными стратегиями анализа объемов были Market Profile и VSA. Market Profile - это попытка отследить рыночную активность в срезе 30-минутных свечей. VSA - анализ действий умных денег через гистограмму объема в рамках временного терминала.

Кластерный анализ - это наиболее точный и наиболее детальный вид анализа, так как основой для анализа является каждая сделка, которая происходит на бирже. Объединение сделок с помощью математических операций за определенный критерий (промежуток времени, величина объема) на выходе будет представлено в виде распределения объема по каждой цене - это кластер.

График цены представленный в виде кластеров называем кластерным графиком (рисунок 1.2).

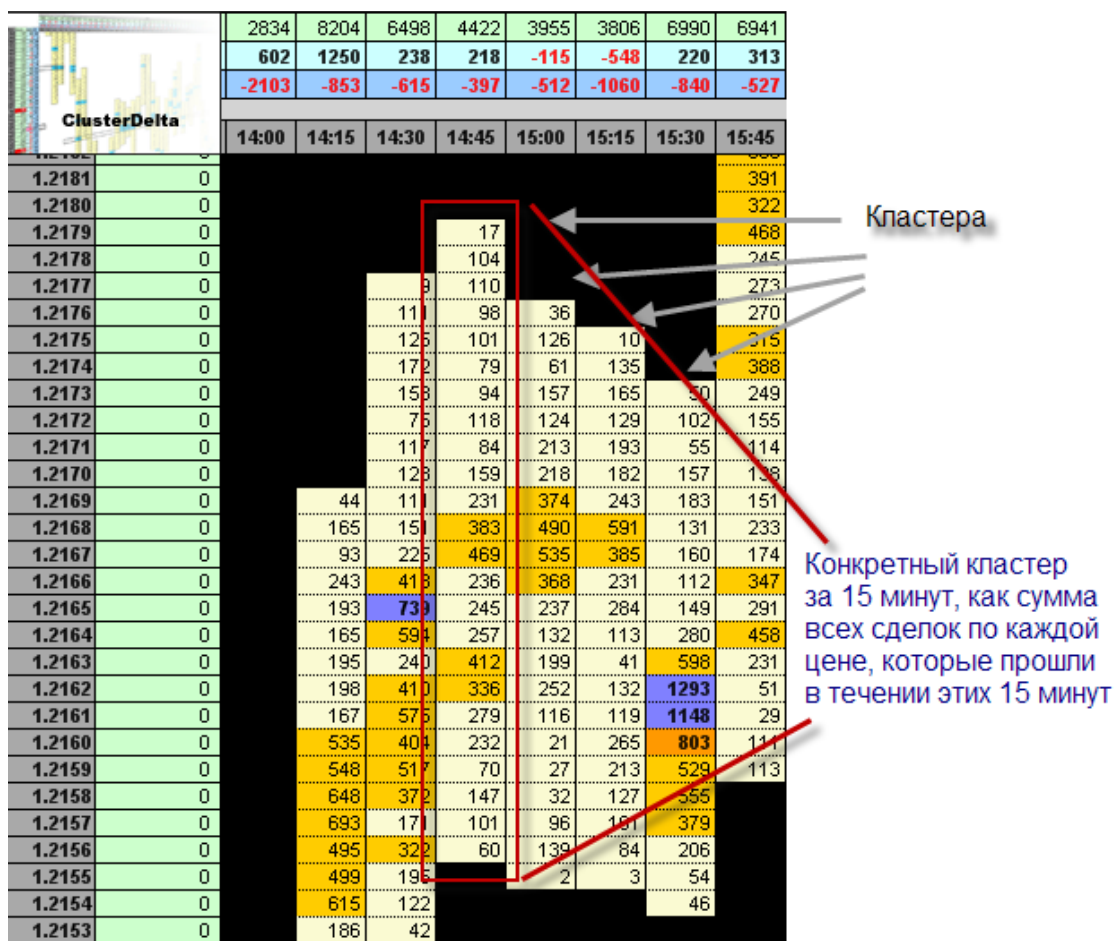


Рисунок 1.2 – Пример кластерного графика ClusterDELTA

В отличие от портала Форекса, все данные по инструментам торгуемым на бирже могут исходить только от одного источника, соответственно эти данные (цена/объем) не могут различаться независимо от того, как и где они были получены, при условии конечно, что получены они были с биржи и не фильтровались.

Однако существующая функциональность биржевых терминалов, ограничения поставщиков данных и прочие технические и финансовые условия практически не дают возможности рядовому трейдеру полноценно делать анализ кластеров на исторических данных на больших периодах.

В связи с этим появилась идея разработать платформу, которая решала бы эти поставленные задачи. Так появились первые наработки будущей платформы ClusterDelta, которая, со временем, из скриптов написанных на VBA в MS Excel перешла в современную фазу и стала полнофункциональной системой, предоставляющей кластерные данные в режиме реального времени.

1.4.2 Сетевая версия программного комплекса STATISTICA

Платформа анализа данных STATISTICA предлагает самую полную и эффективную систему удобных для пользователя инструментов для всего процесса интеллектуального анализа данных - от запросов к базам данных до создания окончательных отчетов [10].

Программное обеспечение StatSoft для интеллектуального анализа данных и интеллектуального моделирования доступно в рамках единой рабочей станции, многопользовательских (одновременное лицензирование пользователей) и выпусков Enterprise edition.

В программном комплексе STATISTICA реализованы классические методы кластерного анализа, включая методы k-средних, иерархической кластеризации и двухвходового объединения (рисунок 1.3).

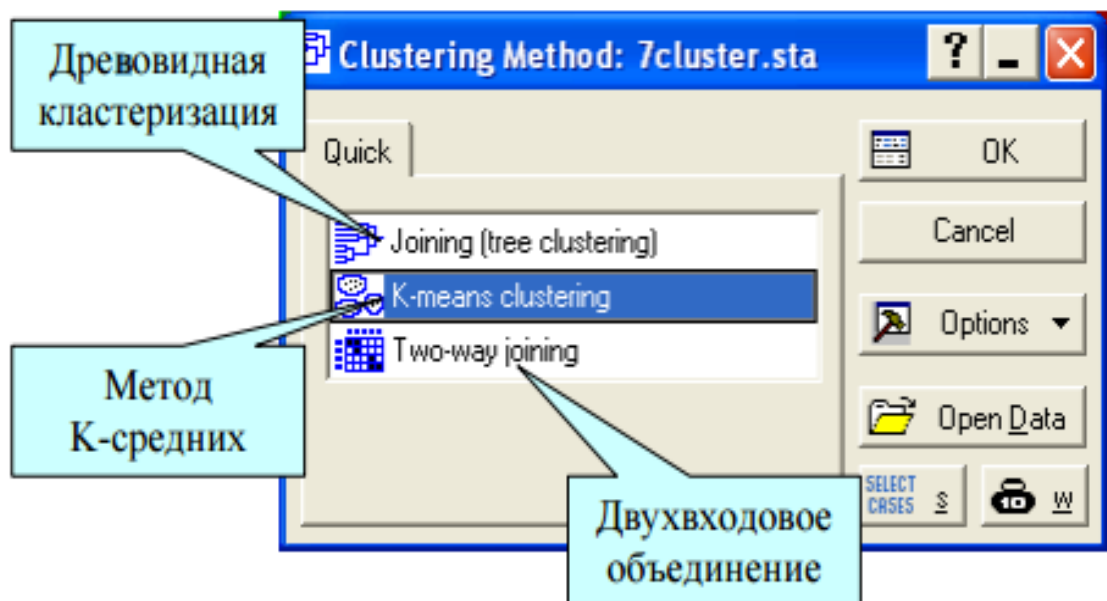


Рисунок 1.3 – Окно выбора метода кластерного анализа программного пакета STATISTICA

Данные могут поступать как в исходном виде, так и в виде матрицы расстояний между объектами.

Наблюдения и переменные можно кластеризовать, используя различные меры расстояния (евклидово, квадрат евклидова, манхэттенское, Чебышева и

др.) и различные правила объединения кластеров (одиночная, полная связь, невзвешенное и взвешенное попарное среднее по группам и др.

STATISTICA является мощным программным решением, предназначенным для анализа процессов, выявления и прогнозирования проблем в различных сферах исследования

1.4.3 Программа Visual Volume

Visual Volume - простая программа, разработанная специально для анализа истории с помощью кластерных графиков, счетчика рейтинговых цен и фильтрации тиков (рисунок 1.4) [13].

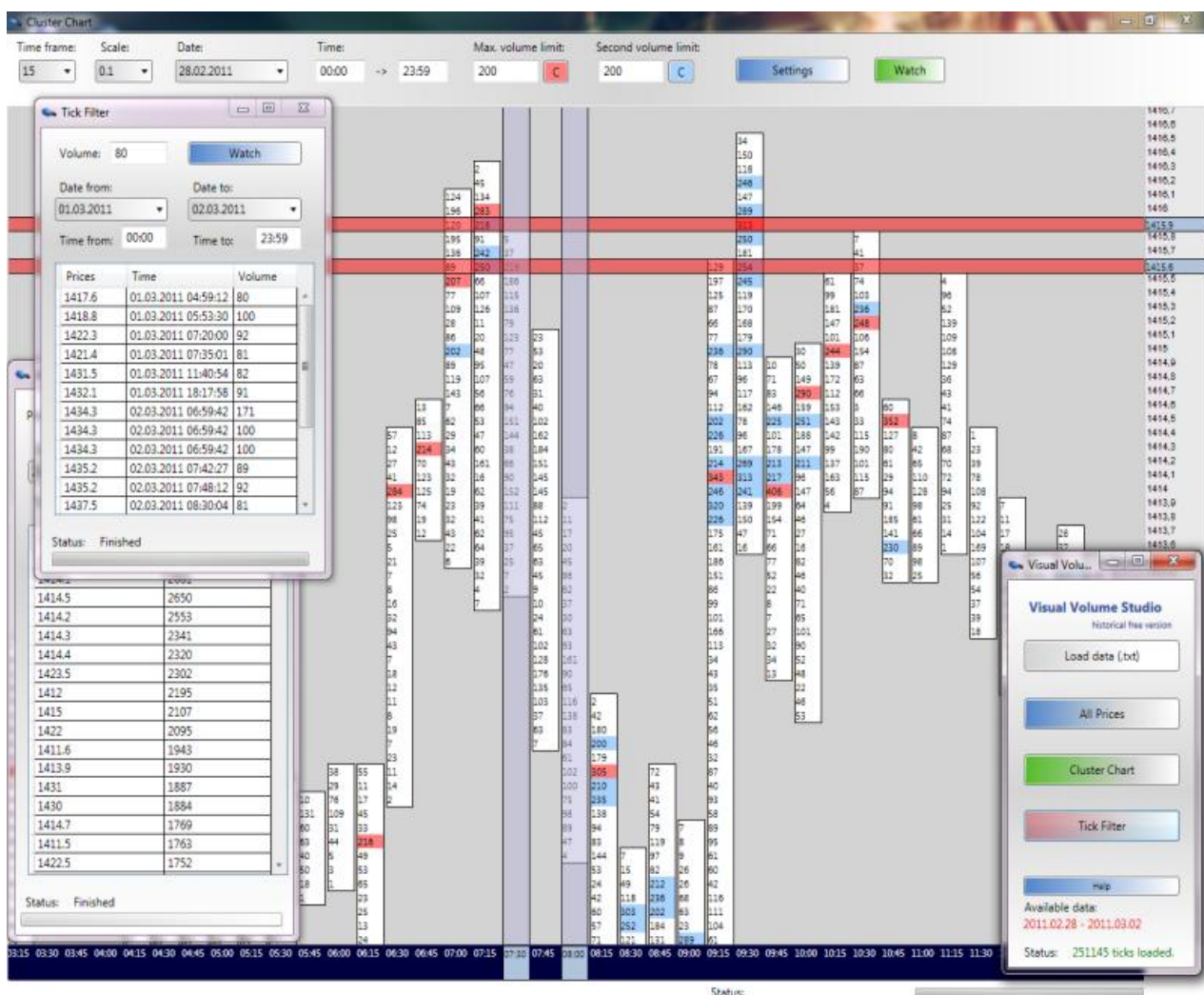


Рисунок 1.4 – Рабочее окно программы Visual Volume

Программа создана именно для анализа истории в онлайн-режиме.

Историческая версия Visual Volume распространяется бесплатно, также можете публиковать программу на ресурсе пользователя, с разрешения автора и обратной ссылкой на сайт автора.

Преимущества программы:

- функциональность программы дает возможность разрабатывать ручные торговые системы, анализировать поведение рынка относительно внутридневных накоплений и рейтинговых цен за длительные промежутки времени;

- по мнению разработчика, это несомненно может дать пользователю преимущество в реальной торговле, так как цена формируется спросом и предложением и исследование таких закономерностей дает более здоровое представление о рыночных процессах, чем классические методы анализа, которые описываются в книгах или интернет-форумах.

Качество данных зависит от поставщика. В данном случае данные выкачиваются либо из платформы NinjaTrader 7 (Zen-Fire data feed), либо через сайт Finam.ru.

Данные сравнительного анализа аналогов приведены в таблице 1.1.

Таблица 1.2 - Сравнительный анализ аналогов сетевой ИС

Характеристика/Аналог	ClusterDELTA	Statistica	Visual Volume
выполнение кластерного анализа в режиме онлайн	+	-	-
реализация алгоритма k-means	+	+	+
использование для разработки ПО сетевой ИС современных Web-технологий	+	-	-
простота эксплуатации	-	+	+
низкая стоимость владения	-	-	+
Итого	3	2	3

Анализ известных ИТ-решений показал, что они не соответствуют в полной мере вышеперечисленным требованиям по архитектуре и функциональности сетевой ИС..

На основании вышеизложенного был сделан вывод о целесообразности разработки новой сетевой ИС для выполнения кластерного анализа.

Выводы к главе 1

Выполнение кластерного анализа без применения вычислительной техники, оснащенной необходимым программным обеспечением, представляет собой сложный и трудоемкий процесс, не гарантирующий высокой точности полученных результатов.

Как показывает практика, метод k-means является самым популярным методом кластерного анализа, так как реализуется с помощью быстрого и относительно простого алгоритма. Потому выбираем в качестве алгоритма кластеризации для сетевой ИС алгоритм метода k-means.

Анализ известных ИТ-решений показал, что они не удовлетворяют в полной мере требованиям, предъявляемым к программному обеспечению сетевой ИС.

В этой связи принято решение о разработке новой сетевой ИС для выполнения кластерного анализа.

Глава 2 РАЗРАБОТКА И РЕАЛИЗАЦИЯ ПРОЕКТНЫХ РЕШЕНИЙ СЕТЕВОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ВЫПОЛНЕНИЯ КЛАСТЕРНОГО АНАЛИЗА

2.1 Разработка программной архитектуры сетевой информационной системы для выполнения кластерного анализа

Программная архитектура или архитектура программного обеспечения - архитектура взаимодействия различных классов в рамках одного приложения.

Архитектура программного обеспечения относится к структурам высокого уровня программной системы, дисциплине создания таких структур и документации этих структур.

Эти структуры необходимы для создания представления о программной системе. Каждая структура содержит программные элементы, отношения между ними, свойства элементов и отношений.

Архитектура программного обеспечения – это фундаментальные структурные решения, которые, как правило, не изменяются.

Выбор архитектуры программного обеспечения включает в себя конкретные структурные решения, позволяющие реализовать специфические особенности разрабатываемого программного обеспечения.

В настоящее время для разработки архитектуры программного обеспечения используются два подхода [8]:

- концепция интеграции приложений EAI (Enterprise Application Integration);
- сервисно-ориентированная архитектура SOA (Service Oriented Architecture).

EAI - это интеграционная структура, состоящая из набора технологий и сервисов, которые образуют промежуточное программное обеспечение или «промежуточную инфраструктуру», что позволяет интегрировать системы и приложения на предприятии.

Многие типы программного обеспечения для бизнеса, такие как ERP-системы, CRM-приложения для управления клиентами, приложения бизнес-аналитики, системы расчета заработной платы и человеческих ресурсов, как правило, не могут связываться друг с другом для обмена данными или бизнес-правилами. Это отсутствие связи приводит к неэффективности, при которой идентичные данные хранятся в нескольких местах, или простые процессы не могут быть автоматизированы.

Интеграция корпоративных приложений – это процесс объединения таких приложений в рамках единой организации, чтобы упростить и автоматизировать бизнес-процессы в максимально возможной степени и в то же время избежать необходимости радикальных изменений в существующих приложениях или структурах данных.

SOA – это технология разработки программного обеспечения, в которой сервисы предоставляются другим компонентам приложения через протокол связи по сети, в том числе по интернет-каналу (Web-сервис).

Основные принципы SOA не зависят от поставщиков, продуктов и технологий.

SOA может быть реализована с помощью Web-сервисов, интегрированных с помощью известных стандартных протоколов (SOAP, WSDL).

Это делается для того, чтобы функциональные блоки-блоки были доступны через стандартные интернет-протоколы, которые не зависят от платформ и языков программирования. Эти службы могут представлять собой либо новые приложения, либо просто обертки вокруг существующих устаревших систем, чтобы сделать их сетевыми.

Web-сервис представляет собой дискретную единицу функциональности, к которой можно получить доступ удаленно, а также использовать и обновлять независимо.

Для выбора технологии разработки программной архитектуры сетевой ИС сформирована таблица 2.1.

Таблица 2.1 - Сравнительный анализ программных архитектур сетевой ИС

Характеристика/Аналог	EAI	SOA
простота модификации при изменении типа алгоритма кластерного анализа	-	+
низкая стоимость владения	-	+
Итого	0	2

Как показывает практика, SOA довольно активно применяется в облачных вычислениях, что позволяет отнести данную технологию к современным мировым трендам в области проектирования программных архитектур распределенных систем [17].

По результатам сравнения в качестве технологии разработки программной архитектуры сетевой ИС для выполнения кластерного анализа выбираем сервисно-ориентированную архитектуру – SOA.

В качестве Web-сервиса используем бесплатно-распространяемый Web-сервис, реализующий алгоритм k-means [14].

2.2 Логическое моделирование сетевой информационной системы для выполнения кластерного анализа

Логическое моделирование используется для формирования подробных требований к системе приложений.

Моделирование сетевой ИС на логическом уровне с помощью UML создает подробную модель процесса выполнения кластерного анализа, используемую для описания функциональных требований прикладной системы, на основе которой может быть получена ее физическая реализация [21].

2.2.1 Разработка диаграммы вариантов использования процесса выполнения кластерного анализа

Диаграмма вариантов использования (use case diagram) – это графическое изображение взаимодействий между элементами системы.

Это поведенческий тип диаграмм UML, часто используемый для анализа различных систем.

Диаграмма вариантов использования позволяет визуализировать разные типы ролей в системе и как эти роли взаимодействуют с системой.

Следует отметить, что диаграмма использования не отражает детализацию вариантов использования, она лишь объединяет некоторые отношения между вариантами использования, актерами и системами.

В процессе выполнения кластерного анализа выделены следующие актеры:

- Пользователь сетевой ИС;
- Web-сервис.

Рассмотрим, какие возможности должно предоставлять разрабатываемое программное обеспечение:

- Пользователь выполняет кластерный анализ.
- Web-сервис предоставляет необходимую функциональность.

В таблице 2.2 выделены прецеденты, которые должны быть реализованы в новой сетевой ИС.

Таблица 2.2 – Краткое описание прецедентов

Прецеденты	Актеры	Краткое описание
Выполнение кластерного анализа	Пользователь	Активизация процедуры кластерного анализа сетевой ИС
Поддержка процесса кластерного анализа	Web-сервис	Предоставление функциональности кластерного анализа

Диаграмма вариантов использования выполнения кластерного анализа, построенная с учетом описанных рекомендаций, изображена на рисунке 2.1.

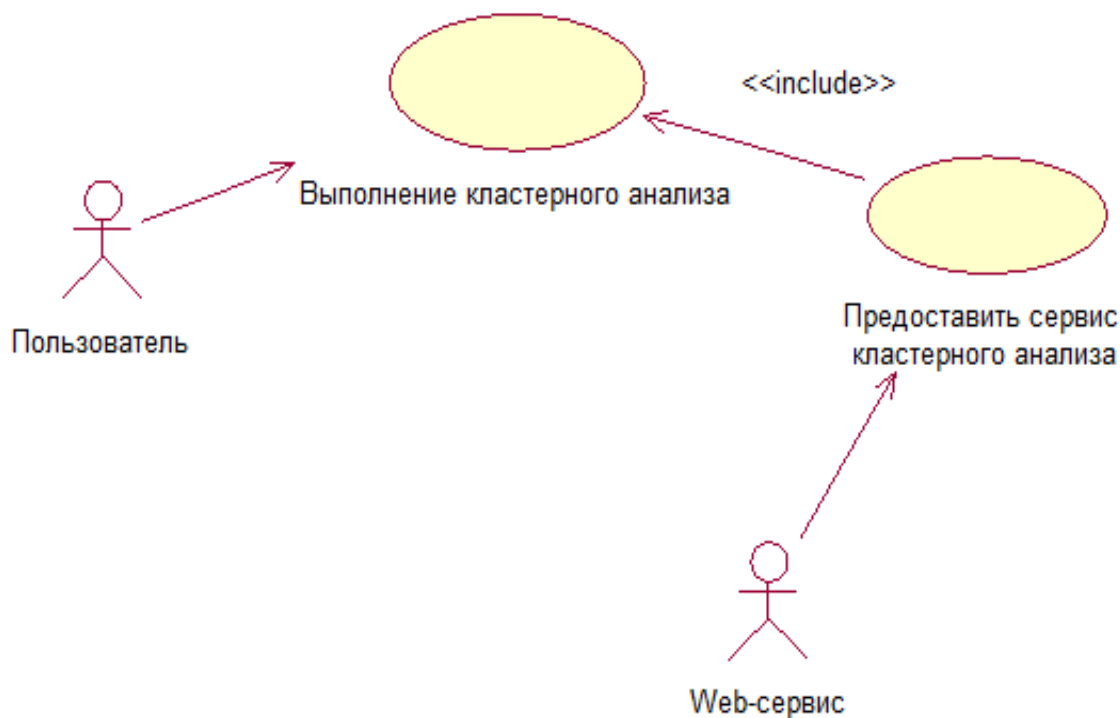


Рисунок 2.1 – Диаграмма вариантов использования выполнения кластерного анализа

Данные прецеденты являются основными технологическими этапами, которые необходимо реализовать.

Диаграмма вариантов использования отражает функциональный аспект программного обеспечения сетевой ИС.

2.2.2 Разработка диаграммы классов сетевой информационной системы для выполнения кластерного анализа

Диаграмма классов (Class diagram) служит для описания состава атрибутов классов, а также для отображения взаимосвязей между классами используется. Эта диаграмма отражает статический аспект системы.

Диаграмма классов сетевой ИС изображена на рисунке 2.2.

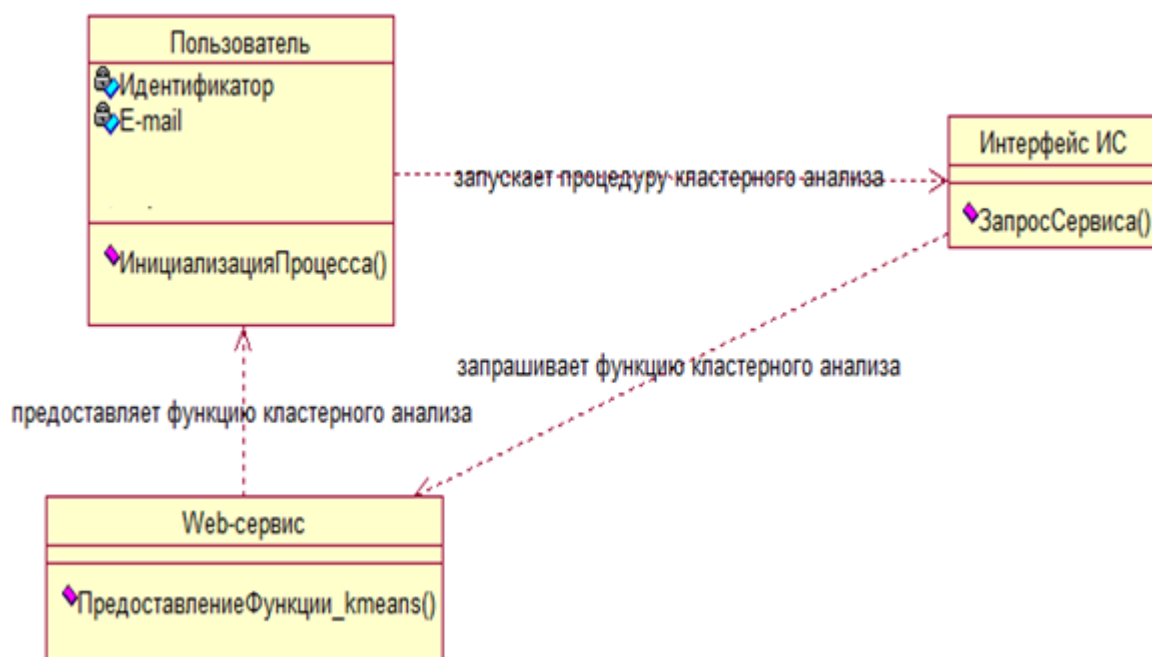


Рисунок 2.2 – Диаграмма классов сетевой ИС

Приведем спецификацию классов, приведенных на диаграмме.

Пользователь – объект, инициализирующий процесс выполнения кластерного анализа в сетевой ИС.

Интерфейс ИС – объект, выполняющий запрос функции кластерного анализа у Web-сервиса;

Web-сервис – объект, предоставляющий функцию кластерного анализа Пользователю.

2.2.3 Разработка диаграммы последовательности выполнения кластерного анализа

Динамический аспект моделирования позволяет описать функционирования процессов во времени, что позволяет всесторонне проанализировать предметную область моделирования.

В методологии UML такие модели могут быть построены с помощью диаграмм последовательности.

На рисунке 2.3 представлена диаграмма последовательности, моделирующая сценарий выполнения кластерного анализа в сетевой ИС.

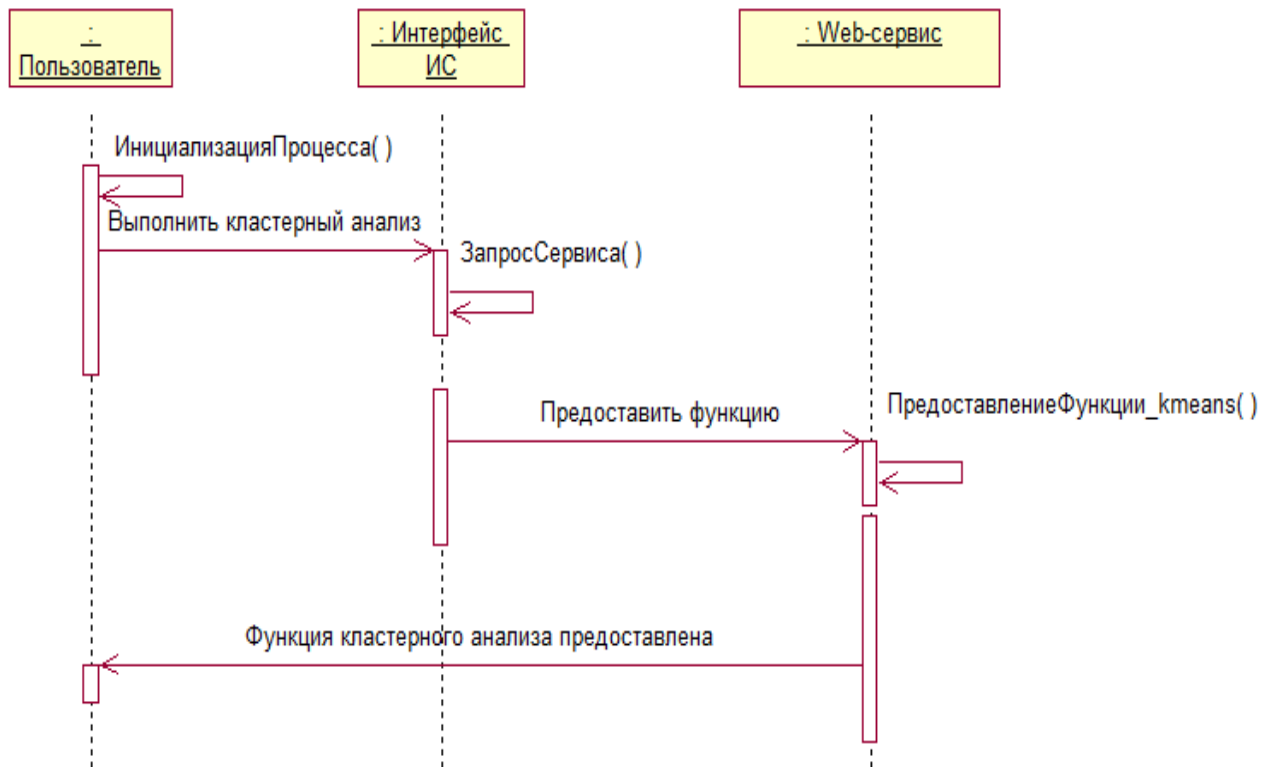


Рисунок 2.3 – Диаграмма последовательности выполнения кластерного анализа

В начальный момент времени объект Пользователь обращается к объекту Интерфейс ИС с сообщением «Выполнить кластерный анализ».

Объект Интерфейс ИС обращается к объекту Web-сервис с запросом предоставить функцию кластерного анализа.

Объект Web-сервис предоставляет функцию кластерного анализа Пользователю ИС.

Процесс выполнения кластерного анализа завершен.

2.2 Разработка логической модели данных сетевой информационной системы для выполнения кластерного анализа

Метод семантического моделирования в настоящее время широко используется для разработки структуры базы данных. Семантическое моделирование позволяет описывать структуры данных, основанные на значении этих данных.

В качестве инструмента семантического моделирования используются различные варианты диаграмм «сущности-связь» ERD (Entity-Relationship Diagram).

На диаграммах, основанных на ER-модели, есть три основных компонента: сущность, отношение и атрибуты.

Сущность – это класс похожих объектов, информация о которых должна учитываться в модели. Каждый объект должен иметь имя, выраженное существительным в единственном числе.

Диаграммы «сущность-связь» используются для логического моделирования данных информационных систем.

Нормализованная логическая модель данных сетевой ИС для выполнения кластерного анализа изображена на рисунке 2.4 [1,5].

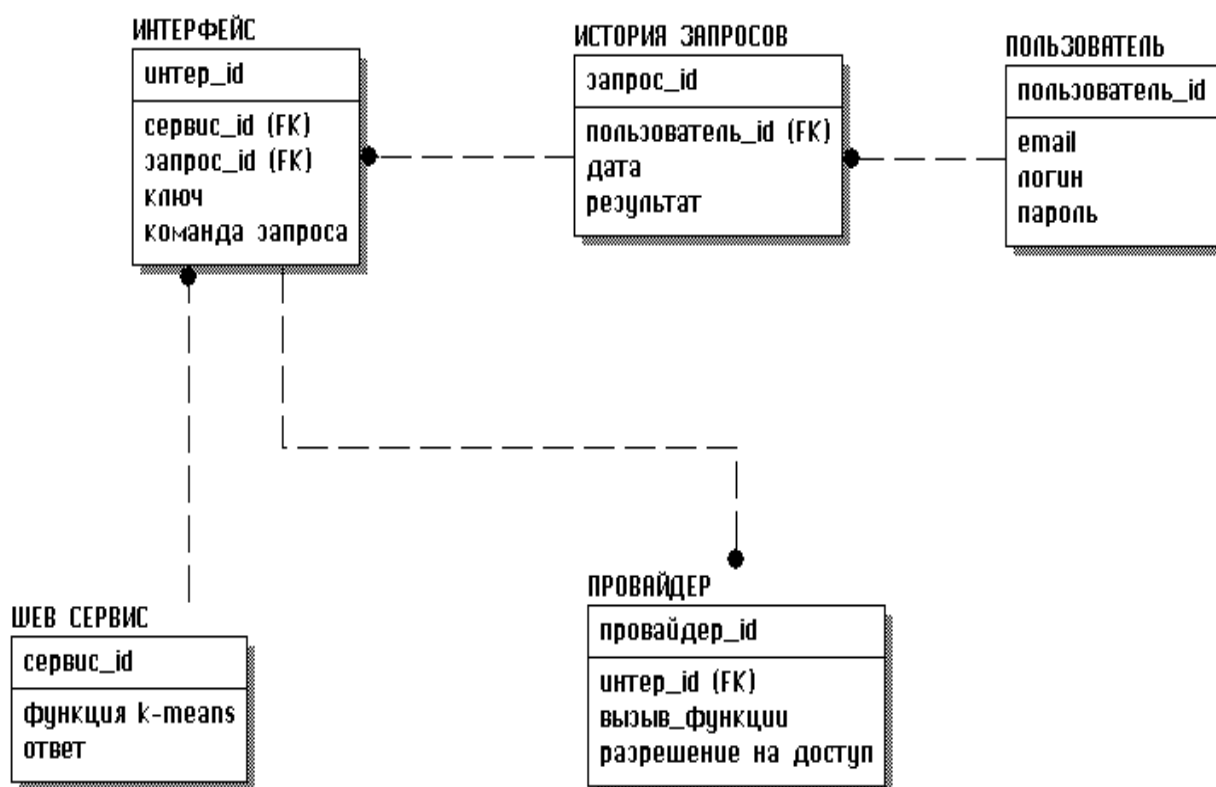


Рисунок 2.4 – Логическая модель данных сетевой ИС для выполнения кластерного анализа

Связи между сущностями описываются следующим образом:

- Пользователь имеет Историю запросов («один ко многим»);

- История запросов связана с Интерфейсом отношением «один ко многим»;
 - Web-сервис связан с Интерфейсом отношением «один ко многим»;
 - Интерфейс связан с Провайдером отношением «один ко многим»;
- Все связи – неидентифицирующие.

2.3 Выбор архитектуры и средств разработки сетевой информационной системы для выполнения кластерного анализа

2.3.1 Выбор архитектуры сетевой информационной системы для выполнения кластерного анализа

Для выбора архитектуры сетевой ИС сравним возможности двух- и трехуровневой архитектуры «клиент-сервер».

Двухуровневая архитектура используется для описания клиент-серверных систем, в которых клиент запрашивает ресурсы, а сервер отвечает непосредственно на запрос, используя свои собственные ресурсы.

Это означает, что сервер не вызывает собственное приложение, чтобы предоставить часть услуги.

В трехуровневой архитектуре существует промежуточный уровень, при этом архитектура обычно разделяется между:

- клиентом-компьютером, который запрашивает ресурсы с помощью пользовательского интерфейса (обычно Web-браузера);
- сервером приложений (или промежуточным программным обеспечением), задачей которого является предоставление запрошенных ресурсов, но путем вызова другого сервера;
- сервером баз данных, который обеспечивает сервер приложений затребованными данными.

Для сравнения описанных архитектур используем таблицу 2.3.

Таблица 2.3 – Сравнительный анализ моделей архитектуры «клиент-сервер»

Критерии оценки	Двухуровневая архитектура	Трехуровневая архитектура
затраты на реализацию	+	-
гибкость	-	+
безопасность	-	+
производительность	-	+
Итого	1	3

По результатам сравнения выбираем трехуровневую архитектуру «клиент-сервер» для реализации сетевой ИС.

2.3.2 Выбор среды разработки программного обеспечения сетевой информационной для выполнения кластерного анализа

Как было отмечено выше, в качестве архитектуры ИС будет использована время трехзвенная архитектура «клиент-сервер» для Web-приложений, в которой:

- сервером базы данных является бесплатная система управления базами данных (СУБД) MySQL 5.x;
- Web-сервер построен на базе свободно распространяемого программного обеспечения Apache 2.x.

В качестве языка программирования Web-приложения ИС выбран язык PHP 5.6 [22].

Web-приложение предназначено для работы в операционной системе Windows.

Таким образом, для разработки ИС необходимо выбрать среду разработки «Windows, Apache, MySQL, PHP» - WAMP.

Для сравнительного анализа выбраны следующие WAMP-платформы: Denwer, XAMPP и Open Server.

Denwer – это отечественная локальная среда разработки для PHP + MySQL.

Denwer обеспечивает эмуляцию Web-сервера Apache с поддержкой PHP и MySQL на машине, работающей под управлением MS Windows. В настоящее время практически не развивается.

Кросс-платформенная среда XAMPP очень популярна за рубежом [15].

XAMPP существует более 10 лет и постоянно обновляется.

Open Server - относительно молодая, обновляемая портативная платформа, доступная без установки на операционную систему.

В таблице 2.4 представлены для сравнения основные показатели описанных платформ.

Таблица 2.4 – Сравнительный анализ WAMP-сред

Критерии оценки	Denwer	XAMPP	Open Server
свободное распространение	+	+	+
русификация	+	+	+
обновление	-	+	+
знание разработчиком	-	+	-
Итого	2	4	3

На основании результатов анализа выбрана среда XAMPP-WIN32-5.6.30-1-VC11.

Обмен с бесплатным Web-сервисом кластерного анализа осуществляется по протоколу SOAP .

SOAP (первоначально Simple Object Access Protocol) - это спецификация протокола для обмена структурированной информацией при реализации веб-сервисов в компьютерных сетях.

Его цель - вызвать расширяемость, нейтральность и независимость. Он использует XML Information Set для своего формата сообщения и опирается на протоколы уровня приложения, чаще всего - протокол передачи гипертекста

(HTTP) или Simple Mail Transfer Protocol (SMTP) для согласования и передачи сообщений.

SOAP позволяет обрабатывать процессы, работающие на разрозненных операционных системах (например, Windows и Linux), с помощью Extensible Markup Language (XML). Поскольку веб-протоколы, такие как HTTP, установлены и запущены во всех операционных системах, SOAP позволяет клиентам вызывать веб-службы и получать ответы независимо от языка и платформ.

2.3.3 Физическая модель данных сетевой информационной системы для выполнения кластерного анализа

Физическое моделирование данных – это создание команд DDL (языка определения данных SQL), содержащих инструкции создания схем, таблиц и индексов, включая ограничения, значения по умолчанию и т.п.

Физическая модель строится на основе логической модели данных и предназначена для конкретной платформы баз данных, в рассматриваемом случае – для СУБД MySQL.

Основные возможности, доступные в MySQL 5.6:

- широкое подмножество ANSI SQL 99, а также расширения;
- кросс-платформенная поддержка;
- хранимые процедуры, использующие процедурный язык, который строго придерживается SQL/ PSM;
- триггеры;
- курсоры;
- обновляемые виды;
- ACID при использовании InnoDB и NDB Cluster Storage Engines;
- поддержка SSL;
- кэширование запросов;
- вложенные запросы

– несколько механизмов хранения, позволяющие выбрать тот, который наиболее эффективен для каждой таблицы в приложении и др.

На рисунке 2.5 изображена физическая модель данных сетевой ИС для выполнения MySQL-кластерного анализа.

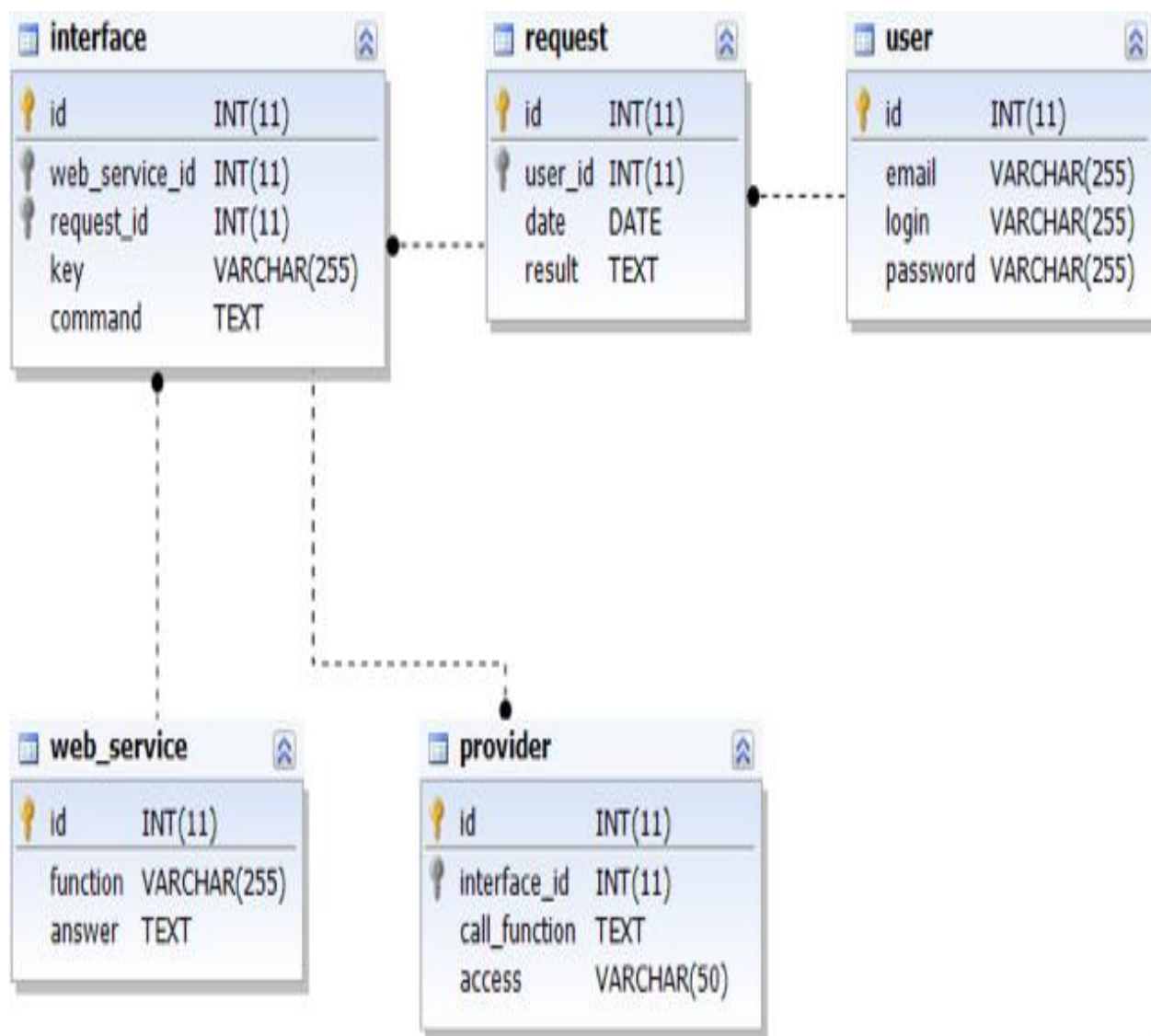


Рисунок 2.5 – Физическая модель данных сетевой ИС для выполнения кластерного анализа

Физическая модель данных построена в дизайнерае dbForge Studio for MySQL [9].

2.3.4 Диаграмм развертывания сетевой информационной системы для выполнения кластерного анализа

Диаграмма развертывания сетевой ИС представлена на рисунке 2.6.

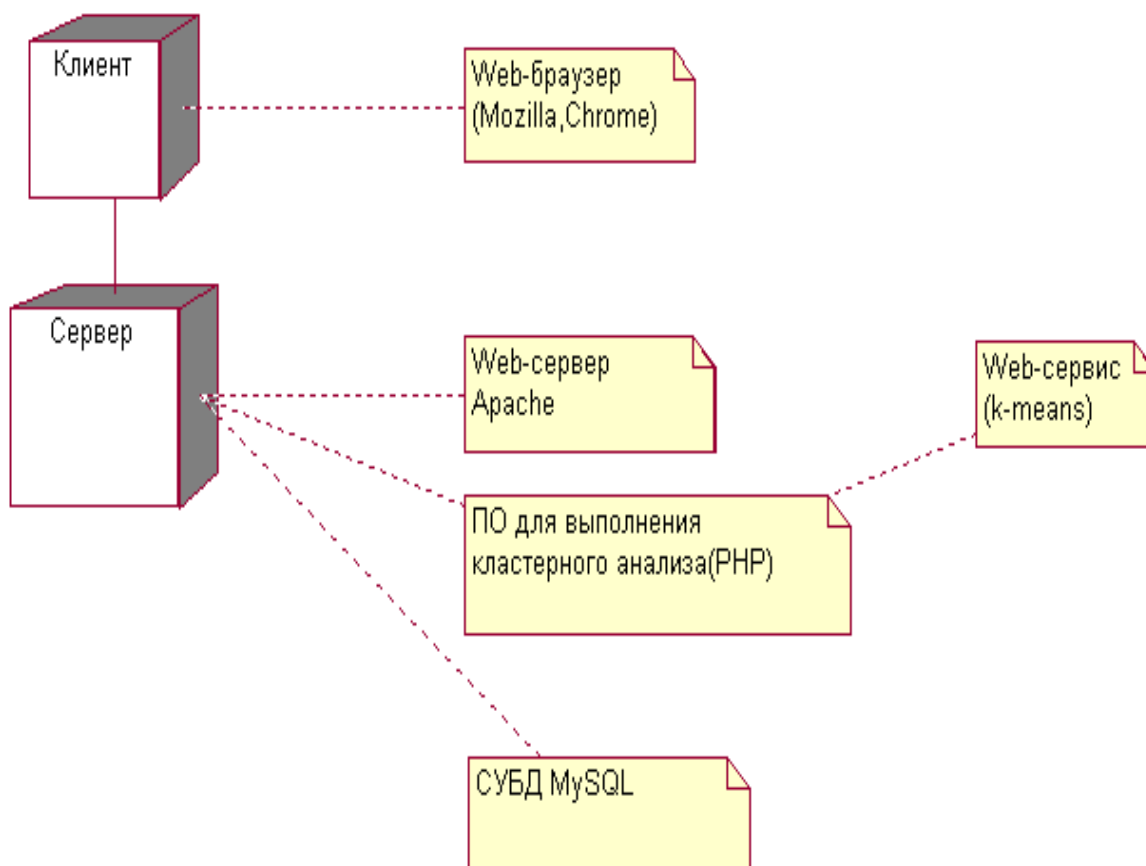


Рисунок 2.6 - Диаграмма развертывания сетевой ИС

ИС управления диспетчерской службы состоит из следующих компонентов:

- клиент – Web –браузер (Mozilla, Chrome);
- Web-сервер – построен на базе программного обеспечения Apache;
- сервер баз данных на основе СУБД MySQL.

Доступ к компонентам ИС осуществляется через интернет-канал.

На рисунке 2.7 изображена диаграмма компонентов программного обеспечение сетевой ИС.

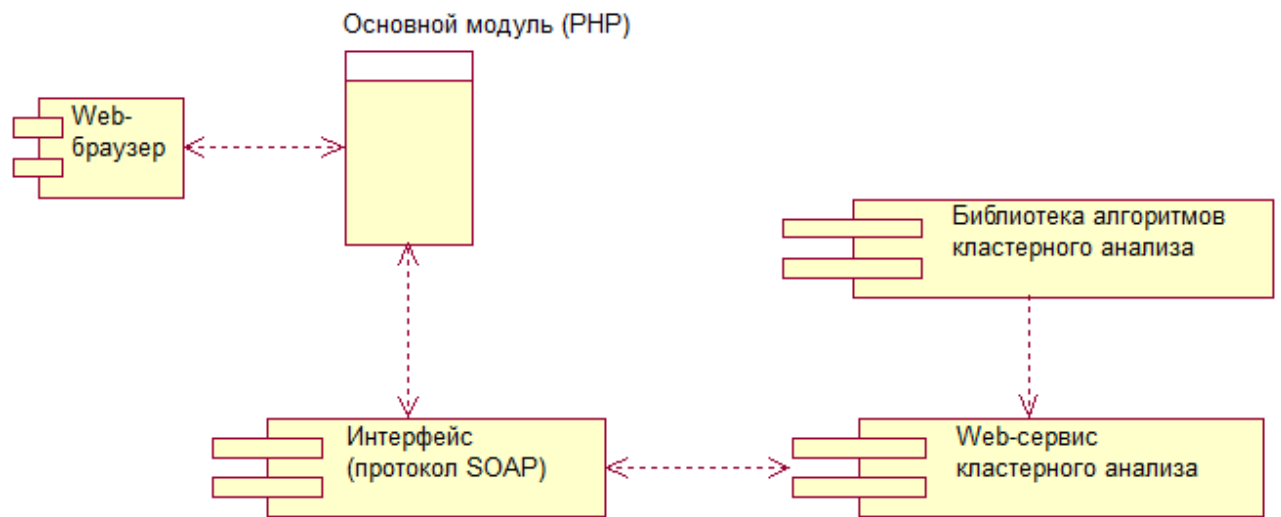


Рисунок 2.7 - Диаграмма компонентов сетевой ИС

Ниже представлен пример PHP-кода классов реализации компонентов ИС.

```

<?
class AnalysisController extends CController
{
    public function actions()
    {
        return array(
            'quote'=>array(
                'class'=>'CWebServiceAction',
                'classMap'=>array(
                    'RequestRows'=>'RequestRows',
                    'RequestRow'=>'RequestRow',
                    'RequestCell'=>'RequestCell',
                ),
            ),
        );
    }
}

/**

```

```

* @param array Исходные данные
* @return string Результат анализа
* @soap
*/
public function getAnalysisText($data)
{
}
}
<?
$exampleTable = [
    1 => [1, 6, 8, 4, 1, 2, 3, 5, 6, 0],
    2 => [20, 189, 7, 9, 5, 4, 12, 7, 66, -90],
    3 => [12, 7, -5, -8, 29, 99, 6, 2, 66, 43],
    4 => [2,4,5,6,7,3,4,5,6,100],
    5 => [0,-4,-7,-4,-1,-1,-1,5,6,7],
];
$client = new SoapClient('http://remote-
analiz.com/nofier/axd/?r=cluser/analysis');
echo $client->getAnalysisText($exampleTable);

```

Фрагмент программного кода сетевой ИС для выполнения кластерного анализа приведен в Приложении А.

2.4 Описание работы сетевой информационной системы для выполнения кластерного анализа

После авторизации на сайте сетевой ИС пользователь получает возможность активизировать процесс выполнения кластерного анализа по методу k-means в онлайн-режиме (рисунок 2.8).

Сетевое программное обеспечение для выполнения кластерного анализа (Малышев К.С.)

Алгоритм К-средних

Метод К-средних – это метод кластерного анализа, целью которого является разделение m наблюдений на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру (центроиду) которого оно ближе всего.

НАЗНАЧЕНИЕ С помощью онлайн-калькулятора можно проводить классификацию объектов **методом К-средних** с построением **дендрограммы** (например, для построения *типологической группировки*).

Решение онлайн **Видеоинструкция** Оформление Word

ИНСТРУКЦИЯ Укажите количество данных, нажмите **Далее**. Полученное решение сохраняется в файле **Word** (см. [пример решения](#)).

Размерность матрицы разбиения x

Далее **Из Excel**

Рисунок 2.8 – Окно сервиса кластерного анализа по методу k-means
Сервис позволяет задать (рисунок 2.9):

- размерность матрицы разбиения;
- количество кластеров.

Алгоритм К-средних

№	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	A	B	C	D	E	F
x_1	-2	-2	-3	4	3	0
x_2	0	-1	0	0	0	0

Количество кластеров $k = 3$

Далее

Рисунок 2.9 – Окно ввода данных для анализа

После ввода данных для активизации анализа нажимаем на кнопку «Далее».

Система в режиме онлайн выдает результаты анализа и строит таблицу кластеров на основе полученных данных (рисунок 2.10).

Сетевое программное обеспечение для выполнения кластерного анализа (Малышев К.С.)

Метод К-средних

$$d(Ee_3) = \sqrt{(3 - (-2.578))^2 + (0 - (-0.281))^2} = 5.585$$

Объект E ближе всех расположен к эталонной точке e_1 .

$$d(Fe_1) = \sqrt{(0 - 3.297)^2 + (0 - 0)^2} = 3.297$$

$$d(Fe_2) = \sqrt{(0 - (-0.25))^2 + (0 - (-0.0938))^2} = 0.267$$

$$d(Fe_3) = \sqrt{(0 - (-2.578))^2 + (0 - (-0.281))^2} = 2.593$$

Объект F ближе всех расположен к эталонной точке e_2 .

e_1	e_2	e_3
DE	F	ABC

Границы кластеров не изменились, т.е. в каждом кластере будет оставаться один и тот же набор записей. Останавливаем процесс кластеризации.

Решение было получено и оформлено с помощью сервиса:

Рисунок 2.10 – Представление результатов кластерного анализа

Предусмотрена возможность импорта данных для анализа из табличного процессора Excel.

Система позволяет экспортировать данные расчетов в файл формата MS Word.

Разработанная программа позволяет достаточно простыми средствами выполнить кластерный анализ в онлайн-режиме без привязки к конкретной области исследования.

ЗАКЛЮЧЕНИЕ

ВКР посвящена актуальной проблеме разработки программного обеспечения сетевого программного обеспечения для выполнения кластерного анализа.

В ходе выполнения ВКР достигнуты следующие результаты:

1) проанализированы методы и алгоритмы кластерного анализа. На основании результатов анализа в качестве алгоритма кластеризации для сетевой ИС выбран алгоритм k-means;

2) разработаны требования к программному обеспечению сетевой ИС для выполнения кластерного анализа;

3) на стадии логического моделирования в нотации языка UML построены диаграмма вариантов использования, диаграмма классов и диаграмма последовательности, отражающие функциональный, статический и динамический аспекты сетевой ИС соответственно;

4) разработана логическая модель сетевой ИС;

5) для реализации сетевой ИС выбраны трехуровневая архитектура «клиент-сервер» и средства разработки XAMPP;

6) разработана физическая модель сетевой ИС для СУБД MySQL;

7) в рамках сетевой ИС реализовано ПО для выполнения кластерного анализа.

ИС в настоящее время находится на стадии опытной эксплуатации. Предполагается доработка и развитие проекта.

Результаты ВКР могут быть рекомендованы для решения задач, связанных с кластерным анализом, в различных сферах деятельности.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

Нормативно-правовые акты

1. ГОСТ 34.320-96. Информационная технология. Система стандартов по базам данных. Концепции и терминология для концептуальной схемы и информационной базы.

Научная и методическая литература

2. Буренин С.Н. Web-программирование и базы данных: учебный практикум / Буренин С.Н. - М.: Московский гуманитарный университет, 2014. - 120 с.

3. Золотов С. Ю. Проектирование информационных систем : учеб. пособие / С. Ю. Золотов ; Томский гос. ун-т систем управления и радиоэлектроники. - Томск : Эль Контент, 2013. - 86 с.

4. Информационные аналитические системы : учебник / Т. В. Алексеева ; под ред. В. В. Дика. - Москва : Синергия, 2013. - 379 с.

5. Карпова И. П. Базы данных : курс лекций и материалы для практ. занятий : учеб. пособие для студентов техн. фак. / И. П. Карпова. – СПб. : Питер, 2013. - 240 с.

6. Крахоткина Е.В. Методы и средства проектирования информационных систем и технологий : учебное пособие / Крахоткина Е.В. - Ставрополь: Северо-Кавказский федеральный университет, 2015. -152 с.

7. Методы математического программирования в задачах оптимизации сложных технических систем: учебное пособие / А.М. Загребаев, Н.А. Крицына, Ю.П. Кулябичев, Ю.Ю. Шумилов. – М.: МИФИ, 2007. – 332 с.

8. Трутнев Д. Р. Архитектуры информационных систем. Основы проектирования: учеб. пособие. – СПб.: НИУ ИТМО, 2012. – 66 с.

Электронные ресурсы

9. Интеллектуальный дизайнер для MySQL dbForge Studio for MySQL. - URL : <https://www.devart.com/ru/dbforge/mysql/studio> (дата обращения 01.06.2018).

10. Компания StatSoft. - URL :<http://statsoft.ru/products>(дата обращения 01.06.2018).

11. Метод k-средних. – URL: https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_k-%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B8%D1%85(дата обращения 01.06.2018).

12. Платформа ClusterDELTA.- URL : <http://clusterdelta.com/platform> (дата обращения 01.06.2018).

13. Программа Visual Volume .- URL : <http://t-traders.com/other/forex-soft/2680-visual-volume-analiz-obema.html> (дата обращения 01.06.2018).

14. Сервис калькулятор кластерного анализа. -URL: <http://axd.semestr.ru/upr/cluster.php> (дата обращения 01.06.2018).

15. Среда разработки XAMPP. - URL:<https://www.apachefriends.org/ru/index.html> (дата обращения 01.06.2018).

16. Тестирование программного обеспечения. -URL: <http://www.protesting.ru> (дата обращения 01.06.2018).

17. CloudTech. -URL: <https://www.cloudcomputing-news.net/news/2014/feb/20/the-path-to-good-cloud-architecture-uses-soa/>(дата обращения 01.06.2018).

Литература на иностранном языке

18. Berkhin P. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, USA, 2002.

19. Cornish R. Statistics: Cluster Analysis, Mathematics Learning Support Centre, 2007.

20. Everitt B.S., Landau S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.

21. Models and Analysis in Distributed Systems / ed. by S. Haddad, F. Kordon, L. Pautet, L. Petrucci. – London-Hoboken: Wiley-ISTE, 2011. -368 p.

22. Nixon R. Learning PHP, MySQL, JavaScript, CSS & HTML5 - 3rd Edition, 2014.

ПРИЛОЖЕНИЕ А

Фрагменты программного кода приложения

```
<?
class RequestRows extends CActiveRecord
{
    /**
     * @var array Список строк
     * @soap
     */
    public $rows;

    public static function model($className=__CLASS__)
    {
        return parent::model($className);
    }
}

class RequestRow extends CActiveRecord
{
    /**
     * @var array Список ячеек
     * @soap
     */
    public $cells;

    public static function model($className=__CLASS__)
    {
        return parent::model($className);
    }

    public static function getSum()
    {
```

```

    $sum = 0;
    foreach ($this->cells as $cell) {
        $sum += $cell->getSum();
    }
    return $sum;
}

public static function getMedian()
{
    $median = 1;
    // ...
    return $median;
}

public static function setCells($cells)
{
    $this->cells = [];
    foreach ($cells as $key => $value) {
        $cell = new RequestCell($value);
        $cell->parent = $this;
        $this->cells[] = $cell;
    }
}

}

class RequestCell extends CActiveRecord
{
    /**
     * @var string Значение
     * @soap
     */

```

```

public $value;

public $parent;

public static function model($className=__CLASS__)
{
    return parent::model($className);
}

public static function getSum()
{
    $sum = $this->getFirstAnalyze() * $this->getMedian();
    return $sum;
}

body > .row > .col-sm-3 {
    opacity: 0;
    width: 23%;
}

body > .row > .col-sm-3:after {
    display: block;
    content: " ";
    width: 100%;
    height: 100%;
    background: red;
    position: absolute;
    top: 0;
    left: 0;
}

body > .row {
    margin: 0;

```

```
}
```

```
body > div:nth-child(2),  
body > .row > .col-sm-6 > .uplink,  
body > .row > .col-sm-6 > .btn-group,  
body > .row > .col-sm-6 > .adsbygoogle,  
#maintop,  
body > .footer,  
#MathLeftBlock,  
#MathRightBlock,  
body > .row > .col-sm-3 > div:nth-child(6),  
#soloway240,  
#MathLeftSpan,  
#MathTopAdBlock,  
#mainleft,  
#forma1,  
#forma2,  
#forma3,  
#disqus_thread {  
    display:none!important;  
}
```

```
.ref_cat h1:before{  
    content: 'Сетевое программное обеспечение для выполнения  
кластерного анализа (Мальшев К.С.)';  
    display: block;  
    padding-bottom: 10px;  
    font-size: 20px;  
}
```

```
#premain {
    position: relative;
}
/*#premain a {
    display: none;
}*/
/*#premain a:last-child {
    display: block;
}*/
/*#premain a:last-child:after {
    display: block;
    content: " ";
    width: 100%;
    height: 195px;
    background: white;
    margin-top: -195px;
    z-index: 5;
    position: relative;
}*/
#balance1 {
    display: none;
}
#premain {
    position: relative;
}

#premain:after {
    display: block;
    content: " ";
    width: 100%;
```

```
    height: 195px;
    background: white;
    margin-top: -195px;
    z-index: 5;
    position: relative;
}

/*@media screen and (min-width: 1377px) {
    #premain:after {
        height: 170px;
        margin-top: -170px;
    }
}*/
```