Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа информационных технологий и робототехники
Направление подготовки 09.04.01 «Информатика и вычислительная техника»
Отделение школы (НОЦ) информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Методика и программное обеспечение для обработки данных ранжирования узлов социального графа для идентификации пользователей-экспертов

УДК 004.42:004.6-048.57:004.056.523

Студент

Группа	ФИО	Подпись	Дата
8ВМ71	Виноградова Дарья Андреевна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Лунева Е. Е.	к.т.н.		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Меньшикова Е.В.	к.ф.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент	Алексеев Н.А.			

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Кочегурова Е.А.	к.т.н.		

Томск 2019

Планируемые результаты обучения

Код результата	Результат обучения (выпускник должен быть готов)
	Общепрофессиональные компетенции
Р1	Воспринимать и самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
Р2	Владеть и применять методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях.
Р3	Демонстрировать культуру мышления, способность выстраивать логику рассуждений и высказываний, основанных на интерпретации данных, интегрированных из разных областей науки и техники, выносить суждения на основании неполных данных, анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.
Р4	Анализировать и оценивать уровни своих компетенций в сочетании со способностью и готовностью к саморегулированию дальнейшего образования и профессиональной мобильности. Владеть, по крайней мере, одним из иностранных языков на уровне социального и профессионального общения, применять специальную лексику и профессиональную терминологию языка.
	Профессиональные компетенции
Р5	Выполнять инновационные инженерные проекты по разработке аппаратных и программных средств автоматизированных систем различного назначения с использованием современных методов проектирования, систем автоматизированного проектирования, передового опыта разработки конкурентно способных изделий.
Р6	Планировать и проводить теоретические и экспериментальные исследования в области проектирования аппаратных и программных средств автоматизированных систем с использованием новейших достижений науки и техники, передового отечественного и зарубежного опыта. Критически оценивать полученные данные и делать выводы.
Р7	Осуществлять авторское сопровождение процессов проектирования, внедрения и эксплуатации аппаратных и программных средств автоматизированных систем различного назначения.
	Общекультурные компетенции
Р8	Использовать на практике умения и навыки в организации исследовательских, проектных работ и профессиональной эксплуатации современного оборудования и приборов, в управлении коллективом.
Р9	Осуществлять коммуникации в профессиональной среде и в обществе в целом, активно владеть иностранным языком, разрабатывать документацию, презентовать и защищать результаты инновационной инженерной деятельности, в том числе на иностранном языке.
Р10	Совершенствовать и развивать свой интеллектуальный и общекультурный уровень. Проявлять инициативу, в том числе в ситуациях риска, брать на себя всю полноту ответственности.
Р11	Демонстрировать способность к самостоятельному обучению новым методам исследования, к изменению научного и научно-производственного профиля своей профессиональной деятельности, способность самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности, способность к педагогической деятельности.

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа информационных технологий и робототехники
Направление подготовки 09.04.01 «Информатика и вычислительная техника»
Уровень образования магистратура
Отделение школы (НОЦ) информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8ВМ71	Виноградова Дарья Андреевна

Тема работы:

Методика и программное обеспечение для обработки данных ранжирования узлов социального графа для идентификации пользователей-экспертов	
Утверждена приказом директора (дата, номер)	Приказ №1325/с от 20.02.2019г.

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	1. Показатели авторитетности пользователей социальной сети, связанных с определенной предметной областью: показатель Боргатти, информационной энтропии, и коммуникационной эффективности. 2. Требование разработать методику обработки данных ранжирования узлов социального графа для решения задачи идентификации пользователей-экспертов социальной сети. 3. Требование разработать ПО для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. Анализ возможности применения методов машинного обучения к задаче идентификации пользователей-экспертов социальной сети в заданной предметной области. 2. Разработка методики обработки данных ранжирования узлов социального графа для решения задачи идентификации пользователей-экспертов социальной сети 3. Проектирование и разработка ПО для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети. 4. Результаты работы разработанного ПО
Перечень графического материала	Презентация в формате *.pptx.
Консультанты по разделам выпускной квалификационной работы	
Раздел	**Консультант**
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Меньшикова Е.В.
Социальная ответственность	Алексеев Н.А.
Названия разделов, которые должны быть написаны на русском и иностранном языках:	
Ведение	
1Анализ возможности применения методов машинного обучения к задаче идентификации пользователей-экспертов социальной сети в заданной предметной области	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Лунева Е.Е.	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ71	Виноградова Дарья Андреевна		

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа информационных технологий и робототехники
Направление подготовки 09.04.01 «Информатика и вычислительная техника»
Уровень образования магистратура
Отделение школы (НОЦ) информационных технологий
Период выполнения (осенний / весенний семестр 2018/2019 учебного года)

Форма представления работы:

магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.09.2018	Получение задания	0
12.02.2019	Исследование темы, выбор методов и средств разработки	30
15.05.2019	Разработка программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети	30
05.06.2019	Оформление пояснительной записки	20
05.06.2019	Социальная ответственность	10
05.06.2019	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Лунева Е.Е.	к.т.н.		

СОГЛАСОВАНО:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Кочегурова Е.А.	к.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ВМ71	Виноградова Дарья Андреевна

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	магистратура	Направление/специальность	09.04.01 Информатика и вычислительная техника

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	1. Оклад доцента – 33664₽ 2. Оклад инженера – 12663₽ 3. Стоимость услуг Internet– 12₽/дн. 4. Стоимость оборудования – 40 000₽
2. Нормы и нормативы расходования ресурсов	5. 6-часовой рабочий день 6. 6-ти дневная рабочая неделя
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	7. Единый социальный налог – 27,1%
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Оценка коммерческого и инновационного потенциала НТИ	8. Потенциальные потребители результатов исследования 9. SWOT-анализ
2. Разработка устава научно-технического проекта	10. Цели и результат проекта 11. Организационная структура проекта 12. Ограничения и допущения проекта
3. Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок	13. План проекта 14. Бюджет научного исследования 15. Реестр рисков проекта
Перечень графического материала	
1. Матрица SWOT 2. Диаграмма Ганта 3. Потенциальные риски	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Меньшикова Е.В.	к.ф.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ71	Виноградова Дарья Андреевна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ВМ71	Виноградова Дарья Андреевна

Школа	ИШИТР	Отделение (НОЦ)	Информационных технологий
Уровень образования	Магистратура	Направление/специальность	09.04.01 Информатика и вычислительная техника

Исходные данные к разделу «Социальная ответственность»:	
Характеристика объекта исследования (вещество, материал, рабочая зона) и области его применения	Система видеомониторинга и идентификации обучающегося в дистанционном обучении
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности	1. Специальные правовые нормы трудового законодательства; 2. Организационные мероприятия при компоновке рабочей зоны.
2. Производственная безопасность 2.1. Анализ выявленных вредных факторов при разработке и эксплуатации проектируемого решения 2.2. Анализ выявленных опасных факторов при разработке и эксплуатации проектируемого решения	**Вредные факторы:** 1. Отклонения показателей микроклимата; 2. Недостаточная освещенность рабочей зоны; 3. Превышение уровня шума; 4. Повышенный уровень электромагнитных излучений; **Опасные факторы:** 1. Электрический ток 2. Опасность возникновения пожара
3. Экологическая безопасность	Источники выбросов в атмосферу; Образование сточных вод и отходов. Мероприятия по снижению вредного воздействия на ОС
4. Безопасность в чрезвычайных ситуациях	Вероятные ЧС при разработке и эксплуатации проектируемого решения и меры по их предупреждению

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ООД ШБИП	Алексеев Н.А.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ71	Виноградова Дарья Андреевна		

Реферат

Выпускная квалификационная работа 149 с., 14 рис., 24 табл., 74 источников, 8 прил.

Ключевые слова: кластеризация, DBScan, Fast Affinity Propagation, классификация, дерево решений, силуэт, V-мера.

Объектом исследования является: задача идентификации пользователей-экспертов социальной сети в заданной предметной области.

Цель работы – разработка программного обеспечения для обработки данных ранжирования узлов социального графа для идентификации пользователей-экспертов социальной сети в заданной предметной области.

В результате данной работы разработано программное обеспечение обработки данных ранжирования узлов социального графа для идентификации пользователей-экспертов социальной сети в заданной предметной области, а также создана методика, позволяющей обработать и интерпретировать данные ранжирования узлов социального графа при помощи нескольких эффективных способов, которая и легла в основу приложения.

Область применения: идентификация пользователей-экспертов социальной сети в заданной предметной области.

В будущем планируется усовершенствование используемого в работе программного обеспечения классификатора, путем повышения качества его обучения.

Определения, обозначения, сокращения

Социальный граф – взвешенный ориентированный граф, построенный на основе данных, выбранных из социальных сетей по заданной предметной области.

Задач KPP-POS (Key player problem) – задача идентификации узла (или группы узлов) социального графа, максимально распространяющего информацию по сети.

Машинное обучение - обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться

Кластеризация, кластерный анализ – задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов.

Классификация – задача распределения объектов по заранее выделенным классом.

ГОСТ - Межгосударственный стандарт.

ПЭВМ – персональная электронно-вычислительная машина.

СанПиН - Санитарные нормы и правила.

Оглавление

Введение

Современный мир невозможно представить без социальных сетей. Высокий уровень вовлеченности современного человека в процесс виртуального общения приводит к огромному скоплению различных данных в социальных сетях, анализ которых может выявить ценную информацию разного рода.

Информация, полученная из социальных сетей, может быть определяющей при решении задач антитеррористической направленности, политической аналитики, прогнозирования репутационных рисков компании или субъекта, оценки спроса на товар или услугу, а также мониторинга общественного мнения.

Социальные сети оказывают значительное влияние на мнение людей. Пользователи социальных сетей часто выражают собственное мнение по актуальным вопросам, событиям, товаром или услугам в интересующих их темах. Однако, можно заметить, что мнение некоторых пользователей в определенных тематиках оказывает высокое влияние на мнение других пользователей.

Информация о влиятельных пользователях (далее пользователи-эксперты) в определенной сфере может быть полезна при решении различных практических задач. Пример решения маркетинговой задачи: компания производства косметики может получить качественную рекламу у пользователя-эксперта социальной сети в сфере красоты.

В настоящее время нельзя определенно назвать единственный лучший метод поиска пользователей-экспертов социальной сети в заданной тематики. Данная работа направлена на создание методики, позволяющей обработать и интерпретировать данные ранжирования узлов социального графа при помощи нескольких эффективных способов. Практическая часть работы связанна с разработкой соответствующей алгоритмической базы и программного обеспечения для идентификации пользователей-экспертов.

1 Анализ возможности применения методов машинного обучения к задаче идентификации пользователей-экспертов социальной сети в заданной предметной области

Прежде, чем приступить к созданию методики, позволяющей обработать и интерпретировать данные ранжирования узлов социального графа для идентификации пользователей-экспертов, или же к проектированию и разработке программного обеспечения для выделения пользователей-экспертов, необходимо провести обзор и анализ литературы и других источников информации по теме идентификации пользователей-экспертов в социальной сети.

В данной главе составлен обзор задачи идентификации пользователей-экспертов в социальной сети, рассмотрена возможность применения машинных методов обучения для ее решения, а также определены цели и задачи данной диссертации.

1.1 Обзор задачи идентификации пользователей-экспертов в социальной сети

Множество литературных источников посвящено проблеме идентификации значимых узлов в социальных сетях [1-3]. Для обозначения таких узлов используются разнообразные термины: влиятельный узел, ключевой игрок, критический узел, жизненно важный узел и т.д. [3].

В работе [1] данная проблема была обозначена как проблема поиска ключевых игроков (KPP, Key player problem) и разделена на два класса: Positive (POS) и Negative (NEG).

Класс KPP-NEG направлен на поиск такого узла (или группы узлов) социального графа, удаление которого сделает граф максимально фрагментированным. Подходы к решению данной задачи можно применять для решения задач в области медицины; например, при эпидемиях это задача идентификации группы риска, вакцинация которой максимально сократит распространение заболевания. Также, данные подходы используются для

задач антитеррористической направленности, в частности, при выявлении участников, исключение которых из террористической группы максимально разобщит оставшихся членов. Позже данному классу задач был присвоен термин CNP (Critical Node Problem) [2] и дано аналитическое описание.

Второй класс задач KPP-POS направлен на идентификацию узла (или группы узлов), максимально распространяющего информацию по сети. Решение данной проблемы необходимо, например, для решения задачи внедрения в террористическую сеть информаторов, за счет которых неверная информация максимально быстро распространится по сети [1].

Пользователи социальных сетей, являющиеся экспертами в некоторой предметной области, задают темы для обсуждения, выказывают свое мнение относительно определенных вопросов по заданной тематики и актуальных событий. При этом, к их мнению прислушиваются другие пользователи и часто меняют свои взгляды относительно оглашенных тем в обсуждаемой предметной области. Таким образом, мнения пользователей-экспертов распространяются по социальной сети за счет различных действий читателей, таких как комментирование, репост, упоминание, «лайк». Значит задача идентификации пользователей-экспертов в социальных сетях имеет некоторые отличия от задачи CNP и, очевидно, относится к классу задач KPP-POS.

Анализ литературы [1, 4-6] позволил выявить следующие способы решения задачи KPP-POS, считающиеся наиболее эффективными:

- способ, основанный на вычислении показателя Боргатти;
- расчет информационной энтропии вершин графа;
- расчет коммуникационной эффективности.

Однако, применение предложенных способов решения проблемы KPP-POS к задаче определения пользователей социальных сетей, являющихся лидерами общественного мнения или рассматриваемых в качестве экспертов по заданной тематике, связано со следующими трудностями:

- В литературе решение данной проблемы апробируется на неориентированных и/или невзвешенных графах [1, 4, 7-9]. Использование таких графов значительно сокращает их способность адекватно представлять реальную группу пользователей социальных сетей, поскольку не позволяет учитывать дополнительную информацию об отношениях между пользователями [10, 11].
- Существует сложность в выборе оптимального способа решения данной задачи, т.к. оценка эффективности решений затруднена отсутствием фактических данных, а также может зависеть от конкретных характеристик анализируемых социальных графов.

Таким образом, для получения качественного результата решения задачи выявления пользователей-экспертов в социальных сетях необходимо использовать все вышеперечисленные способы решения задачи KPP POS. После чего полученные разрозненные результаты по каждому способу необходимо объединить и интерпретировать для получения единого решения.

1.2 Обзор эффективных способов решения задачи KPP-POS применительно к задаче идентификации пользователей-экспертов социальной сети

Как было сказано ранее, наиболее эффективными способами решения задачи KPP-POS считаются следующие способы:

- способ, основанный на вычислении показателя Боргатти;
- расчет информационной энтропии вершин графа;
- расчет коммуникационной эффективности.

Каждый из перечисленных способов в качестве входных данных использует так называемый социальный граф.

1.2.1 Методика построения социального графа при решении задачи идентификации пользователей-экспертов

Входными данными для поиска пользователей-экспертов в социальных сетях по заданной предметной области служит количественная информация, отражающая заинтересованность пользователей социальных сетей мнением экспертов. К такой информации можно отнести количество репостов – *r*, комментариев – *c*, упоминаний пользователя-эксперта – *m*, репостов с комментариями – *x*. Очевидно, заинтересованность пользователя *j* публикациями пользователя *i* можно выразить в виде функции $f\left(r_{ij}, c_{ij}, m_{ij}, x_{ij}\right)$. Аналитическое описание данной функции предложено в работе [12].

Социальный граф определяется, как взвешенный ориентированный граф [10]:

$$G = (V, A, w) \tag{1}$$

Где

$V = \{v_i\}$ – множество узлов графа ($i = 1, 2, \ldots, N$),

$A = \{(v_i, v_j)\} \subset V \times V$ – множество направленных ребер (дуг),

$w : A \to (0,1]$ – весовая функция, ставящая в соответствие каждой дуге $a_{ij} = (v_i, v_j) \in A$ некоторое значение $w(a_{ij}) \in (0,1]$.

Тогда на основе данных, выбранных из социальных сетей по заданной предметной области, можно построить социальный граф, узлы которого связаны ребрами по следующему принципу: существует направленное ребро (дуга) a_{ij}, связывающее узлы *i* и *j*, если *j*-й пользователь комментировал публикации *i*-го пользователя, делал их репост или репост с комментарием, упоминал в своих комментариях пользователя *i*. При этом $w(a_{ij}) = f\left(r_{ij}, c_{ij}, m_{ij}, x_{ij}\right)$ [10]

1.2.2 Обзор выделенных способов решения задачи KPP-POS

1.2.2.1 Способ, основанный на вычислении показателя Боргатти

Принципиальное отличие способа, предложенного Боргатти в работе [1], от других выделенных способов, состоит в том, что автор предлагает заранее задавать количество ключевых игроков, которое следует найти в социальном графе. Для этого из множества узлов V выделяется подмножество $K \subset V$ потенциальных ключевых игроков с требуемым количеством элементов и определяется показатель C_K по следующей формуле [1]:

$$C_K = \sum_{v_j \in V \setminus K} \max_{v_i \in K} \left(\frac{1}{d_{ij}} \right), \quad (2)$$

Где

K – множество, в которое попала некоторая проверяемая комбинация узлов,

V – множество всех узлов графа,

d_{ij} – кратчайшее расстояние от узла v_i до узла v_j, при этом из множества K выбирается тот узел, от которого кратчайшее расстояние до узла v_j минимально.

Рассчитав значения C_K для всех возможных множеств K, можно выбрать комбинацию узлов, составляющую искомое множество ключевых игроков: чем больше значение C_K для проверяемой комбинации узлов, тем большее влияние пользователи из этой группы могут оказывать на своих читателей. Для корректного применения способа Боргатти в качестве веса дуги между узлами социального графа используется значение, вычисленное по формуле [1]:

$$w(a_{ij}) = \frac{1}{f(r_{ij}, c_{ij}, m_{ij}, x_{ij})}. \quad (3)$$

1.2.2.2 Расчет информационной энтропии вершин графа

Способ, основанный на расчете показателя информационной энтропии, предполагает последовательное исключение из социального графа узлов и

инцидентных им дуг. При этом, энтропия социального графа вычисляется по следующей формуле [4]:

$$H(G,P)=\sum_{i=1}^{|V|} p_i \cdot \log_2\left(\frac{1}{p_i}\right) \qquad (4),$$

Где

$V=\{v_i\}$ – множество узлов графа *G*,

$P=\{p_i\}$ – плотность распределения вероятности на множестве V, задающая вероятность того, что узел v_i представляет самого значимого пользователя-эксперта.

Значение p_i зависит от заинтересованности пользователей социальной сети сообщениями *i*-го пользователя, выраженной как отношение количества репостов, комментариев и т.п., порожденных сообщениями *i*-го пользователя ко всему объему сообщений социальной сети. Исходя из этого, значения $p_i,\ i\in[1,N]$ можно вычислить следующим образом [4]:

$$p_i=\frac{\sum_{v_j\in V^{[i]}} w_{ij}}{\sum_{v_k\in V}\sum_{v_l\in V^{[k]}} w_{kl}}, \qquad (5)$$

где

$V^{[s]}$ – множество конечных узлов дуг, исходящих из узла v_s, $w_{yz}=f(r_{yz},c_{yz},m_{yz},x_{yz})$ – вес дуги a_{yz}.

1.2.2.3 Расчет коммуникационной эффективности

В теории графов известна задача ранжирования игроков по итогам кругового турнира [13, 14, 15]. Для решения этой задачи составляется полный ориентированный граф без петель с числом вершин *n* равным количеству игроков. При этом любые две вершины соединены дугой, и дуга из *i*-й вершины в *j*-ю означает победу *i*-го игрока над *j*-м. В матрице смежности A

турнира элемент $a_{ij} = 1$, если существует дуга из *i*-й вершины в *j*-ю; иначе, $a_{ij} = 0$.

Сумма элементов *i*-й строки $s_i^{\langle 1 \rangle} = \sum_{j=1}^{n} a_{ij}$ соответствует количеству побед, одержанных *i*-м игроком в ходе турнира, и является мерой силы *i*-го игрока. Вектор $\mathbf{S}^{\langle 1 \rangle}$, составленный из элементов $s_i^{\langle 1 \rangle}$, называется вектором сил первого порядка [16].

В тривиальном случае все элементы вектора $\mathbf{S}^{\langle 1 \rangle}$ различны, и ранжирование игроков не представляет сложности. Однако, если по итогам турнира несколько игроков одержали одинаковое количество побед (набрали одинаковое количество очков), однозначное ранжирование игроков может быть затруднительным [13, 14].

Процедура РКУ предполагает вычисление итерированных сил *k*-го порядка игроков по одной из следующих формул [16]:

$$\mathbf{S}^{\langle k \rangle} = \mathbf{A} \cdot \mathbf{S}^{\langle k-1 \rangle} \tag{6}$$

$$\mathbf{S}^{\langle k \rangle} = \mathbf{A}^k \cdot \mathbf{S}^{\langle 1 \rangle} \tag{7}$$

Где

$k \geq 2$ - номер итерации;

$\mathbf{A}$ - матрица смежности;

$\mathbf{S}^{\langle j \rangle}$ - вектор итерированных сил *j*-го порядка.

1.3 Обзор и анализ машинных методов обучения применительно к задаче идентификации пользователей-экспертов социальной сети.

Для выявления пользователей-экспертов в определенной предметной области необходимо объединить и интерпретировать данные решения задачи KPP POS каждого пользователя социальной сети, кто связан тем или иным образом с заданной предметной областью. Объем пользовательской базы в социальных сетях на сей день очень велик, а значит данные решения задачи

КРР POS всех пользователей, объединенных интересом к одной предметной области, также будут не малочислены, что очень осложняет процесс их анализа. Сегодня для анализа большого скопления данных часто используют машинное обучение (Machine Learning).

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться [17]. Целю машинного обучения является предсказание результатов по входным данным.

Проанализировав источники [17-19] можно выделить два основных вида машинного обучения:

- Обучение с учителем –способ машинного обучения, когда машина принудительно обучается на представляемых ей примерах. Примеры содержат в себе входы (входные данные объектов) и эталонные выходы (ожидаемые результаты). Машине нужно установить между входами и эталонными выходами некоторую зависимость. Далее машина, используя выведенную зависимость, будет подбирать выходные значения для подаваемых ей новых входов [20, 21].
- Обучение без учителя –способ машинного обучения, когда машина самостоятельно обучается выполнять поставленную задачу. Обычно, данный способ пригоден для решения задачи, когда известны только входы, и требуется установить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами [22, 23].

Все остальные виды машинного обучения, представленные в [17-19] представляют из себя либо частные случаи выделенных видов машинного обучения, либо их смешивание.

В [19] выделены классические задачи, решаемые машинным обучением.

Классические задачи, решаемые обучением с учителем:

- Классификация – задача распределения объектов по заранее выделенным классом.
- Регрессия – задача прогнозирования нужного значения.

Классические задачи, решаемые обучением без учителя:

- Кластеризация – задача разделения объектов на заранее не известные группы.
- Обобщение – задача понижения размерности данных.
- Ассоциация – задача выявления последовательностей, построение зависимостей объектов друг от друга.

Задача идентификации пользователей-экспертов предполагает, что всех пользователей социальной сети, связанных с заданной предметной областью, нужно разделить на две заранее известные группы: на писателей (непосредственно сами пользователи-эксперты) и на читателей (остальные пользователи). Таким образом задача идентификации пользователей-экспертов относится к задачи классификации, которая решается обучение с учителем. Однако, классификация данных требует обучающую выборку (примеры входов с эталонными выходами для них), создание которой заберет много сил и времени.

Для решения задачи идентификации пользователей-экспертов узлы социального графа могут быть ранжированы не только при помощи классификации. Если для решения данной задачи воспользоваться кластеризацией, решаемая обучением без учителя, между объектами будут установлены внутренние взаимосвязи, зависимости и закономерности, на основании которых рассматриваемые пользователи будут разделены на более частные группы. Использование данного способа позволит получить дополнительные сведения о группе читателей, находящихся под влиянием пользователей-экспертов, задающих общественное мнение, а также информацию о потенциальных скрытых лидерах-экспертах.

Таким образом, чтобы максимально полезно решить задачу идентификации пользователей-экспертов в социальной сети, необходимо для ее решения использовать кластерный анализ.

1.4 Цель работы и задачи

Приведенный обзор и анализ литературы и других источников информации по теме идентификации пользователей-экспертов в социальной сети показал, что для решения заданной задачи необходимо выполнить следующие шаги:

- Для каждого пользователя социальной сети, связанного с определенной предметной областью, получить данные, предназначенные для решения задачи KPP POS: показатель Боргатти, показатель информационной энтропии и показатель коммуникационной эффективности.
- Произвести ранжирование полученных показателей с использованием кластерного анализа.

Целью данной работы является создание методики, позволяющей обработать и интерпретировать данные ранжирования узлов социального графа программным путем с целью идентификации пользователей-экспертов.

Для достижения поставленной цели необходимо решить следующие задачи:

- Разработать методику обработки данных ранжирования узлов социального графа для задачи идентификации пользователей-экспертов социальной сети.
- Спроектировать программное обеспечение для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети по разработанной методики.
- Реализовать спроектированное программное обеспечение.

2 Разработка методики обработки данных ранжирования узлов социального графа для задачи идентификации пользователей-экспертов социальной сети

2.1 Исследование методов кластерного анализа применительно к задаче идентификации пользователей-экспертов социальной сети

Кластерный анализ – задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов. Сходство объектов определяется на основе метрики, выбранной в соответствии с критерием кластеризации [24].

Таким образом, кластерный анализ представляет собой многомерную статистическую процедуру, выполняющую сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы [25].

На данный момент существуют сотни алгоритмов кластерного анализа данных. Можно выделить алгоритмы кластерного анализа, для которых число кластеров задано, а также алгоритмы, в которых число кластеров определяется в ходе решения кластеризации [26].

Если к задачи идентификации пользователей-экспертов в социальной сети применять алгоритмы кластеризации, для которых число кластеров нужно задавать, то результатом кластеризации будут два кластера: один с читателями, другой с писателями. Так, используя кластеризацию для интерпретации данных ранжирования узлов социального графа, никакой дополнительной информации ни о читателях, ни о потенциальных писателях, получено не будет. Следовательно, для решения поставленной задачи необходимо применение алгоритмов кластерного анализа, в которых число кластеров определяется в течение выполнения кластеризации.

Далее рассмотрены некоторые алгоритмы кластерного анализа, не требующие явного задания количества кластеров.

2.1.1 Обзор некоторых существующих алгоритмов кластерного анализа, не требующих явного задания количества кластеров

2.1.1.1 FOREL

FOREL (Формальный Элемент) – алгоритм кластеризации, основанный на идее объединения в один кластер объектов в областях их наибольшего сгущения [27].

На каждом шаге случайным образом выбирается объект из выборки, раздувается вокруг него сфера радиуса R, внутри этой сферы рассчитывается центр тяжести. Полученный центр тяжести становится центром новой сферы. То есть на каждом шаге сфера фиксированного радиуса двигается в сторону локального сгущения объектов выборки, то есть старается захватить как можно больше объектов выборки. После того как центр сферы стабилизируется, все объекты внутри сферы с этим центром помечаются как кластеризованные и выкидываются из выборки. Этот процесс повторяется до тех пор, пока вся выборка не будет кластеризована [28].

Алгоритм

На вход задается множество объектов X, для которых задана метрическая функция расстояния ρ и параметр R.

1. Случайно выбираем объект $\boldsymbol{x_0}$ из выборки некластеризованных объектов $\boldsymbol{X^l}$ и помечаем его.
2. Помечаем объекты $\boldsymbol{x_i}$ выборки некластеризованных объектов $\boldsymbol{X^l}$, находящиеся на расстоянии менее, чем $\boldsymbol{R}$ от текущего $\boldsymbol{x_0}$, или по другому все объекты $\boldsymbol{x_i}$ для которых справедливо неравенство $\boldsymbol{\rho(x_i, x_0) \geq R}$;
3. Вычисляем центр тяжести всех помеченных объектов, помечаем этот центр как новый текущий объект $\boldsymbol{x_0}$;
4. Повторяем шаги 2-3, пока новый текущий объект не совпадет с прежним;

5. Помечаем объекты внутри сферы радиуса R вокруг текущего объекта как кластеризованные, выкидываем их из выборки;
6. Повторяем шаги 1-5, пока не будет кластеризована вся выборка, то есть пока $\boldsymbol{X^l} = \emptyset$.

Для описания алгоритма FOREL использовались источники [28] и [29].

2.1.1.2 Алгоритм выделения связных компонент.

Алгоритм выделения связных компонент основан на представлении выборки в виде графа. Вершинам графа соответствуют объекты выборки, а рёбрам – попарные расстояния между объектами $\rho_{ij} = \rho(x_i, x_j)$, которое определяется по выбранной метрике [29].

В алгоритме выделения связных компонент задается входной параметр R и в графе удаляются все ребра, для которых «расстояния» больше R. Соединенными остаются только наиболее близкие пары объектов. Смысл алгоритма заключается в том, чтобы подобрать такое значение R, лежащее в диапазон всех «расстояний», при котором граф «развалится» на несколько связных компонент. Полученные компоненты и есть кластеры [30].

Для подбора параметра R обычно строится гистограмма распределений попарных расстояний. В задачах с хорошо выраженной кластерной структурой данных на гистограмме будет два пика – один соответствует внутрикластерным расстояниям, второй – межкластерным расстояния. Параметр R подбирается из зоны минимума между этими пиками [31].

Алгоритм выделения связных компонент имеет свои недостатки.

Алгоритм выделения связных компонент наиболее подходит для выделения кластеров типа сгущений или лент. Наличие разреженного фона или «узких перемычек» между кластерами приводит к неадекватной кластеризации [30].

2.1.1.3 Алгоритм минимального покрывающего дерева

Как и алгоритм выделения связных компонент, алгоритм минимального покрывающего дерева основан на представлении выборки в виде графа. Каждый объект в заданном множестве представляется вершиной в графе, а ребро между двумя вершинами представляет собой расстояние между объектами $\rho_{ij} = \rho(x_i, x_j)$, которое определяется по выбранной метрике [32].

Основная идея алгоритма заключается в том, чтобы на полученном графе построить минимальное покрывающее дерево, а после этого удалить из него все ребра, вес которых больше заданного параметра R . Чем больше R, тем меньше в результате получится кластеров, а также, тем меньше внутри кластеров объекты будут похожи друг на друга.

Покрывающее дерево – ациклический связный подграф данного связного неориентированного графа, в который входят все его вершины [33]. Минимальное покрывающее дерево в связанном взвешенном неориентированном графе – это покрывающее дерево этого графа, имеющее минимальный возможный вес, где под весом дерева понимается сумма весов входящих в него рёбер [33].

Для построение минимального покрывающего дерева можно использовать такие алгоритмы как:

- алгоритм Краскала;
- алгоритм Прима;
- алгоритм Борувки.

2.1.1.4 Алгоритмы семейства Affinity propagation

2.1.1.4.1 Affinity propagation

Affinity propagation – это алгоритм кластеризации, основанный на концепции «передачи сообщений» между точками данных [34]. Affinity propagation находит «образцы» среди точек данных и формирует кластеры вокруг них. [35].

Под «образцом» понимается точка данных, которая является репрезентативным представителем определенного кластера, то есть точка данных, которая хорошо отражает себя и другие точки данных, входящие в один кластер и является образцовым экземпляром этого кластера.

Affinity propagation одновременно рассматривает все точки данных как потенциальных образцы. Точки данных обмениваются сообщениями между собой до тех пор, пока постепенно не появится высококачественный набор образцов, обросших похожих на них самих точек данных, другими словами набор кластеров [35].

На выбор образцов влияют три параметра: сходство, ответственность и доступность.

Сходство (similarity) определяет сходство между любыми двумя точками. При этом должна существовать функция $\mathrm{s}\,(i,k)$, определяющая сходство двух точек количественно. Сходство представляет из себя таблицу S, где записаны все $\mathrm{s}\,(i,k)$. При этом диагональ этой таблицы, то есть сходство точки с собой все $\mathrm{s}\,(i,i)$ особенно важна, поскольку оно представляет собой собственное предпочтение, что определяет, насколько вероятно для нее стать образцом. Медианное значение всех $\mathrm{s}\,(i,i)$ определяет количество классов, генерированных алгоритмом: значение чем меньше значение, тем меньше классов получится.

Ответственность (responsibility) определяет, насколько точка хочет видеть своим образцом другую точку. Ответственность определяется

функцией $r(i,k)$. Ответственность возлагается каждой точкой на каждого кандидата в образцы кластера.

Доступность (availability) определяет, насколько кандидат в образцы готов стать образцом для другой точки. Доступность определяется функцией $\mathrm{a}(i,k)$. Доступность возлагается от кандидатов в образцы на каждую точку.

Ответственность и доступность точки вычисляют в том числе и сами для себя. Именно по наибольшей сумме самоответственности и самодоступности и определяют образцы. Рядовые точки в конечном итоге присоединяются к образцу, для которого у них наибольшая сумма $a(i,k) + r(i,k)$.

Алгоритм

Пусть $\{x_1, \ldots x_n\}$ - множество точек данных без внутренней структуры;

s - функция, которая количественно определяет сходство между любыми двумя точками, так что выражение $\mathrm{s}\,(i,j) > \mathrm{s}\,(i,k)$ справедливо тогда и только тогда, когда x_iбольше похож на x_j, чем на x_k.

Алгоритм выполняется путем чередования двух шагов передачи сообщения, для обновления двух матриц R и A:

Обе матрицы инициализируются ко всем нулям и могут рассматриваться как лог-вероятностные таблицы.

$$R, A \leftarrow 0, 0$$

Затем алгоритм выполняет следующие обновления:

- Сообщения об ответственности отправляются [35]:

$$r(i,k) \leftarrow s(i,k) - \leftarrow \max_{k' \neq k}\{a(i,k') + s(i,k')\} \quad (8)$$

- Доступность обновляется следующим образом [35]:

$$a(i,k) \leftarrow \min\big(0, r(k,k) + \textstyle\sum_{i' \notin \{i,k\}} \max(0, r(i',k))\big)\ for\ i \neq k\ and \quad (9)$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max\big(0, r(i',k)\big).$$

Итерации выполняются до тех пор, пока границы кластера не останутся неизменными по целому ряду итераций или после некоторого заданного

количества итераций. Образцами из окончательных матриц считаются те, чья «ответственность + доступность» для себя положительна, то есть $r(i,i) + a(i,i) > 0$

Для описания алгоритма использовались в основном источники [35] и [36].

Алгоритм Affinity Propagation – это один из современных методов кластеризации, который успешно применяется в широких областях исследований компьютерных наук. Однако данный алгоритм не масштабируются для больших наборов данных, поскольку это требует квадратичное процессорное время в количестве точек данных, используемых в вычислениях.

Однако существуют более быстрые версии алгоритма Affinity Propagation, как Fast Affinity Propagation. В дальнейшем будет рассматриваться только Fast Affinity Propagation.

2.1.1.4.2 Fast Affinity Propagation

В основе Fast Affinity Propagation лежит идея обрезать ненужный обмен сообщениями в течении выполнения итераций в Affinity Propagation. Подобная идея родилась из факта, что объект играет важную роль в подборе образцов рядом с ним, но не имеет ничего общего с образцами, расположенными далеко от него. Поэтому обмен сообщениями между удаленными объектами может быть опущен.

У самого алгоритма Fast Affinity Propagation также существует некоторое множество вариантов, которые различаются тем, каким конкретно образом осуществляется выбор того, какие обмены сообщениями являются излишними, а в следствии и не нужными, и отбрасываются.

Так в [37] и [38] были построены разреженные факторные графы, а затем была реализована передача сообщений по разреженным графам. Наблюдение заключается в том, что объект играет важную роль в подборе

образцов рядом с ним, но имеет ограниченное влияние на выбор образцов, находящихся далеко от него.

В [37] был построен разреженный факторный граф с использованием k-ближайших соседей (k-NN). Однако применение k-NN привело к неожиданно слишком большому количеству фрагментов, что привело к слишком большому количеству примеров на следующем этапе.

Аналогично, в [38] также был принят метод k-NN для построения разреженного факторного графа, но тут стратегия состоит в том, чтобы добавить некоторые ребра на полученном методом k-NN разреженном факторном графе. После добавления ребер передача сообщений снова реализуется на новом разреженном графе, чтобы получить окончательный результат кластеризации. Однако результаты не сильно улучшились, и эффективность кластеризации все еще была ниже, чем у исходного алгоритма Affinity Propagation.

В [39] были оценены верхняя и нижняя границы обменных сообщений между каждой парой объектов, а затем построен разреженный факторный граф в соответствии с этими границами. Конечный результат кластеризации вычислялся с использованием построенного разреженного факторного графа. В результате кластеризации, по мнению авторов источника, должны получаться кластеры абсолютно идентичные кластерам, полученных с помощью алгоритма Affinity Propagation, но быстрее.

Таким образом для дальнейшего рассмотрения была выбрана вариация Fast Affinity Propagation, представленная в [39].

Алгоритм

Пусть $\{x_1, \dots x_n\}$ - множество точек данных без внутренней структуры;

s - функция, которая количественно определяет сходство между любыми двумя точками, так что выражение $s(i,j) > s(i,k)$ справедливо тогда и только тогда, когда x_iбольше похож на x_j, чем на x_k.

Первым шагом нужно вычислить верхние/нижние границы обменных сообщений между каждой парой объектов. Следующие определения вычисляют верхние / нижние ограничивающие оценки $\underline{a}$, $\overline{r}$ и $\overline{a}$ [39]:

$$\underline{a}[i,j] = \begin{cases} min(\,0, r[j,j]) & (i \neq j\,) \\ 0 & (i = j) \end{cases} \quad (10)$$

$$\overline{r}[i,j] = \begin{cases} s[i,j] - \ max_{k \neq j}\big(\underline{a}[i,k] + \ s[i,k]\big) & (i \neq j) \\ s[i,j] - \ max_{k \neq j}(\,s[i,k]) & (i = j) \end{cases} \quad (11)$$

$$\overline{a}[i,j] = \begin{cases} min\big(\,0, \overline{r}[j,j] + \ \sum_{k \neq i,j} max(0, \overline{r}[k,j])\big) & (i \neq j\,) \\ \sum_{k \neq i,j} max(0, \overline{r}[k,j]) & (i = j) \end{cases} \quad (12)$$

Заметим, что $\underline{a}[i,j]$ можно вычислить только из парных совпадений $s\,[i,j]$, так как [39]:

$$r\,[j,j] \ = \ s[j,j] - \ max_{k \neq j}(\,s[j,k]) \quad (13)$$

Следующий шаг, оценить верхнюю и нижнюю границы обменных сообщений между каждой парой точек, используя верхние / нижние ограничивающие оценки $\underline{a}$, $\overline{r}$ и $\overline{a}$, и построить разреженный факторный граф в соответствии с этими границами. Используя условия, прописанные в (14), строится разреженный факторный граф следующим образом: все пары точек, для которых заданное условие является истинным, соединяются ребрами [39].

$$\begin{cases} \overline{r}[i,j] \geq 0 \\ \overline{a}[i,j] + s[i,j] \geq max_{k \neq j}\big(\underline{a}[i,k] + s[i,k]\big) \end{cases} \quad (14)$$

Таким образом получается разреженный факторный граф.

Далее для всех соединенных пар точек в разреженном факторном графе применяется классический Affinity Propagation, который выполняется путем чередования двух шагов передачи сообщения, для обновления двух матриц R и A, пользуясь уравнениями (8) – (9). Итерации выполняются до тех пор, пока таблицы R и A не останутся неизменными по целому ряду итераций или после некоторого заданного количества итераций.

Далее для всех несоединенных пар точек в разреженном факторном графе применяются следующие выражения [39]:

$$r[i,j] = \rho[i,j] \quad (15)$$

$$a[i,j] = \gamma[i,j] \quad (16)$$

Образцами из окончательных матриц считаются те, чья «ответственность + доступность» для себя положительна, то есть $r(i,i) + a(i,i) > 0$.

2.1.1.5 DBSCAN

DBSCAN (Density-based spatial clustering of applications with noise, плотностной алгоритм пространственной кластеризации с присутствием шума) – это алгоритм кластеризации данных, предложенный Мартином Эстер, Хансом-Питером Кригелем, Йоргом Сандером и Xiaowei Xu в 1996 году. Это алгоритм кластеризации основан на плотности распределения точек данных. DBSCAN является одним из наиболее распространенных алгоритмов кластеризации, а также наиболее часто цитируется в научной литературе [40]

Основной идеей алгоритма DBSCAN является представление объектов кластера в виде группы точек в метрическом пространстве, являющихся вершинами одного связного графа. Причём две точки в таком графе соединяются ребром только в том случае, если расстояние между ними в заданной метрике не превышает определённого расстояния. Если рядом с некоторым объектом нет достаточно близких соседей, то он признаётся выбросом [41].

Рассмотрим множество точек в некотором пространстве, которое нужно кластеризавать. В рассматриваемом алгоритме кластеризации DBSCAN все точки данных по окончанию будут разделены на три типа: основные точки, достижимые точки и выбросы [43].

Распределение точек данных на типы будет зависеть от двух входных параметров: количество соседних точек, необходимое, чтобы создать толпу - ε; и максимальное расстояние между точками, при котором они могут считаться соседями - $minPts$.

Основными точки (core points) будут считаться, когда они имеют соседей в количестве хотя бы $minPts - 1$. Другими словами, точка называется основной, если в ε –окрестности точки находятся MinPts объектов (включая саму точку). Все соседи этой точки называются напрямую достижимыми для нее.

Достижимыми точки (reachable points) будут считаться, когда они не имеют достаточного количества соседей, чтобы считаться основными, но среди их имеющихся соседей присутствуют основные точки, что позволяет им быть достижимыми для толпы.

Выбросами (outliers) называются все точки, кто имеет мало соседей, и среди которых нет основных точек.

После распределения всех точек по типом формируются кластеры: сначала объединяются скопление основных точек, которые связаны друг с другом через соседей, а затем к образовавшемуся кластеру присоединяются и все достижимы точки, которые связаны с основными точками, уже относящихся к данному кластеру. Точки-выбросы считаются шумом и не объединяются ни с одним кластером. На рисунке ниже показано графическое представление процесса выполнения алгоритма DBSCAN.

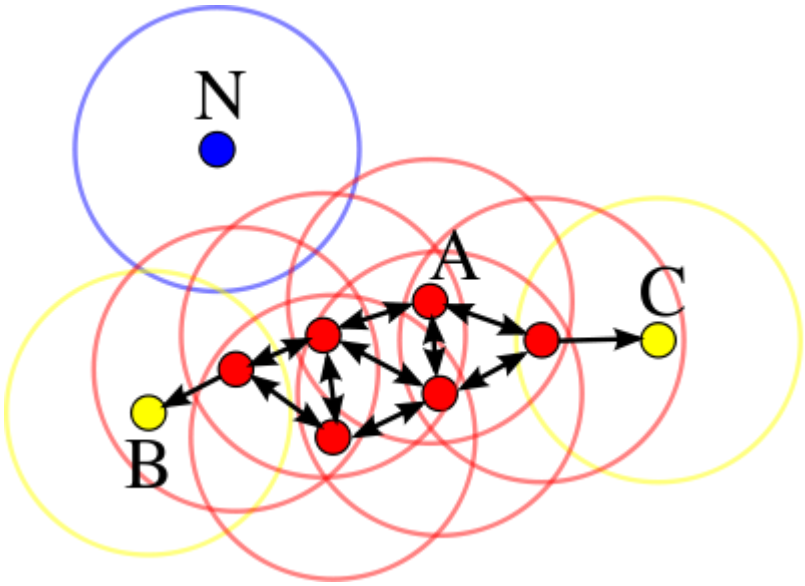

Рисунок 1 – формирование кластера алгоритмом DBSCAN

Алгоритм

Входные данные: множество объектов X, для которых задана метрическая функция расстояния ρ.

Выходные данные: размеченное множество объектов, где каждому объекту сопоставлен порядковый номер кластера, в который попал данный объект, либо объект отмечается как выброс (шум).

ε - максимальное расстояния между соседними объектами,

$minPts$ - минимальное количество соседних объектов, необходимых для образования кластера.

Пошаговая инструкция:

1. Выбираем необработанный объект p.
2. Отмечаем объект p как обработанный.
3. Находим соседние объекты в ε-окрестности объекта p.
4. Сравниваем количество соседних объектов с MinPts, определяя, является ли p ядровым объектом
   - Если объект p является ядровым, то создаём новый кластер и запускаем поиск в ширину из данного объекта по другим не посещённым объектам, находя все объекты кластера.
   - Если объект p не является ядровым, то отмечаем его как выброс (или шум).
5. Если присутствуют необработанные объекты, то возвращаемся к шагу 1.

2.1.2 Сравнение рассмотренных алгоритмов кластерного анализа, не требующих явного задания количества кластеров

Сравнение всех рассмотренных алгоритмов кластерного анализа, не требующих явного задания количества кластеров рассмотрено в формате таблицы в приложение А.

Таблица содержит такие столбцы для сравнения, как сложность алгоритма, форма получающихся кластеров, результат алгоритма, чувствительность алгоритма и специфики алгоритма.

В столбце специфики алгоритма указаны некоторые плюсы и минусы алгоритмов.

2.2 Выбор показателей качества для оценки результатов применения различных методов кластерного анализа

Использование кластерного анализа требует решения задачи измерения качества кластеризации. Задача оценки качества кластеризации является более сложной по сравнению с оценкой качества классификации. Во-первых, такие оценки не должны зависеть от самих значений меток, а только от самого разбиения выборки. Во-вторых, не всегда известны истинные метки объектов, поэтому также нужны оценки, позволяющие оценить качество кластеризации, используя только неразмеченную выборку[43].

Для решения задачи определения качества результатов кластерного анализа используются метрики оценки эффективности кластеризации. Далее рассматриваются некоторые метрики оценки эффективности кластеризации.

Выделяют внешние и внутренние метрики качества. Внешние используют информацию об истинном разбиении на кластеры, в то время как внутренние метрики не используют никакой внешней информации и оценивают качество кластеризации, основываясь только на наборе данных [43].

В качестве внешней метрики рассмотрена V-мера, а в качестве внутренней метрики рассмотрена мера под названием силуэт.

2.2.1 Однородность, полнота, V-мера

Задача кластеризации состоит в разделении всех объектов выборки на кластеры таким образом, чтобы каждый кластер содержал только те объекты, которые являются членами одного и того же класса. При наличии истинных меток классов у объектов выборки, определить была ли достигнута идеальная

кластеризация, довольно просто. Тем не менее, оценка того, насколько далёко полученное кластерное решение от идеала, является более сложной задачей и предлагаемые подходы часто не являются строгими, как например в [44].

V-мера – мера оценки качества кластеризации на основе энтропии, предназначенная для решения проблемы количественной оценки кластерного несовершенства [45]. Как и все внешние меры, V-мера сравнивает истинную классификацию - например, аннотированное вручную представительное подмножество данных - с автоматически сгенерированной кластеризацией этого подмножества данных, и определяет их схожесть между собой количественно.

V-мера включает в себя две оценочных меры: однородность (homogeneity) и полнота (completeness).

Однородность измеряет, насколько каждый кластер состоит из объектов одного класса, а полнота - насколько объекты одного класса относятся к одному кластеру. Эти меры не являются симметричными. Обе величины принимают значения в диапазоне [0, 1], и чем ближе значение к 1, тем точнее кластеризация. Эти меры не являются нормализованными, что значит, что они зависят от числа кластеров [43].

Однородность и полнота считаются по следующим формулам [43]:

$$h = \begin{cases} 1 & (H(C,K) = 0) \\ 1 - \frac{H(C|K)}{H(C)} & (H(C,K) \neq 0) \end{cases} \quad (17)$$

$$c = \begin{cases} 1 & (H(K,C) = 0) \\ 1 - \frac{H(K|C)}{H(K)} & (H(K,C) \neq 0) \end{cases} \quad (18)$$

где

h, c – однородность и полнота соответственно;

C, K – истинное разбиение на классы и результаты кластеризации соответственно;

$H(C|K), H(C)$ – это функции энтропии классов и условной энтропии классов с учетом кластерных назначений соответственно, которые рассчитываются по следующим формулам [43]:

$$H(C|K) = -\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{n_{c,k}}{n} * \log(\frac{n_{c,k}}{\sum_{c=1}^{C} n_{c,k}}) \quad (19)$$

$$H(C) = \sum_{c=1}^{C}\frac{\sum_{k=1}^{k} n_{c,k}}{n} * \log(\frac{\sum_{k=1}^{k} n_{c,k}}{n}) \quad (20)$$

где

n - общее количество образцов в выборке;

$n_{c,k}$ - количество образцов из класса c назначен на кластер k.

Условная энтропия кластеров данного класса $H(K|C)$ и энтропия кластеров $H(K)$ определены симметричным образом.

V-мера представляет собой взвешенное среднее гармоническое значение однородности и полноты [43]:

$$V = 2 * \frac{h*c}{(h+c)} \quad (21)$$

2.2.2 Силуэт

Если основные метки истинности неизвестны, оценка должна выполняться с использованием самой модели. Силуэт позволяет оценить качество кластеризации, используя только саму (неразмеченную) выборку объектов и результат кластеризации.

Силуэт является оценочной мерой кластерного анализа, которая отдаст более высокий показатель среди нескольких кластеризаций модели с более четко определенными кластерами. Силуэт определяется для каждого образца и состоит из двух показателей:

a – среднее расстояние между образцом и всеми другими точками в том же кластере;

b – среднее расстояние между образцом и всеми остальными точками в следующем ближайшем кластере.

Тогда силуэт объекта рассчитывается следующим образом [43]:

$$s = \frac{b-a}{\max(a,b)} \quad (22)$$

Силуэт показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров [46].

Силуэтом выборки называется средняя величина силуэта объектов данной выборки [43]:

$$S = \frac{\sum_{i=1}^{N} s_i}{N} \quad (23)$$

Данная величина лежит в диапазоне [-1, 1]. Значение силуэта близкое к -1 соответствует плохой (разрозненной) кластеризации. Значение силуэта близкое к нулю, говорят о том, что кластеры пересекаются и накладываются друг на друга, значения, близкие к 1, соответствуют четко выделенным кластерам. Таким образом, чем больше силуэт, тем более четко выделены кластеры, и они представляют собой компактные, плотно сгруппированные облака точек [43].

2.3 Исследование подходов к выбору наиболее эффективного метода кластеризации в зависимости от исходных данных

Разные выборки данных обладают разными параметрами (кучность данных, корреляция данных, разброс данных и т.п), поэтому один и тот же алгоритм кластеризации на двух разных выборках, обладающие непохожими параметрами, дают результаты разного качества. Таким образом, для получения более качественной кластеризации для каждой выборки необходимо подбирать подходящий ей алгоритм кластерного анализа.

Выделены два пути определения наиболее эффективного алгоритма кластеризации для исходных данных:

1. Выделение значимых параметров выборок данных, подверженных кластеризации.

2. Использование классификаторов для поиска закономерностей в параметрах выборок, для которых нужно применять одинаковые алгоритмы кластеризации.

2.4 Выделение значимых параметров выборки данных

Выделение значимых параметров является очень кропотливой и сложной работой и является нетривиальной задачей. Выделенные параметры должны четко описывать состояние выборки, что легко позволит подобрать правильный алгоритм кластеризации.

Выделение производится смесью практического и аналитического метода.

Необходимо подобрать некоторое количество выборок, обладающих разными характеристиками. Затем к каждой выборки применить все рассматриваемые алгоритмы кластеризации и определить какой алгоритм дает более качественные результаты. Обобщить результаты всех выборок, которые имеют одинаковые характеристики.

2.4.1 Использование классификаторов для поиска закономерностей

При использовании классификатора по сути происходит тоже выделение значимых параметров кластеризуемых выборок, только программным путем.

Для использования классификатора нужно также подобрать некоторое количество выборок, обладающих разными характеристиками. Затем к каждой выборки применить все рассматриваемые алгоритмы кластеризации и определить какой алгоритм дает более качественные результаты. Составить обучающую выборку для классификатора. Обучить классификатор, и потом просто использовать его для определения алгоритма кластеризации.

Проблемой использования классификаторов заключается в том, какие данные подавать на вход классификатора, как при его обучении, так и при его использовании. Ведь в зависимости от числа узлов в социальном графе выборки имеют разную размерность.

Выделим два варианта:

1. Выделение n-ого количества значимых точек из рассматриваемых выборок.

Выделение значимых точек крайне лёгкий способ. Однако при его использовании решается одна важная проблема. Дело в том, что, во многих случаях данные выборок имеют высокую корреляцию менее значительных узлов, что создает высокую корреляцию данных и для всей выборки (Рисунок 2). В таких случаях классификаторы переобучаются и не дают качественных результатов в применении.

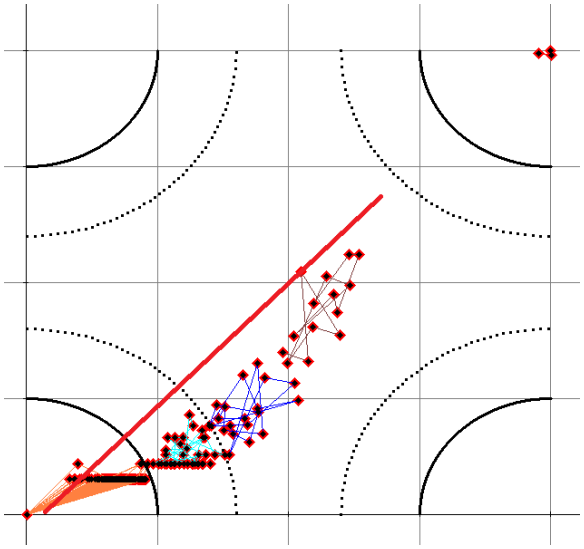

Рисунок 2 – Наблюдение общей корреляции данных выборки

Это связано с тем, что данными являются метрики Боргатти, Кендалла-уэя и Энтропии, выделенные при обработки социального графа пользователей социальных сетей, которые часто имеют линейную зависимость.

Однако, в некоторых случаях, подобная корреляция не наблюдается для значимых узлов (Рисунок 3). Таким образом, если обучать классификаторы только на небольшой группе некоррелированных узлов, то можно получить от классификатора более достойный результат.

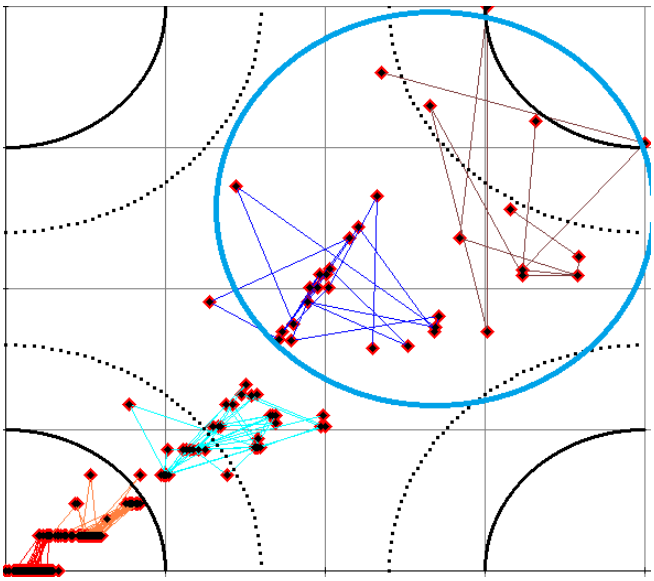

Рисунок 3 –Наблюдение корреляции данных отдельных значимых узлов

1. Трансформировать входные данные в вектор заданной размерности.

Для этого исходные данные для кластерного анализа можно представить, как полносвязный граф, и трансформировать в граф меньшего заранее заданного масштаба, после чего подавать на вход классификатору.

Идея этих методов состоит в том, что алгоритмы трансформируют данные таким образом, что схожи данные оказываются в новом d-мерном пространстве где-то рядом, и наоборот разные данные, где-то далеко. На Рисунок 4 визуализирована общая суть таких методов. Примерами таких подходов являются такие методы, как doc2vec или graph2vec.

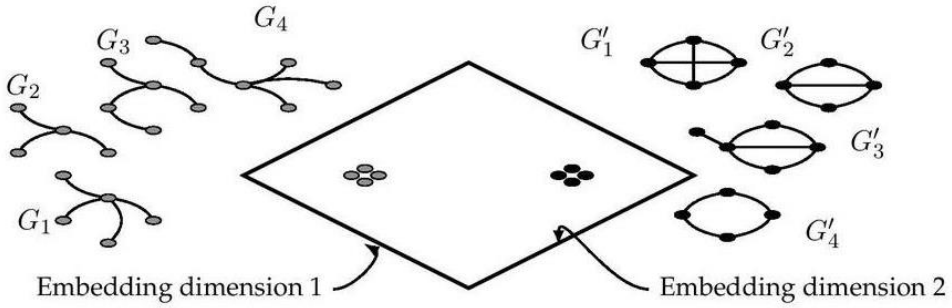

Рисунок 4 – Визуализация методов трансформации графа в меньший заданный размер

2.4.2 Подбор подхода выбора наиболее эффективного метода кластеризации в зависимости от исходных данных

Из выделенных путей определения наиболее эффективного алгоритма кластеризации для исходных данных использование классификатора является более простым в реализации. Также классификатор может выявить некоторые характеристики, которые человек не сможет выявить при самостоятельном анализе, выборок данных, анализируемые одним и тем же алгоритмом кластеризации.

Поэтому выбран путь использования классификатора для выбора наиболее эффективного метода кластеризации в зависимости от исходных данных.

Был выбран линейный классификатор дерево решений из-за своей простоты и понятности.

Деревья решений является методом классификации, который позволяет предсказывать принадлежность наблюдаемого объекта к одному из выделенных ранее классу. Определение принадлежности к тому или иному классу зависит от соответствующих значений выделенных предикторных переменных [47].

Деревья решений предоставляют прямой и интуитивно понятный способ получения классификации нового экземпляра из набора простых правил. Из-за этой характеристики деревья решений находят широкое применение в ситуациях, когда интерпретация полученных результатов и процесса рассуждения имеет решающее значение, например, в компьютерной диагностике (CAD) и в анализе финансового кредита [47].

Дерево решений является логическим алгоритмом классификации, основанный на поиске конъюнктивных закономерностей. При этом, все конъюнкции строятся одновременно при генерации дерева решений [48].

Дерево представляет собой конечный связный граф с множеством вершин V, который не содержит циклов, но имеет выделенную вершину $v_0 \in$

V, в которую не входит ни одно ребро. Такая вершина называется корнем дерева решений. Вершина, которая не имеет выходящих рёбер, называется терминальной или листом. Остальные вершины называются внутренними [48].

Дерево называется бинарным, когда из всех внутренних вершин выходит ровно два ребра. Выходящие рёбра связывают каждую внутреннюю вершину v с левой дочерней вершиной L_v и с правой дочерней вершиной R_v. [48].

Бинарное решающее дерево –классификатор, рассчитывающийся на основе бинарного дерева, в котором корню дерева и каждой внутренней вершине $v \in V$ приписан предикат $\beta_v : X \to \{0, 1\}$, а каждой терминальной вершине $v \in V$ приписано имя класса $c_v \in Y$ [48].

Вовремя расчета классификатора бинарного решающего дерева объект $x \in X$ проходит по дереву путь от корневой вершины до одного из листа дерева по следующему алгоритму: попадая на очередную вершину графа, если она не является листом, вычисляется приписанный ей предикат β_v, если расчитанный предикат $\beta_v = 1$, то объект переходит на правую дочернюю вершину R_v., иначе объект переходит на левую дочернюю вершину L_v. [48].

Таким образом, алгоритм классификации, реализуемый бинарным решающим деревом, можно записать в виде простого голосования конъюнкций. Математическая модель классификатора бинарного дерева решений выглядит следующим образом [48]:

$$a(x) = \arg\max_{y \in Y} \sum_{\substack{v \in T \\ c_v = y}} K_v(x) \qquad (24)$$

Где

x – классифицируемый объект;

$a(x)$ – классификатор, реализуемый бинарным решающим деревом, для объекта x;

$K_v(x)$ – конъюнкция, составленная приписанным вершине v предикатами, примененная к объекту x;

T – множество всех терминальных вершин дерева.

2.4.3 Определение метрики качества обучения классификатора для определения наиболее эффективного метода кластеризации в зависимости от исходных данных

В области машинного обучения и, в частности, в области классификации, существует проблема оценки качества машинного обучения.

Для решения данной проблемы используются метрики оценки качества обучения. В машинном обучении существует большое количество метрик качества, каждая из которых имеет свою прикладную интерпретацию и направлена на измерение конкретного свойства решения.

Расчет многих метрик оценки качества обучения классификации строится на матрице ошибок (confusion matrix) [49].

Матрица ошибок, также известная как матрица путаницы, представляет собой конкретную схему таблиц, которая позволяет визуализировать производительность алгоритма контролируемого обучения. Каждая строка матрицы представляет вариант ответа классификатора, в то время, как каждый столбец представляет варианты истинных меток объектов. На пересечении строки и столбца записывают количество объектов, которые имеют данную столбцом истинную метку, определенных по мнению классификатора в данный строкой класс [49]. Таким образом, используя матрицу ошибок, легко увидеть, смешивает ли система классы (маркирует один как другой).

В Таблица 1 представлена общая схема матрицы ошибок при бинарной классификации:

Таблица 1 - Общая схема матрицы ошибок при бинарной классификации

	Истинная метка = 1	Истинная метка = 0
Определенная классификатором метка =1	True Positive (TP) – количество объектов, имеющих истинную метку равную 1, и определенные классификатором в класс с меткой 1	False Positive (FP) - количество объектов, имеющих истинную метку равную 0, но определенные классификатором в класс с меткой 1
Определенная классификатором метка =0	False Negative (FN) - количество объектов, имеющих истинную метку равную 1, но определенные классификатором в класс с меткой 0	True Negative (TN) - количество объектов, имеющих истинную метку равную 0, и определенные классификатором в класс с меткой 0

Таким образом при бинарной классификации существует два типа ошибок: False Positive (FP) и False Negative (FN).

На основе матрицы ошибок высчитывают такие показатели эффективности обучения, как:

- Истиноположительная норма (true positive rate, TPR) – измеряет долю фактических положительных результатов, которые правильно определены как таковые [50]. Определяется по следующей формуле:

$$TPR = \frac{TP}{TP+FN} \quad (24)$$

- Истиноотрицательная норма (true negative rate, TNR) – измеряет долю фактических отрицательных значений, которые правильно определены как таковые [50]. Определяется по формуле:

$$TNR = \frac{TN}{TN+FP} \quad (25)$$

- Ложноположительная норма (false positive rate, FPR) - измеряет долю фактических положительных результатов, которые неправильно определены как таковые [49]. Определяется по следующей формуле:

$$FPR = \frac{FP}{FP+TN} \quad (26)$$

- Ложноотрицательная норма (false negative rate, FNR) – измеряет долю фактических отрицательных значений, которые неправильно определены как таковые [49]. Определяется по следующей формуле:

$$FNR = \frac{FN}{FN+TP} \quad (27)$$

- Положительная прогностическая ценность (positive predictive value, PPV) - измеряет долю положительных результатов определения класса классификатором, которые являются истиноположительными [51]. Определяется по следующей формуле:

$$PPV = \frac{TP}{TP+FP} \quad (28)$$

- Отрицательно прогностическая ценность (negative predictive value, NPV) - измеряет долю отрицательных результатов определения класса классификатором, которые являются истиноотрицительными [51]. Определяется по следующей формуле:

$$NPV = \frac{TN}{TN+FN} \quad (29)$$

На основе матрицы ошибок рассчитываются следующие показатели оценки качества обучения бинарного классификатора:

- Точность (Accuracy, ACC) – измеряет долю правильных ответов классификации. ACC возвращает значение от 0 до 1, где 1 – это идеальная точность обучения бинарного классификатора, а 0 – низкая

точность обучения, когда наблюдается полное расхождение истинных меток объекта и присвоенных классификатором [52]. Определяется по следующей формуле:

$$ACC = \frac{TR+TN}{TP+TN+FN+FP} \quad (32)$$

- F-мера (F) - является средним гармоническим истиноположительной нормы и положительной прогностической ценности. F-мера возвращает значение от 0 до 1, где 1 – это идеальная точность обучения бинарного классификатора, а 0 – низкая точность обучения, когда наблюдается полное расхождение истинных меток объекта и присвоенных классификатором [49]. Определяется по следующей формуле:

$$F = \frac{2*PPV*TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN} \quad (33)$$

- Коэффициент корреляции Мэтьюса (Matthews correlation coefficient, MCC) – считается сбалансированной мерой оценки качества обучения бинарной классификации, которую можно использовать, даже если классы очень разных размеров [53]. MCC по сути является коэффициентом корреляции между истинном разделением на классы и разделением на классы бинарным классификатором. MCC возвращает значение от -1 до +1. Коэффициент +1 представляет собой идеальное предсказание, 0 не лучше, чем случайное предсказание, а -1 указывает на полное расхождение между предсказанием и наблюдением [54]. Определяется по следующей формуле:

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (34)$$

2.5 Описание методики обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

Обобщая все вышесказанное, выведена следующая методика обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети:

1. Берем выборку данных, содержащую показателей Боргатти, информационной энтропии, и коммуникационной эффективности всех пользователей социальной сети, кто связан с заданной предметной областью;
2. Выбираем первые 20 значимых узла взятой выборки данных и передаем их обученному классификатору;
3. Получаем результат классификатора, передаем полную выборку данных подобранному классификатором алгоритму кластерного анализа;
4. Получаем результаты кластеризации, переводим их в доступную для понимания человеком форму, выдаем результат.

3 Проектирование и разработка программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

В данной главе описываются такие важные этапы создания программного приложения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети, как его проектирования и реализация.

3.1 Проектирование программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

3.1.1 Описание используемых технологий

В качестве инструмента разработки программного обеспечения выбрана среда разработки Microsoft Visual Studio 2015.

В качестве основного языка разработки был выбран объектно-ориентированный язык C#.

Для реализации классификатора дерево решений использована библиотека Accord.NET Framework.

Accord.NET Framework - это расширение платформы популярной платформы AForge.NET , добавляющее и предоставляющее новые алгоритмы и методы для машинного обучения, компьютерного зрения и не только [54].

Для реализации визуализации результатов кластеризации, представленных в виде трехмерной диаграммы рассеивания, была использована библиотека nzy3d-api.

Nzy3d-api – это инструментарий, основанный на OpenTK для интеграции OpenGL. Nzy3d-api позволяет легко рисовать 3d научные данные: поверхности, точечные диаграммы, гистограммы и множество других трехмерных примитивов, обеспечивает поддержку многофункциональных интерактивных диаграмм с цветными полосами, подсказками и наложениями. Ось и макет диаграммы могут быть полностью настроены и улучшены [55].

При это Nzy3d-api дает возможность:

- Взаимодействие мыши с объектами;
- Взаимодействие мыши с графиком (вращение, масштабирование);
- Взаимодействие клавиш с графиком (вращение, масштабирование);
- Анимация структур объектов.

3.1.2 Выделение функциональных требований

Программное обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети должно соответствовать следующим функциональным требованиям:

- Наличие средств для загрузки данных;
- Возможность кластеризации данных алгоритмом DBScan;
- Возможность кластеризации данных алгоритмом Fast Affinity Propagation;
- Возможность автоматической кластеризации данных, организованной по выведенной методике;
- Наличие средств определения качества выполненной кластеризации;
- Наличие средств визуализации результатов кластеризации.

3.1.3 Определение вариантов использования программного обеспечения

Чтобы проектирование программного обеспечения было выполнено корректно, прежде всего нужно определить его функциональное назначение. Другими словами, необходимо определить, что данное программное обеспечения будет делать во время своего функционирования. Для этого используется диаграмма вариантов использования UML (Use case diagrams).

Диаграмма вариантов использования отображает отношения между актерами и вариантами использования. Актерами называют внешнюю по

отношению к программному обеспечению сущность, которая взаимодействует с системой. Вариант использования программного обеспечения представляет из себя полное описание поведения программной продукта в ответ на определенное действие актера [57].

Для проектируемой системы в качестве актера выступает пользователь системы, который может выполнять ограниченное число манипуляций с проектированным продуктом.

На рисунке ниже представлен сценарий вариантов использования проектируемой системы, оформленный в виде диаграммы вариантов использования UML.

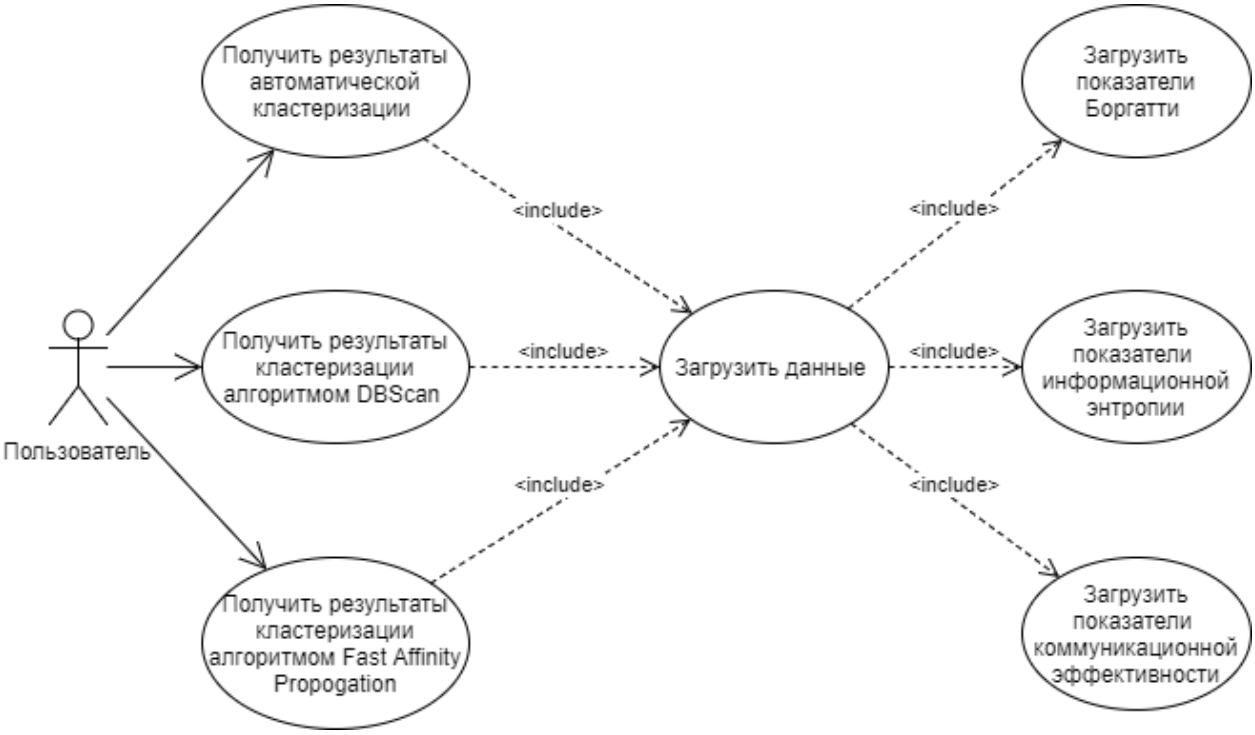

Рисунок 5 – Диаграмма вариантов использования

На представленной диаграмме вариантов использования для проектируемого программного обеспечения показано, что пользователь может использовать приложение либо для автоматической кластеризации данных, либо для кластеризации данных алгоритмом DBScan, либо для кластеризации данных алгоритмом Fast Affinity Propagation. Каждый из перечисленных

вариант использования включает в себя загрузку данных, которая, в свою очередь, не обходится без загрузки данных по каждому из используемых в разработанной методики показателю.

3.1.4 Описание архитектуры программного обеспечения

Проектируемое программного обеспечение создается для дальнейшего встраивания в сервис для загрузки и обработки данных социальной сети в качестве отдельного компонента.

Архитектура развертывания этого программного сервиса с встроенным проектируемым программным обеспечением в качестве компонента для идентификации пользователей-экспертов представлена ниже (Рисунок 6). Она основана на клиент-серверном архитектурном стиле.

Сервер содержит компоненты, на которых проходят вычислительные процессы, связанные с выгрузкой данных из социальной сети Twitter, их обработкой и анализом. К таким компонентам относится и разрабатываемое программное обеспечение. Там же на сервере хранятся загруженные коллекции данных пользователей.

На стороне клиента происходит получение входных данных от пользователя сервисом (например, введенное им ключевое слово для поиска данных социальной сети, запрос на проведения анализа.

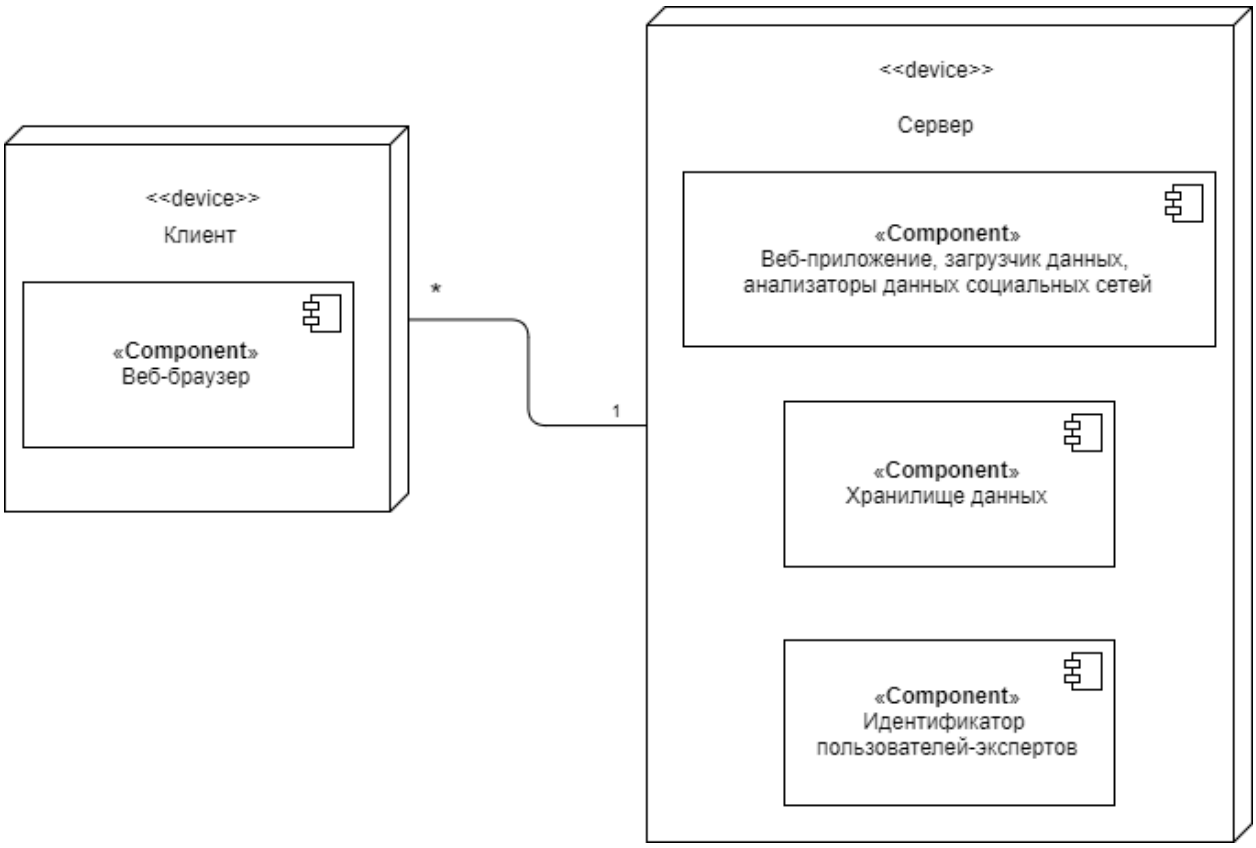

Рисунок 6 – Диаграмма развертывания веб-приложения для загрузки и обработки данных социальной сети с встроенным разрабатываемым компонентом

Рисунок 7 демонстрирует детализированную компонентную архитектуру серверной части. В разработанной архитектуре используются два хранилища данных:

1) MongoDB для хранения загруженных данных социальной сети Twitter и результата их анализа;

2) Microsoft SQL Server для хранения учетных записей пользователей и запланированных задач загрузки и анализа данных.

Компонент «Идентификатор пользователей-экспертов» будет извлекать данные результатов анализа данных пользователей социальной сети Twitter из хранилища MongoDB в виде показателей Боргатти, информационной энтропии, и коммуникационной эффективности и анализировать их.

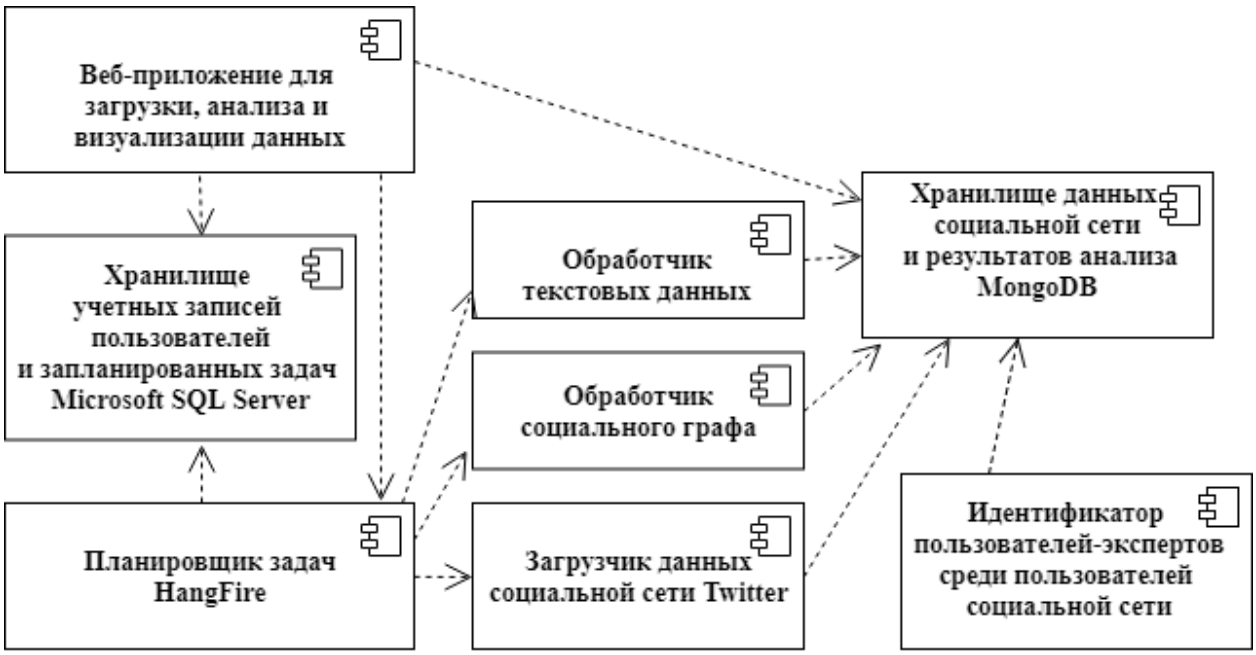

Рисунок 7 - Детализированная компонентная архитектура серверной части веб-приложения для загрузки и обработки данных социальной сети с встроенным разрабатываемым компонентом.

Компонент «Идентификатор пользователей-экспертов» также можно детализировать. Он включает в себя такие компоненты, как:

- Компонент обработки данных и демонстрации результатов;
- Компонент загрузки и обработки данных из документа формата XLSX;
- Компонент кластеризации;
    - Компонент кластеризации алгоритмом DBScan;
    - Компонент кластеризации алгоритмом Fast Affinity Propagation;
- Компонент классификации;
- Компонент оценки результатов;
- Компонент визуализации
    - Компонент организации визуализации;
    - Компонент Nzy3d.

Компонент обработки данных и демонстрации результатов является зависим от всех остальных. Он использует функции других компонентов, чтобы получить и обработать данные о пользователях социальной сети для выявления пользователей-экспертов среди них.

Рисунок 8 показывает описанную детализированную компонентную архитектуру разрабатываемого компонента «Идентификатор пользователей-экспертов».

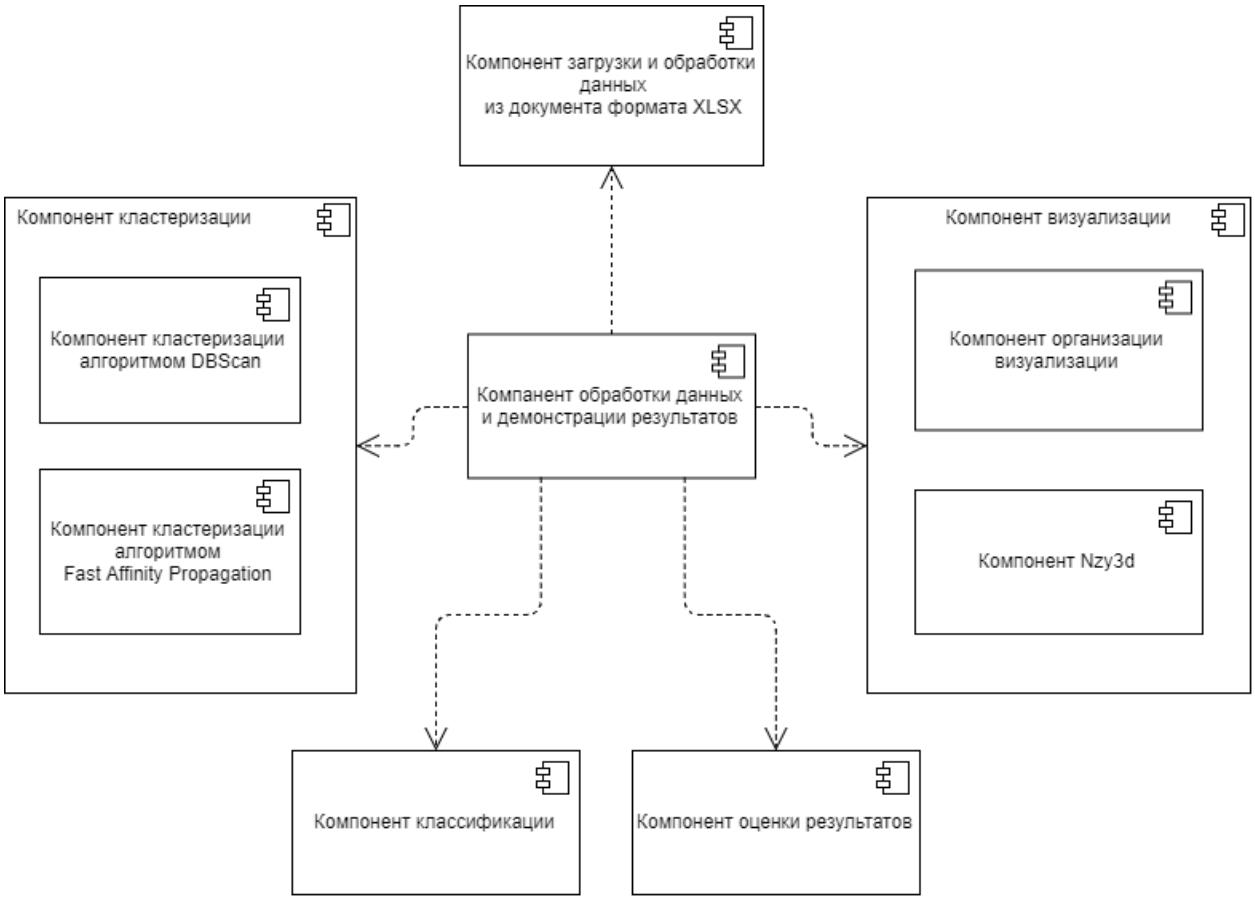

Рисунок 8 - Детализированная компонентная архитектура разрабатываемого компонента «Идентификатор пользователей-экспертов»

3.1.5 Описание микроархитектуры программного обеспечения

Микроархитектура проектируемого программного компонента представлена в виде диаграммы классов в приложении Б.

Представленная диаграмма классов включает в себя 11 классов. Таблица 2 содержит краткие описания каждого класса.

Таблица 2 - Описание классов проектируемого компонента

Наименование класса	Краткое описание класса
Data	Является классом, хранящий данные о пользователях.
FuncForLoadDate	Представляет из себя набор функций для реализации загрузки данных из документов формата XLSX
LoadDateForm	Представляет из себя пользовательское окно. Загружает данные о пользователях социальной сети из документов формата XLSX, используя функции класса FuncForLoadDate. Загруженные данные представляются в виде экзепляров класса Data
ForWorkData	Представляет из себя пользовательское окно. Получает данные о пользователях социальной сети от класса LoadDateForm. Выполняет процессы обработки данных и демонстрации результатов, используя функции следующих классов: FSAPAlgorithm<T>, DbscanAlgorithm<T>, DecTree, AccuracyMetrics , Visualization.
FSAPPoint<T>	Является классом, хранящий данные о пользователях в формате пригодного для дальнейшего применения при кластеризации алгоритмом Fast Affinity Propagation
DbscanPoint<T>	Является классом, хранящий данные о пользователях в формате пригодного для дальнейшего применения при кластеризации алгоритмом DBScan
FSAPAlgorithm<T>	Представляет из себя набор функций для кластерного анализа алгоритмом Fast Affinity Propagation. При этом используются экземпляры класса FSAPPoint<T>.

Продолжение Таблица 2

DbscanAlgorithm<T>	Представляет из себя набор функций для кластерного анализа алгоритмом DBScan. При этом используются экземпляры класса DbscanPoint <T>.
DecTree	Представляет из себя набор функций для реализации классификатора дерева решений. Используется для организации автоматической кластеризации.
AccuracyMetrics	Представляет из себя набор функций для реализации оценки качества кластеризации.
Visualization	Представляет из себя набор функций для реализации трехмерной диаграммы рассеивания. Используется для визуализации результатов кластеризации.

3.2 Разработка программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

На этапе разработки программного обеспечения проводятся работы по реализации запланированного функционала. Разработка программного обеспечения ведется согласно вышеописанному проектированию системы.

3.2.1 Описание реализации функций загрузки и обработки данных показателей Боргатти, информационной энтропии, и коммуникационной эффективности из документа формата XLSX

Функция загрузки и обработки данных показателей Боргатти, информационной энтропии, и коммуникационной эффективности из документа формата XLSX состоит из нескольких этапов. Для наглядного представления общего алгоритма вычисления, этапы расчета представлены в виде диаграммы деятельности, представленной в приложении В.

3.2.2 Описание реализации функций кластеризации данных определенными алгоритмами

При реализации функций кластеризации данных алгоритмами DBScan и Fast Affinity propagation используются несколько классов.

Для наглядности в приложении Г приведена диаграмма последовательности функции кластеризации данных алгоритмом Fast Affinity propagation. Функция кластеризации данных алгоритмом DBScan выполняется аналогичным способом.

3.2.3 Описание реализации функции автоматической кластеризации

Функция автоматической кластеризации применяет классификатор дерева решений.

Дерево решений реализовано с использованием Accord.NET Framework.

В качестве входных данных подаются первые 20 значимых точек данных.

С начала производится обучение дерева. Для этого загружается составленная заранее обучающая выборка, после чего преобразуется в матрицу, удобную для дальнейшей работы:

```
ExcelReader db = new ExcelReader(@"Resources\res10.xlsx", true, false);
DataTable tableSource = db.GetWorksheet("res10");
double[,] sourceMatrix = tableSource.ToMatrix(out columnNames);
double[,] tableForLearning = (tableSource).ToMatrix(out columnNames);
```

После идет вычленение из полученной матрицы входных и выходных значений для обучения:

```
List<int> index = new List<int>();
for (int i = 1; i < 61; i++)
{
    index.Add(i);
}
double[][] inputsForLearning =
tableForLearning.GetColumns(index.ToArray()).ToJagged();

int[] outputsForLearning = tableForLearning.GetColumn(0).ToInt32();
```

Затем необходимо задать входные переменные для дерево решений:

```
List<DecisionVariable> list = new List<DecisionVariable>();
for (int i = 1; i <= 60; i++)
{
    string x = "x" + i;
    list.Add(new DecisionVariable(x, DecisionVariableKind.Continuous));
}
```

Далее дерево решений обучается:

```
var c45 = new C45Learning(list.ToArray());
tree = c45.Learn(inputsForLearning, outputsForLearning);
```

Для использования дерева решений нужно взять значения 3 параметров у 20 первых значимых точек классифицируемой выборки, привести к определенному виду и подать на вход обученного дерева решений:

```
var SortPoints = Points.OrderByDescending(p => p.X).ToList();

double[,] table = new double[1, 60];
for (int i = 0; i < 60; i++)
{
    int a = (int)Math.Floor((i + 1) / 3.0);
    if ((i + 1) % 3 == 1) { table[0, i] = SortPoints[a].X; }
    if ((i + 1) % 3 == 2) { table[0, i] = SortPoints[a].Y; }
    if ((i + 1) % 3 == 0) { table[0, i] = SortPoints[a].Z; }

}
double[][] inputs = table.ToJagged();

// Вычислить фактические выходные данные дерева
int[] actual = tree.Decide(inputs);
```

На выходе обученное дерево дает одно число – 0 или 1. При нуле подходящий алгоритм кластеризации – DBScan, при единице - Fast Affinity Propagation.

При автоматической реализации сначала используются классификатор для определения алгоритма кластеризации, после чего выполняется функция кластеризации определенным классификатором алгоритмом.

В приложение Д представлена диаграмма последовательности функции автоматической кластеризации, на которой хорошо видно, как применение классификатора, так и выполнение выбранного алгоритма.

3.2.4 Описание реализации функций определения качества кластеризации

После выполнения кластеризации, мера силуэт рассчитывается автоматически. На рисунке ниже представлена диаграмма деятельности функции расчета меры определения качества кластеризации силуэт.

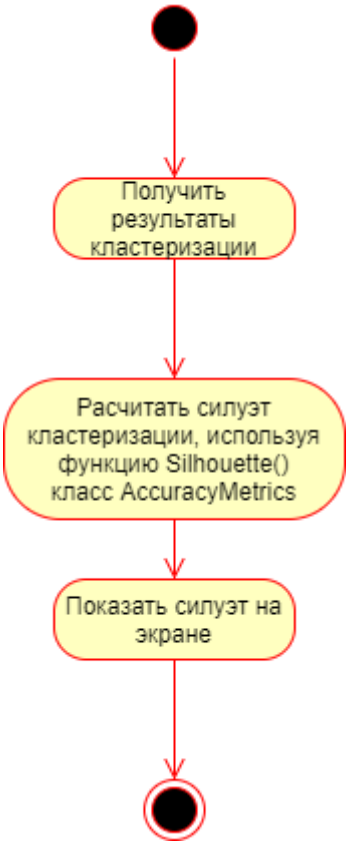

Рисунок 9 - Диаграмма деятельности функции расчета меры определения качества кластеризации силуэт

Для запуска функции расчета V-меры пользователю приложения нужно загрузить метки классов для анализируемых классов. Для это используется функция Excel() класса FuncForLoadDate. После чего результаты кластеризации сравниваются с загруженными метками классов, используя для этого функцию V_measure() класса AccuracyMetrics. При расчете V-меры функцию V_measure() использует приватные функции H1() и H2() класса AccuracyMetrics.

Функции H1() и H2() являются функциями для расчета функций энтропии (19) и условной энтропии (20) соответственно.

```
private static double H1(int[] Nc, int N)
        {
            double H_c = 0;
            for (int i = 0; i < Nc.GetLength(0); i++)
            {
                H_c = H_c + ((double)Nc[i] / N * Math.Log((double)Nc[i] / N));
            }
            return H_c * (-1);
        }
```

```
        private static double H2(int[,] Ncc, int[] Nc, int N)
        {
            double H_cc = 0;
            int m = Nc.Length;

            for (int i = 0; i < Ncc.GetLength(0); i++)
            {
                for (int j = 0; j < Ncc.GetLength(1); j++)
                {
                    if (Ncc[i, j] != 0)
                    {
                        if (m == Ncc.GetLength(0))
                            H_cc = H_cc + ((double)Ncc[i, j] / N *
Math.Log((double)Ncc[i, j] / Nc[i]));
                        else
                            H_cc = H_cc + ((double)Ncc[i, j] / N *
Math.Log((double)Ncc[i, j] / Nc[j]));
                    }
                }
            }
            return H_cc * (-1);
        }
```

3.2.5 Описание реализации функции визуализации

Визуализация реализована в классе Visualization в функции Visual().

Функционал визуализации в функции Visual() реализован с использованием библиотеки Nzy3d-api, которая встроена в приложение отдельным модулем.

Визуализация происходит автоматически по окончанию процесса выполнения кластеризации. Функция Visual() получает от функций кластеризации результаты. После чего, основываясь на полученных данных, происходит генерация диаграммы рассеивания таким образом, что все точки одного кластера имеют один цвет и объединены между собой линиями, создавая тем самым на диаграмме рассеивания явный сгусток. Затем уже для сформированной диаграммы рассеивания в функции Visual() происходит настройка взаимодействия диаграммы с мышью: настройка вращения графики, настройка масштабирования. И наконец, в функции Visual() происходит отрисовка сформированной диаграммы рассеивания в окне приложения.

Рисунок 10 представлена диаграмма деятельности функции визуализации.

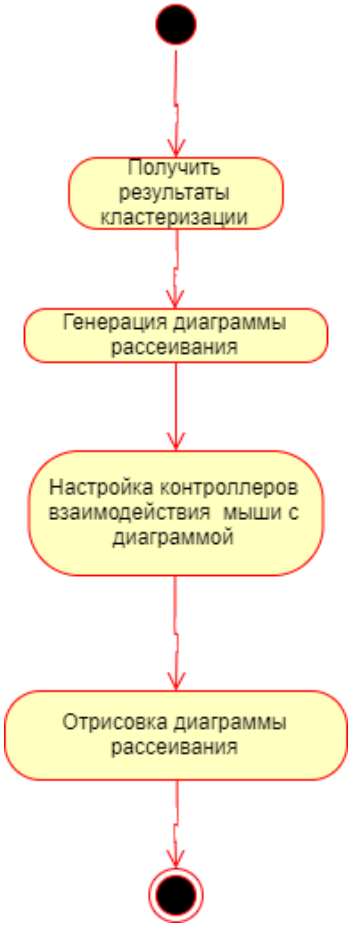

Рисунок 10 –Диаграмма деятельности функции визуализации

4 Результаты работы разработанного программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

Программное приложение для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети было реализовано и полностью удовлетворяет все выявленные к нему требования.

4.1 Общие результаты работы разработанного программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети

При запуске разработанного программного обеспечения для обработки данных ранжирования узлов социального графа при идентификации пользователей-экспертов социальной сети пользователь попадает на форму для загрузки и обработки данных показателей Боргатти, информационной энтропии, и коммуникационной эффективности из документа формата XLSX.

Показатели Боргатти, информационной энтропии, и коммуникационной эффективности должны предоставляться приложению в виде столбиков в документе формата XLSX.

Функционал для загрузки и обработки данных из документа формата XLSX в выборку пригодного для дальнейшей кластеризации содержит 3 текстовых боксов для ввода и 4 кнопки.

В текстовых боксах необходимо вводить номера столбцов в документе XLSX из которого будут взяты данные. Кнопки «Загрузить X», «Загрузить Y» и «Загрузить Z» недоступны пока в текстовых боксах, что рядом, не введут число целого типа. Сами кнопки «Загрузить X», «Загрузить Y» и «Загрузить Z» реализуют загрузку показателей Боргатти, информационной энтропии, и коммуникационной эффективности из документа формата XLSX.

После загрузки данных X, Y и Z, которые являются загружанными показателями, кнопка «Создать выборку из загруженных XYZ» становится доступной. Выполняет кнопка «Создать выборку из загруженных XYZ» действия соответствующие названию. После нажатия на кнопку «Создать выборку из загруженных XYZ» появляется окно для кластеризации данных, и созданная выборка загруженных данных передается ему, а форма загрузки данных закрывается.

Рисунок 11 демонстрирует интерфейс пользователя формы, предназначенной для загрузки и обработки в выборку, пригодную для дальнейшей кластеризации, показателей Боргатти, информационной энтропии, и коммуникационной эффективности из документа формата XLSX.

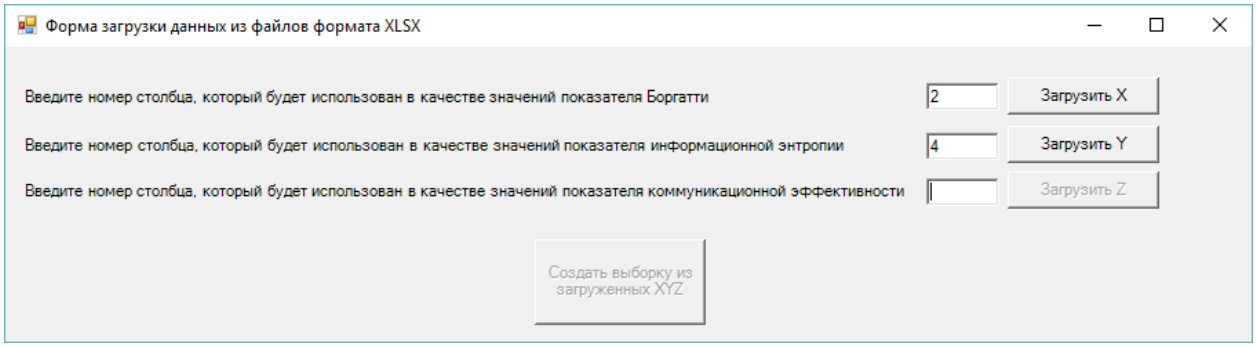

Рисунок 11 - Пользовательский интерфейс формы загрузки данных из файла формата XLSX

Окно для кластеризации данных разделено на 2 части: одна часть предназначена для визуализации данных, вторая часть содержит 3 кнопки запуска кластеризации тем или иным способом и информационный блок, который заполняется такой информацией по результатом кластеризации, как время выполнения, силуэт результатов кластеризации, примененный алгоритм кластеризации.

Также во второй части окна для кластеризации находится кнопка для расчета V-меры. Пока не проведена ни одна кластеризация, данная кнопка недоступна. Нажав на кнопку расчета V-меры, пользователю приложения откроется диалоговое окно, в котором он должен будет загрузить данные о

истинных метках классов пользователей социальной сети по тому же принципу, как он загружал показатели авторитетности, описанный выше. Загруженные истинные метки классов тут же будут использованы в расчете V-меры, а результат расчета выведен на экран.

Рисунок 12 демонстрирует интерфейс пользователя окна кластеризации.

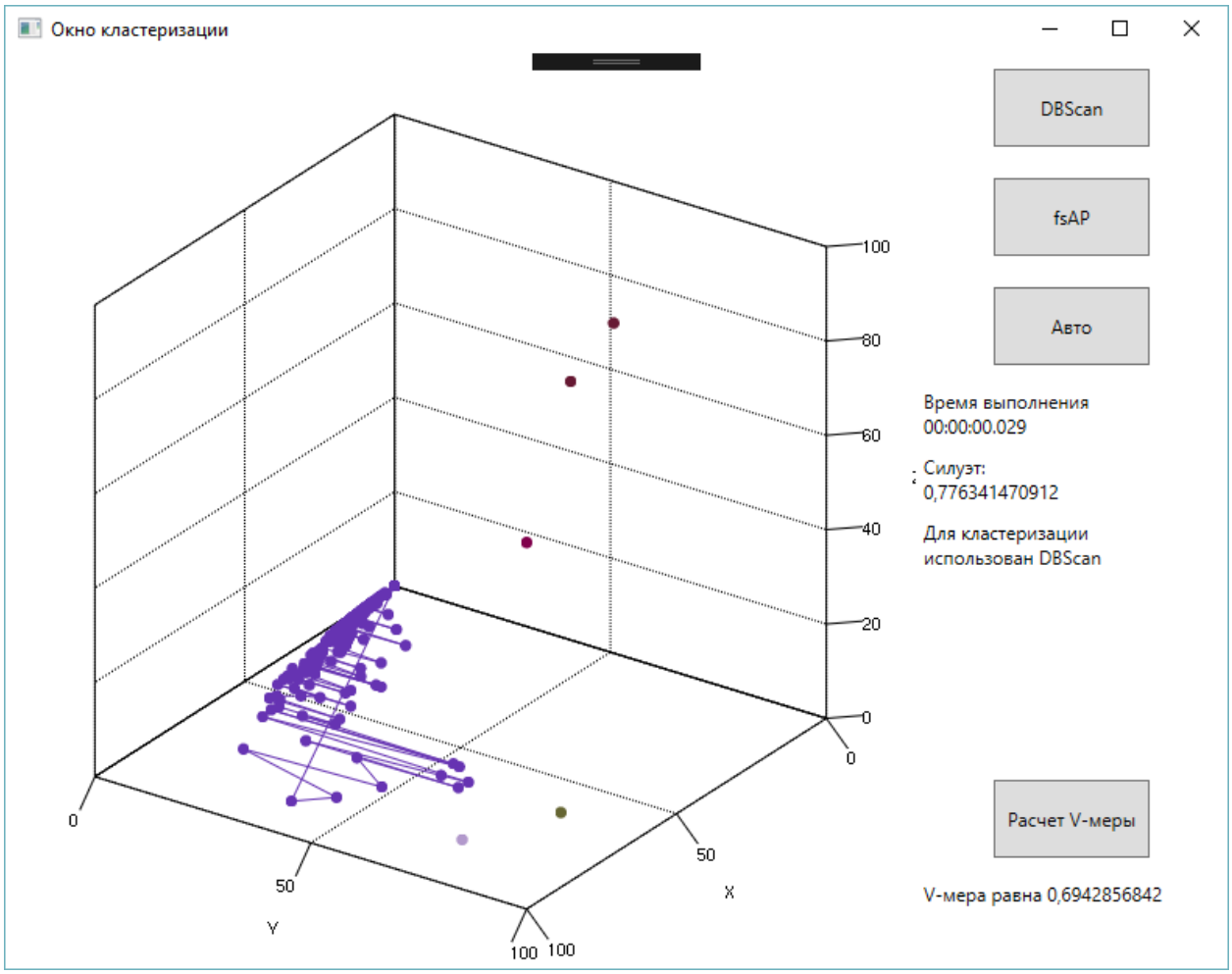

Рисунок 12 – Пользовательский интерфейс окна кластеризации

4.2 Анализ качества обучения классификатора дерева решений применяемого к определению наиболее эффективного метода кластеризации в зависимости от исходных данных

Для обученного классификатора дерева решений, который используется для определения наиболее эффективного метода калстерного анализа при автоматической кластеризации, рассчитаны все показатели

эффективности обучения и меры оценки качества обучения бинарного классификатора, упомянутые в главе 2 в разделе 3.4.

Рассчитанные показатели эффективности обучения и меры оценки качества обучения бинарного классификатора приведены в Таблица 3.

Таблица 3 - Рассчитанные показатели эффективности обучения и меры оценки качества обучения бинарного классификатора обученного классификатора дерева решений

Показатели соотношения распределения по классом и истинных меток	
TP	2
FP	2
FN	3
TN	293
Показатели эффективности обучения	
TPR	0,4
TNR	0,9932
FPR	0,0068
FNR	0,6
PPV	0,5
NPV	0,9898
Меры оценки качества обучения бинарного классификатора	
ACC	0,9833
F-мера	0,4444
MCC	0,4388

Из Таблица 3 можно сделать вывод, что обучен классификатор приемлемо. В будущем необходимо переобучить классификатор, повысив его качество обучения.

4.3 Анализ качества применения алгоритмов кластеризации DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети

Для вычисления эффективности применения алгоритмов кластеризации DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети взяты размеченные данные - так называемые графы с «писателями» и «читателями». В каждой выборке всем точкам заранее присвоены истинные метки одного из двух классов: метка писателя (в дальнейшем П) и читателя (в дальнейшим Ч). Писатели – это узлы графа, которые эмитируют влиятельных-активных пользователей в сети. Читатели – это узлы графа которые читают «писателей» с разной активностью.

Всего использовано пять выборок: выборка с 3 писателями и 500 читателями (3П500Ч), выборка с 5 писателями и 200 читателями (5П200Ч), выборка с 10 писателями и 300 читателями (10П300Ч), выборка с 10 писателями и 200 читателями (10П200Ч) и выборка с 20 писателями и 200 читателями (20П200Ч).

Вычисления эффективности применения алгоритмов кластеризации DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети представлены в приложение Приложение Е.

Расчет среднего показателя силуэта при автоматической кластеризации:

$$S = \frac{1 + 1 + 0{,}9975 + 0{,}9956 + 0{,}9966}{5} = 0{,}998$$

Расчет среднего показателя V-меры при автоматической кластеризации:

$$V = \frac{0{,}9176 + 0{,}8538 + 0{,}7160 + 0{,}6981 + 0{,}6481}{5} = 0{,}77$$

Основываясь на полученных значениях показателей качества для оценки результатов применения методов кластерного анализа, можно сделать вывод, что применения алгоритмов DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети путем автоматической кластеризации является верным решением, так как показатели говорят о качественной кластеризации данных.

5 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Целью раздела «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» является определение перспективности и успешности данной диссертационной работы, разработка механизма управления и сопровождения конкретных проектных решений на этапе реализации.

5.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности

5.1.1 Потенциальные потребители результатов исследования

Цель данной диссертационной работы - создание программного продукта, реализующего автоматизированный подбор алгоритма кластеризации данных в задаче идентификации пользователей-экспертов в социальных сетях в заданной предметной области.

Используя данный программный продукт, потребитель получает группу самых авторитетных пользователей определенной социальной сети в предметной области, заданной непосредственно самим потребителем программы. Полученный список людей можно использовать для разных целей таких, как: поиск высококвалифицированных работников с глубоким пониманием заданной предметной области; поиск авторитетных людей, к чьему мнению в заданной предметной области прислушиваются, для предложения дать определенной организации рекламу; выбор информации в заданной предметной области от самых авторитетных авторов; и т.п.

Использовать данный программный продукт могут, как коммерческие организации, например, для поиска сотрудников или рекламодателей, так и физические лица, например, для использования данной информации в собственных научно-исследовательских работах.

В коммерческих организациях подобную информацию чаще используют в специализированных отделах (отдел кадров, отдел маркетинга, отдел исследований и т.д.), которых в мелких организациях просто не

существует. Следовательно, основная направленность на предприятия средних и крупных размеров. В мелких организациях рассматриваемый программный продукт также может использоваться для тех же целей, но реже.

Так как физическим лицам подобный программный продукт нужен будет в основном для исследований, следовательно, физическое лицо должно быть связано с наукой и иметь или получать высшее образование.

5.1.2 QuaD

Технология QuaD (QUality ADvisor) представляет собой гибкий инструмент измерения характеристик, описывающих качество новой разработки и ее перспективность на рынке и позволяющие принимать решение целесообразности вложения денежных средств в научно-исследовательский проект.

В основе технологии QuaD лежит нахождение средневзвешенной величины выделенных определенных показателей.

В соответствии с технологией QuaD каждый показатель оценивается экспертным путем по стобалльной шкале, где 1 – наиболее слабая позиция, а 100 – наиболее сильная. Веса показателей, определяемые экспертным путем, в сумме должны составлять 1.

Оценка перспективности и качество по QuaD технологии определяется по формуле:

$$П_{ср} = \sum Вi Бi \quad (35)$$

Где

$П_{ср}$ – средневзвешенное значение показателя качества и перспективности научной разработки,

Вi – вес показателей (в долях единицы),

Бi – средневзвешенное значение i-го показателя.

Таблица 4 - технология QuaD

Критерии оценки	**Вес**	**Баллы**	**макс. Балл**	**Отн. Знач**	**Ср.-взвеш. Значение**
Показатели оценки качества разработки					
Автоматизация процесса кластеризации	0,3	70	100	0,7	0,21
Получение качественного результата кластеризации	0,25	70	100	0,7	0,175
Простота использование	0,2	80	100	0,8	0,16
Ремонтопригодность	0,15	80	100	0,8	0,12
Показатели оценки коммерческого потенциала разработки					
Перспективность рынка	0,05	70	100	0,7	0,035
Доступность	0,03	80	100	0,8	0,024
Законченность работы	0,02	60	100	0,6	0,012
				Итог:	0,736

Полученная оценка, равная 0,736, соответствует проекту, перспективность которого выше среднего. При дальнейшей модернизации проекта этот показатель может быть увеличен.

5.1.3 SWOT-анализ

SWOT – Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы) – представляет собой комплексный анализ научно-исследовательского проекта. SWOT-анализ применяют для исследования внешней и внутренней среды проекта.

Он проводится в несколько этапов.

Первый этап заключается в описании сильных и слабых сторон проекта, в выявлении возможностей и угроз для реализации проекта, которые проявились или могут появиться в его внешней среде.

Второй этап состоит в выявлении соответствия сильных и слабых сторон научно-исследовательского проекта внешним условиям окружающей среды. Это соответствие или несоответствие должны помочь выявить степень необходимости проведения стратегических изменений. В рамках второго этапа были составлены таблицы 2-5.

В рамках третьего этапа составляется итоговая матрица SWOT-анализа представленная в таблице 6.

Таблица 5 - Интерактивная матрица сильных сторон и возможностей

		Сильные стороны проекта				
Возможности проекта		С1	С2	С3	С4	С5
	В1	-	-	-	-	-
	В2	-	-	-	-	-
	В3	-	-	-	+	-

Таблица 6 - Интерактивная матрица слабых сторон и возможностей

		Слабые стороны проекта				
Возможности проекта		Сл1	Сл2	Сл3	Сл4	Сл5
	В1	+	-	-	0	-
	В2	+	-	--	+	-
	В3	-	+	+	-	-

Таблица 7 - Интерактивная матрица сильных сторон и угроз

		Сильные стороны проекта				
Угрозы проекта		С1	С2	С3	С4	С5
	У1	+	-	-	-	-
	У2	+	-	-	-	-
	У3	-	-	+	-	-
	У4	-	-	-	-	-
	У5	-	-	-	-	-

Таблица 8 - Интерактивная матрица слабых сторон и угроз

		Слабые стороны проекта				
Угрозы проекта		Сл1	Сл2	Сл3	Сл4	Сл5
	У1	+	+	+	+	-
	У2	-	-	-	-	-
	У3	-	-	-	+	-
	У4	-	-	-	-	+
	У5	+	+	+	+	-

Таблица 9 - Матрица SWOT

	Сильные стороны: С1. Уникальный функционал. С2. Автоматизация процесса кластеризации. С3. Широкая целевая аудитория пользователей. С4. Легко масштабируемая архитектура приложения. С5. Учет особенностей кластеризации данных пользователей социальной сети.	Слабые стороны: Сл1. Низкая скорость работы программного продукта. Сл2. Ограниченные функциональные возможности. Сл3. Узкая специализация применения разработки. Сл4. Система не оптимальна на данном этапе. Сл5. Необходимость финансирования для дальнейшей разработки.
Возможности: В1. Усовершенствование алгоритма автоматической кластеризации. В2. Устранение функциональных недостатков программного приложения. В3. Расширение функционала.	В3С4 – Легко масштабируемая архитектура приложения позволяет быстро и без сложностей внедрять в приложение новый функционал.	В1Сл1 – Усовершенствование алгоритма автоматической кластеризации ускорит ее выполнение. В2Сл1Сл4 – Устранение функциональных недостатков программного приложения приведет к повышению таких показателей приложения, как скорость выполнения алгоритма и оптимальность системы. В3Сл2Сл3 – расширение функционала поможет расширить границы применения приложения.

Продолжение таблицы 9

Угрозы: У1. Отсутствие интереса пользователей к продукту. У2. Появление конкурентных приложений с тем же функционалом. У3. Несовместимость продукта с какими-то операционными системами. У4. Отсутствие финансирования для дальнейшего развития продукта. У5. Негативные отзывы о продукте.	У1У2С1 – уникальность разработки не даст продукту остаться незамеченным, а также из-за уникальности данный продукт не с чем сравнивать, а значит люди не будут оставлять плохие отзывы. У3С3С4 – Приложение имеет широкую целевую аудиторию, поэтому, если несовместимость продукта с некоторыми операционными системами приведет к суживанию круга пользователей, количество пользователей все равно будет высоко.	У1У5Сл1Сл2Сл3Сл4 – Технические и функциональные недостатки продукта отпугнут пользователей, и заставит их оставить неприятный отзыв, поэтому необходимо как можно быстрее исправить недостатки системы У3Сл4 – неоптимальность системы может привести к несовместимости продукта с определенными ОС, нужно быстрее оптимизировать систему У4Сл5 – Отсутствие финансирования не даст продукту развиваться, поэтому необходимо выпустить первую версию ПО для привлечения инвестиций.

5.1.4 Оценка готовности проекта к коммерциализации

На какой бы стадии жизненного цикла не находилась научная разработка полезно оценить степень ее готовности к коммерциализации и выяснить уровень собственных знаний для ее проведения (или завершения).

Для этого была заполнена специальная форма, содержащая показатели о степени проработанности проекта с позиции коммерциализации и компетенциям разработчика научного проекта, Приведенная в таблице 7.

Таблица 10 - Бланк оценки степени готовности научного проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1.	Определен имеющийся научно-технический задел	4	4
2.	Определены отрасли и технологии для предложения на рынке	4	3
3.	Определена товарная форма научно-технического задела для представления на рынок	3	4
4.	Определены авторы и осуществлена охрана их прав	4	3
5.	Проведена оценка стоимости интеллектуальной собственности	1	2
6.	Проведены маркетинговые исследования рынков сбыта	2	3
7.	Разработан бизнес-план коммерциализации научной разработки	1	2
8.	Определены пути продвижения научной разработки на рынок	3	4
9.	Разработана стратегия (форма) реализации научной разработки	4	4
10.	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	1	1
11.	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	1	1
12.	Проработаны вопросы финансирования коммерциализации научной разработки	4	3
13.	Имеется команда для коммерциализации научной разработки	1	2
14.	Проработан механизм реализации научного проекта	5	5
	ИТОГО БАЛЛОВ	38	41

Исходя из результатов таблицы 7 можно сказать, что степень готовности и научной разработки, и ее разработчика к коммерциализации, находится на среднем уровне.

5.1.5 Методы коммерциализации результатов научно-технического исследования

Время продвижения товара на рынок во многом зависит от правильности выбора метода коммерциализации. Задача данного раздела магистерской диссертации – это выбор метода коммерциализации объекта исследования и обоснование его целесообразности.

Для данного проекта в качестве метода его коммерциализации научной разработке был подобран метод передачи ноу-хау компаниям, которые занимаются сбором и анализом данных.

Полученный программный продукт предполагает анализ предоставляемых ему данных. Однако, данный продукт не имеет никаких средств для сбора необходимых данных. Компании, которые занимаются сбором и анализом данных, уже имеют собственные разработки по сбору данных и предоставлению их в нужной форме.

Таким образом, компании, занимающиеся анализом и сбором данных, могут полностью раскрыть потенциал разработанного программного продукта.

5.2 Инициация проекта

В данном разделе составляется устав научного проекта магистерской работы.

Устав проекта документирует бизнес-потребности, текущее понимание потребностей заказчика проекта, а также новый продукт, услугу или результат, который планируется создать.

В Уставе проекта закрепляется такая информация как: изначальные цели и содержание; изначальные финансовые ресурсы; внутренние и внешние

заинтересованные стороны проекта, которые будут взаимодействовать и влиять на общий результат научного проекта.

5.2.1 Цели и результат проекта

В данном разделе необходимо привести информацию о заинтересованных сторонах проекта, иерархии целей проекта и критериях достижения целей.

Информацию по заинтересованным сторонам проекта представлена в таблице ниже:

Таблица 11 - Заинтересованные стороны проекта

Заинтересованные стороны проекта	**Ожидания заинтересованных сторон**
НИ ТПУ	Разработанная методика и разработанное корректно работающее ПО, что поможет университету далее более глубинно изучать работу с данными пользователей социальной сети
Научное сообщество	Любой результат исследований, как отрицательный, так и положительный. Необходимо для понимания сферы работы с данными пользователей социальной сети.

В таблице 9 представлена информация о иерархии целей проекта и критериях достижения целей. Цели проекта включают цели в области ресурсоэффективности и ресурсосбережения.

Таблица 12 - Цели и результат проекта

Цели проекта:	Создание программного продукта, реализующего автоматизированный подбор алгоритма кластеризации данных в задаче идентификации пользователей-экспертов в социальных сетях в заданной предметной области.
Ожидаемые результаты проекта:	Описанный программный продукт создан и работает. Автоматический подбор алгоритма кластеризации подпитано соответствующими исследованиями.
Критерии приемки результата проекта:	При создании описанного программного продукта учтены и выполнены все требования к проекту.
Требования к результату проекта:	**Требование:**
	Необходимо провести исследования в задаче автоматического пробора алгоритма кластеризации данных в задаче идентификации пользователей-экспертов в социальных сетях в заданной предметной области.
	Созданный программный проект должен реализовывать разработанную методику.
	Программный продукт должен производить качественную кластеризацию данных пользователей социальных сетей.
	Программный продукт должен производить автоматический подбор алгоритма кластеризации данных пользователей социально сети.

5.2.2 Организационная структура проекта

На данном этапе работы необходимо решить следующие вопросы: кто будет входить в рабочую группу данного проекта, определить роль каждого участника в данном проекте, а также прописать функции, выполняемые каждым из участников и их трудозатраты в проекте.

Данная информация представлена в таблице 10.

Таблица 13 – Рабочая группа проекта

№ п/п	ФИО, основное место работы, должность	Роль в проекте	Функции	Трудо-затраты, час
1	Доцент ОИТ ИШТР НИТПУ Лунева Е.Е.	Руководитель проекта	Координация деятельности участников проекта	25
2	Магистр ОИТ ИШИТР НИТПУ Виноградова Д.А.	Исполнитель проекта	Исполнение всех работ по проекту	558

5.2.3 Ограничения и допущения проекта.

Ограничения проекта – это все факторы, которые могут послужить ограничением степени свободы участников команды проекта, а так же «границы проекта» - параметры проекта или его продукта, которые не будут реализованных в рамках данного проекта.

Данная информация отображена в таблице ниже:

Таблица 14 – Ограничения проекта

Фактор	Ограничения/ допущения
3.1. Бюджет проекта	135 970,39руб
3.1.1. Источник финансирования	НИ ТПУ
3.2. Сроки проекта:	6.02.2019 - 6.06.2019
3.2.1. Дата утверждения плана управления проектом	6.02.2019
3.2.2. Дата завершения проекта	6.06.2019
3.3. Прочие ограничения и допущения*	Ограниченный рабочий график научного руководителя и магистранта.

5.3 Планирование управления научно-техническим проектом

5.3.1 План проекта

В рамках планирования научного проекта построен календарный график проекта, представленный в таблице 12.

Приняты следующие сокращения:

НР – научный руководитель, Лунева Е.Е.

И – исполнитель, Виноградова Д.А.

Таблица 15. Календарный план проекта в рабочих днях

Код работы (из ИСР)	**Название**	**Длительность, раб. дни**	**Дата начала работ**	**Дата окончания работ**	**Состав участников**
1	Определение целей и задач	4	6.02.2019	9.02.2019	НР
2	Составление и утверждение технического задания	9	11.02.2019	20.02.2019	НР, И
3	Разработка календарного плана	7	21.02.2019	1.03.2019	НР, И
4	Подбор и изучение материалов по теме	24	2.03.2019	31.03.2019	И
5	Обсуждение литературы	1	1.04.2019	1.04.2019	НР, И
6	Процесс реализации приложения	33	2.04.2019	15.05.2019	И
7	Оформление расчетно-пояснительной записки	15	16.05.2019	1.06.2019	И
8	Оформление графического материала	2	3.06.2019	4.06.2019	НР, И
9	Подведение итогов	2	5.06.2019	6.06.2019	НР, И
ИТОГ:		97	6.02.2019	6.06.2019	-

В приложении Ж приведена диаграмма Ганта.

Диаграмма Ганта – это тип столбчатых диаграмм (гистограмм), который используется для иллюстрации календарного плана проекта, на

котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

5.3.2 Бюджет научного исследования

Целью данного пункта является расчет величины расходов на выполнение проекта. Определение общих затрат производится путем суммирования расходов по следующим статьям:

- Материальные затраты на сырье, материалы, покупные изделия и полуфабрикаты;
- Затраты на специальное оборудование для научных работ;
- Основная заработная плата исполнителей проекта;
- Дополнительная заработная плата исполнителей проекта;
- Отчисления во внебюджетные фонды (страховые отчисления);
- Затраты на научные и производственные командировки;
- Контрагентные расходы;
- Накладные расходы.

5.3.2.1 Материальные затраты на сырье, материалы, покупные изделия и полуфабрикаты

Данная статья включает стоимость всех материалов, используемых при разработке проекта.

Затраты по данной статье расходов отсутствуют.

5.3.2.2 Затраты на специальное оборудование для научных работ

В данную статью включают все затраты, связанные с приобретением специального оборудования, необходимого для проведения работ по данному проекту.

Специальным оборудованием для выполнения данного проекта являются персональные электронные вычислительные машины (ПЭВМ).

Используемое ПЭВМ является неновым. Его стоимость учтена в виде амортизационных отчислений.

Амортизационные отчисления – инструмент компенсации полученного износа. Величина отчислений зависит от первоначальной стоимости используемого оборудования и от выбранной на предприятии амортизационной стратегии. Будем считать, что на НИТПУ вычисление амортизационных отчислений осуществляется линейным способом. Линейная стратегия амортизации подразумевает равномерное накопление амортизационных отчислений во время эксплуатации оборудования. Амортизационные отчисления при линейной амортизации рассчитывается следующим образом:

$$АО = ПСО * НА \quad (36)$$

Где

АО – сумма амортизационных отчислений, руб.;

ПСО – первоначальная стоимость оборудования, руб.;

НА – норма амортизации.

Норма амортизации рассчитывается исходя из срока полезного использования оборудования по следующей формуле:

$$НА = \frac{1}{СПИО} * 100\% \quad (37)$$

Где

СПИО - срок полезного использования оборудования, мес.

Срок полезного использования ПЭВМ составляет 3 года. Первоначальная стоимость ПЭВМ равна 40 тыс.руб.

$$НА = \frac{1}{3 * 365} * 100\% \approx 0{,}091\%$$

Ежедневные амортизационные отчисления равны:

$$А = 40\,000 * 0{,}00091 = 36{,}4 \text{ руб.}$$

Амортизационные отчисления за 93 рабочих дня:

$$36{,}4 * 93 = 3\,385{,}2 \text{ руб.}$$

5.3.2.3 Заработная плата исполнителей проекта

В настоящую статью включается основная заработная плата всех участников проекта, кто принимал участие в выполнении работ по данному проекту. Величина расходов по заработной плате определяется исходя из трудоемкости выполняемых работ и действующей системы оплаты. В состав основной заработной платы включается премия, выплачиваемая ежемесячно из фонда заработной платы.

$$С_{зп} = З_{осн} + З_{доп} \quad (38)$$

Где

$С_{ЗП}$ - статья заработной платы, руб.;

$З_{осн}$ – основная заработная плата, руб.;

$З_{доп}$ – дополнительная заработная плата, руб.

Основная заработная плата ($З_{осн}$) работника рассчитывается по следующей формуле (4):

$$З_{осн} = З_{дн} \cdot Т_р \quad (39)$$

Где

$З_{осн}$ – основная заработная плата работника, руб.;

$Т_р$ – продолжительность работ, выполняемых научно-техническим работником, раб.дн.;

$З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м \cdot М}{F_д} \quad (40)$$

Где

$З_м$ – месячный должностной оклад работника, руб;

М – количество месяцев работы без отпуска в течение года:

При отпуске в 48 раб. дней М = 10,4 месяца, 6-дневная неделя;

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб.дн.

Таблица 16 - Баланс рабочего времени

Показатели рабочего времени	**Руководитель проекта**	**Исполнитель проекта**
Календарное число дней	365	365
Количество нерабочих дней - выходные дни - праздничные дни	 52 14	 52 14
Потери рабочего времени - отпуск - невыходы по болезни	48	48
Действительный годовой фонд рабочего времени	251	251

Месячный должностной оклад работника:

$$З_М = З_Б \cdot (1 + k_{пр} + k_Д) \cdot k_р \quad (41)$$

Где

$З_Б$ – базовый оклад, руб.;

$k_{пр}$ – премиальный коэффициент, ;

$k_д$ – коэффициент доплат и надбавок, не предусмотрен;

$k_р$ – районный коэффициент, равный 1,3 (для Томска).

Основная заработная плата научного руководителя рассчитывается на основании отраслевой оплаты труда. Отраслевая система оплаты труда в НИТПУ предполагает следующий состав заработной платы:

- оклад – определяется предприятием. В ТПУ оклады распределены в соответствии с занимаемыми должностями, например, ассистент, старший преподаватель, доцент, профессор;
- стимулирующие выплаты – устанавливаются руководителем подразделений за эффективный труд, выполнение дополнительных обязанностей и т.д.;
- иные выплаты: районный коэффициент.

Руководителем данной научно-исследовательской работы является сотрудник с должностью доцент. Оклад доцента составляет 33 664 рубля.

Оклад магистранта – 12 664 рублей.

Расчёт основной заработной платы приведён в таблице

Таблица 17 - Расчёт основной заработной платы

Исполнители	**$З_б$, руб.**	**$k_{пр}$**	**$k_д$**	**$k_р$**	**$З_м$, руб**	**$З_{дн}$, руб.**	**$Т_р$, раб. дн.**	**$З_{осн}$, руб.**
Руководитель проекта	33 664	0,3	0,2	1,3	65 644,8	2 719,9	4,2	11 423,58
Исполнитель проекта	12 664	-	-	1,3	16 463,2	682,14	93	63 439,02

Дополнительная заработная плата рассчитывается по формуле:

$$З_{доп} = k_{доп} * З_{осн} \quad (42)$$

Где

$З_{доп}$ - дополнительная заработная плата, руб.;

$k_{доп}$ – коэффициент дополнительной зарплаты, 10%;

$З_{осн}$ – основная заработная плата, руб.

Таблица 18 - Заработная плата исполнителей проекта

Заработная плата	**Руководитель проекта**	**Исполнитель проекта**
Основная зарплата	11 423,58	63 439,02
Дополнительная зарплата	1 142,36	6 343,9
Зарплата	12 565,94	69 782,92
Итого по статье $С_{зп}$	82 348,86	

5.3.2.4 Отчисления во внебюджетные фонды (страховые отчисления)

Статья включает в себя отчисления во внебюджетные фонды.

Отчисления во внебюджетные фонды рассчитывается следующим образом:

$$С_{ВбФ} = k_{ВбФ} * (З_{осн} + З_{доп}) \quad (43)$$

Где

$К_{ВбФ}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.), 27,1%

$$С_{ВбФ} = 0{,}271 * 82\ 348{,}86 = 22\ 234{,}19 \text{ руб.}$$

5.3.2.5 Затраты на научные и производственные командировки

В эту статью включаются расходы по командировкам научного и производственного персонала, связанного с непосредственным выполнением конкретного проекта.

Затраты по данной статье расходов отсутствуют.

5.3.2.6 Контрагентные расходы

Контрагентные расходы включают затраты, связанные с выполнением каких-либо работ по теме сторонними организациями (контрагентами, субподрядчиками).

Расчет величины этой группы расходов зависит от планируемого объема работ и определяется из условий договоров с контрагентами или субподрядчиками.

При выполнении проекта затраты по данной статье расходов относится использование Internet исполнителем проекта.

По договору оплата услуги Internet составляет 12 руб./дн. Исполнитель проекта выполняет работу в течении 96 дней. Рассчитываем Размер контрагентных расходов:

$$С_{КР} = 12 * 86 = 1\ 032 \text{ руб.}$$

5.3.2.7 Накладные расходы

Накладные расходы составляют 30-100% от суммы основной и дополнительной заработной платы, работников, непосредственно участвующих в выполнение проекта.

Расчет накладных расходов ведется по следующей формуле:

$$С_{НР} = k_{НР} * (З_{осн} + З_{доп}) \quad (44)$$

Где

$k_{НР}$ – коэффициент накладных расходов, 30%.

$$С_{ВбФ} = 0{,}3 * 82\,348{,}86 = 24\,704{,}66 \text{ руб.}$$

5.3.2.8 Прочие прямые затраты

В данном пункте рассчитываются затраты на электричество за время работы над проектом. Вычисление производится по следующей формуле:

$$C_{эл.об} = Р_{об} * t_{об} * Ц_Э \quad (45)$$

где $Р_{об}$ - мощность, потребляемая оборудованием, кВт, равно приблизительно 700 Вт;

$t_{об}$ – время работы оборудования, час;

$Ц_Э$ - тариф на 1 кВт*час.

Для ТПУ тариф одного кВт*час равен 5,8 рублей.

Тогда расчет потребленной электроэнергии за время работы над проектом выглядит следующим образом:

$$C_{накл} = 0{,}7 * 558 * 5{,}8 = 2\,265{,}48 \text{ руб.}$$

5.3.2.9 Формирование бюджета затрат проект

Рассчитанная величина затрат научно-исследовательской работы является основой для формирования бюджета затрат проекта, который при формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции.

На основании полученных данных по отдельным статьям затрат составляется калькуляция плановой себестоимости проекта. Определение бюджета затрат на научно-исследовательский проект приведен в таблице ниже:

Таблица 19 - Калькуляция затрат по статьям

Наименование статьи	Сумма, руб
2. Затраты на специальное оборудование	3 385,2
3. Затраты по заработной плате исполнителей проекта	82 348,86
5. Отчисления во внебюджетные фонды	22 234,19
6. Прочие прямые затраты	2 265,48
7. Контрагентные расходы	1 032
8. Накладные расходы	24 704,66
ИТОГО:	**135 970,39**

5.3.3 Реестр рисков проекта

Идентифицированные риски проекта включают в себя возможные неопределенные события, которые могут возникнуть в проекте и вызвать последствия, которые повлекут за собой нежелательные эффекты.

Реестр рисков проекта представлен в приложение З.

5.4 Заключение главы финансовый менеджмент, ресурсоэффективность и ресурсосбережение

В процессе выполнения данного раздела диссертационной работы магистранта были определены перспективность и успешность данной работы, а также были разработан механизм управления и сопровождения конкретных проектных решений на этапе реализации.

Подробно были произведены:

- Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности;
- Инициация проекта;
- Планирование управления научно-техническим проектом.

Получены следующие результаты:

- Определены цели и результаты проекта, рабочая группа проекта, ограничения и допущения проекта;
- Составлен календарный график-план;
- Подсчитан бюджет проекта, который составляет 135 970,39 руб.;
- Выделены риски проекта.

6 Социальная ответственность

В данном разделе рассматриваются особенности организации рабочего места и рабочей среды студента при написании данной диссертационной работы магистра.

Цель данной диссертационной работы - создание программного продукта, реализующего автоматизированный подбор алгоритма кластеризации данных в задаче идентификации пользователей экспертов в социальных сетях в заданной предметной области.

Проектирование системы осуществляется в закрытом, отапливаемом и проветриваемом помещении, за рабочим столом, оснащенном персональным компьютером (ПК).

Далее будут рассмотрены факторы рабочей зоны и рабочего места, влияющие на состояние студента, а также влияние проектной деятельности на состояние окружающей среды.

6.1 Производственная безопасность

Производственная безопасность – система организационных мероприятий и технических средств, предотвращающих или уменьшающих вероятность воздействия на работающих опасных травмирующих производственных факторов, возникающих в рабочей зоне в процессе трудовой деятельности [58].

На человека в процессе трудовой деятельности могут воздействовать опасные и вредные производственные факторы (ГОСТ 12.0.003-2015 «Опасные и вредные производственные факторы. Классификация»). Вредными производственные факторы считаются, если они могут привести к заболеванию человека. Опасными производственные факторы считаются, если они могут привести к травме человека.

В таблице 1 представлены возможные вредные и опасные производственные факторы, возникающие в процессе разработки данного программного продукта за ПК.

Таблица 20 - Возможные вредные и опасные производственные факторы при разработке программного продукта за ПК

Факторы	Этап разработки			Характеристика фактора	Нормативные документы
	Проектирование	Разработка	Эксплуатация		
Отклонение показателей микроклимата (температуры и влажности воздуха)	+	+	+	Вредный	СанПиН 2.2.2/2.4.1340-03 [59] СанПиН 2.2.4.548-96 [60]
Недостаточная освещенность рабочей зоны	+	+	+	Вредный	СанПиН 2.2.2/2.4.1340-03
Опасность поражения электрическим током	+	+	+	Опасный	ГОСТ 12.1.038–82 [61]
Пожароопасность	+	+	+	Опасный	ФЗ от 22.07.2008 №123-ФЗ [62]

6.1.1 Анализ выявленных вредных факторов при разработке программного продукта.

6.1.1.1 Отклонение показателей микроклимата

Одним из главных критериев, по которому определяется «пригодность» рабочего места для постоянного пребывания человека в нем, является микроклимат окружающей среды. Выполнение определенных требований к микроклимату в производственном помещении позволяет поддерживать на рабочем месте здоровую, благоприятную для организма человека обстановку. Следует отметить, что микроклимат в производственных помещениях зависит от специфики производственного процесса и внешних

условий таких, как: условий вентиляции и отопления, категории работ, периода года.

В соответствии с пунктом 4.3 Санитарных правил микроклимат производственного помещения измеряется при помощи следующих показателей [60]:

- температура воздуха;
- температура поверхностей;
- относительная влажность воздуха;
- скорость движения воздуха;
- интенсивность теплового облучения.

При длительной работе любая вычислительная техника имеет свойства нагреваться, в том числе и ПК. Таким образом температура его поверхности повышается и может переходить за пределы благоприятной для работы, что повышает эмоциональное напряжение человека, работающего за ПК. Также, нагрев ПК во время работы за ним способствует повышению температуры воздуха вокруг. Из-за того, что во время работы человек находится в непосредственной близости от нагретого ПК, это становится вредным фактором, который стимулирует перегрев работающего и его быструю утомляемость [63].

Влажность оказывает большое влияние на терморегуляцию организма человека. Повышенная влажность при высокой температуре воздуха приводит к перегреванию организма. Пониженная температура увеличивает теплоотдачу с поверхности кожи, что может привести к переохлаждению. Низкая влажность вызывает неприятные ощущения в виде сухости слизистых оболочек дыхательных путей [64].

Выполняемые работы по разработки программного продукта по интенсивности энергозатрат попадает в категорию Iа, так как выполняются сидя и без значительных физических напряжений. Согласно СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных

помещений», показатели микроклимата на рабочих местах для работников категории работ Iа должны соответствовать значениям, представленным в таблице 3 (допустимые значения), однако более комфортны для работы условия, соответствующие оптимальным значениям, представленным в таблице 2 [60].

Таблица 21 - Оптимальные значения показателей микроклимата на рабочих местах производственных помещений согласно СанПиН 2.2.4.548-96 для категории работ Iа

Период года	**Категория работ**	**Температура воздуха, °С**	**Относительная влажность воздуха, %**	**Скорость движения воздуха, м/с**
Холодный	Iа	22 – 24	60 – 40	0,1
Теплый	Iа	21 – 23	60 – 40	0,1

Таблица 22 – Допустимые значения показателей микроклимата на рабочих местах производственных помещений согласно СанПиН 2.2.4.548-96 для категории работ Iа

Период года	**Категория работ**	**Температура воздуха, °С**	**Относительная влажность воздуха, %**	**Скорость движения воздуха, м/с**
Холодный	Iа	20 – 25	15 – 75	0,1
Теплый	Iа	21 – 28	15 – 75	0,1 – 0,2

Соблюдение оптимальных микроклиматических условий в производственном помещении обеспечивает общее и локальное ощущение теплового комфорта при минимальном напряжении механизмов терморегуляции, не вызывают отклонений в состоянии здоровья, создают предпосылки для высокого уровня работоспособности и являются предпочтительными на рабочих местах.

Для поддержания оптимальных параметров микроклимата необходимо применять системы отопления, вентиляции и увлажнения воздуха. В рабочих

помещениях, оснащённых ПК, необходимо ежедневно проводить влажную уборку и каждый час проветривать помещение.

6.1.1.2 Недостаточная освещенность рабочей зоны

Освещение рабочего места имеет большое значение и оказывает влияние на работоспособность человека, а также на его физическое состояние. Недостаточное освещение влияет на функционирование зрительного аппарата, а также на психику человека, вызывая усталость центральной нервной системы. Недостаточный уровень освещенности приводит к головным болям, снижению концентрации внимания и, как следствие, к ухудшению производительности труда.

Освещение рабочего места складывается из естественного и искусственного освещения.

Естественное освещение достигается оснащением оконных проемов, при этом коэффициент естественного освещения должен быть не ниже 1,2% в зонах с устойчивым снежным покровом и не ниже 1,5% на остальной территории. Приоритетно падение светового потока из окна на рабочее место должно располагаться с левой стороны. Окна в помещениях, где эксплуатируется вычислительная техника, преимущественно должны быть ориентированы на север и северо-восток. При этом оконные проемы должны быть оборудованы регулируемыми устройствами типа: жалюзи, занавесей, внешних козырьков и другие [59].

Искусственное освещение достигается применением люминесцентных ламп типа ЛБ и компактных люминесцентных ламп. Коэффициент пульсации при работе с ПК не должен превышать 5% [56].

Также, согласно СанПиН 2.2.2/2.4.1340-03, нужно выполнить следующие правила [59]:

- освещенность поверхности рабочего стола пользователя ПК должна быть в пределах 300 – 500 лк.;

- освещенность поверхности дисплея не должна превышать 300 лк.;
- яркость светящихся поверхностей (стекло окона, светильников) должна быть не более 200 кд/м^2.

Соблюдение вышеуказанных мер позволит избежать пагубного влияния на зрение пользователя ПК.

Освещение – получение, распределение и использование световой энергии для обеспечения благоприятных условий видения предметов и объектов. Оно влияет на настроение и общее самочувствие, определяет эффективность труда. Нерационально организованное освещение может явиться причиной травматизма: плохо освещенные опасные зоны, слепящие источники света и блики от них, резкие тени и пульсации освещенности ухудшают видимость и могут вызвать неадекватное восприятие наблюдаемого объекта [65].

Длина рассматриваемого помещения составляет 6 метров, ширина – 3 м, высота – 2,5 м. Высота рабочей поверхности 0,7 м.

Размещение светильников в помещении определяется следующими параметрами (Рисунок 13):

H – высота помещения;

h_c – расстояние светильников от перекрытия (свес);

$h_п = H - h_c$ – высота светильника над полом, высота подвеса;

$h_{рп}$ – высота рабочей поверхности над полом;

$h = h_п - h_{рп}$ – расчётная высота, высота светильника над рабочей поверхностью.

L – расстояние между соседними светильниками или рядами;

l – расстояние от крайних светильников или рядов до стены;

Оптимальное расстояние l от крайнего ряда светильников до стены рекомендуется принимать равным L/3.

Интегральным критерием оптимальности расположения светильников является величина λ = L/h, уменьшение которой удорожает устройство и обслуживание освещения, а чрезмерное увеличение ведёт к резкой неравномерности освещённости.

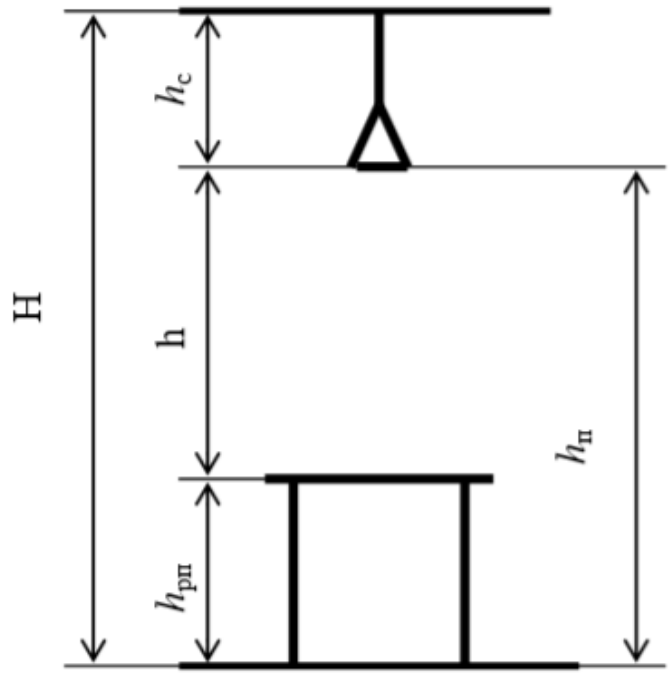

Рисунок 13 - Основные расчетные параметры

Расчет параметров освещения осуществлен для светильника ШОД – 2-40, характеристики которого приведены в таблице 4.

Таблица 23 – Характеристика светильника ШОД 2-40

Количество и мощность лампы, Вт	Размеры, мм			КПД, %
	Длина	Ширина	Высота	
2 х 40	1230	266	-	85

Исходные характеристики помещения:

H = 2,5 м; $h_{рп}$ = 0,7 м; λ = 1,1;

Тогда, расчётная высота, высота светильника над рабочей поверхностью равна: $h = h_п - h_{рп} = H – h_c – h_{рп} = 2,5 – 0,7 = 1,8$ м.

Соответственно, расстояние между соседними светильниками или рядами равно: L = λ * h = 1,1 * 1,8 = 1,98 м.

Тогда, расстояние от крайних светильников или рядов до стены:

l = L / 3 = 1,98 / 3 = 0,66 м.

Исходя из рассчитанных значений размещения светильников для люминесцентных ламп, определено, что в помещении с заданными характеристиками может быть размещено 3 светильника типа ШОД 2-40 (Рисунок 14).

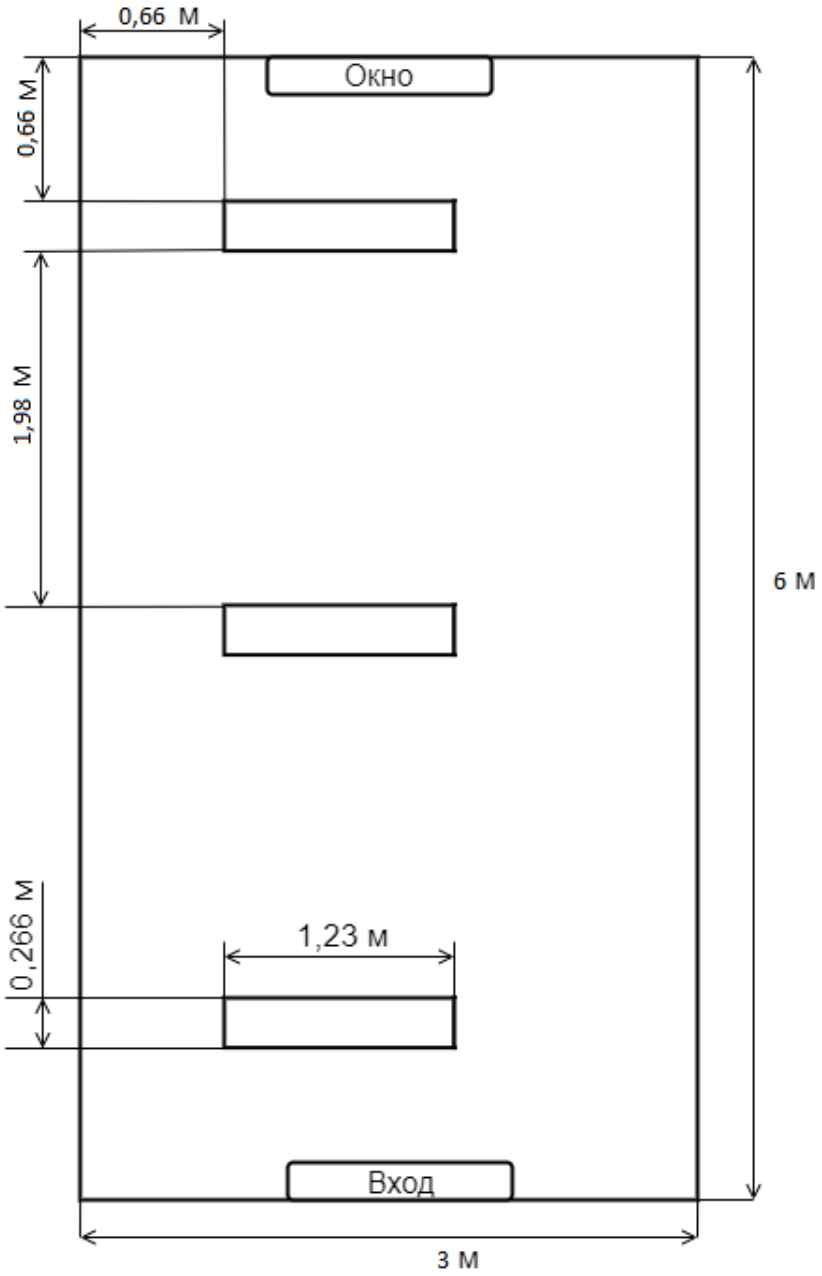

Рисунок 14 - План размещения общего освещения (вид сверху)

Работа оператора ЭВМ относится к третьему разряду зрительных работ и считается работой высокой точности. Согласно СНиП 23-05-95 норма освещенности для рассматриваемого помещения при системе комбинированного освещения оставляет 400 Лк, в том числе от общего освещения – 200 Лк [65].

Расчёт общего равномерного искусственного освещения горизонтальной рабочей поверхности выполняется методом коэффициента светового потока, учитывающим световой поток, отражённый от потолка и стен.

Световой поток лампы накаливания или группы люминесцентных ламп светильника определяется по формуле:

$$Ф = \frac{E_Н * S * К_З * Z * 100}{n * \eta} \quad (46)$$

где

$E_Н$ – нормируемая минимальная освещённость по СНиП 23-0595, лк;

S – площадь освещаемого помещения, м2;

$К_З$ - коэффициент запаса, учитывающий загрязнение светильника (источника света, светотехнической арматуры, стен и пр., т. е. отражающих поверхностей), наличие в атмосфере цеха дыма, пыли;

Z – коэффициент неравномерности освещения, отношение $E_{ср}/E_{min}$. Для люминесцентных ламп при расчётах берётся равным 1,1;

N – число ламп в помещении;

η – коэффициент использования светового потока.

Коэффициент использования светового потока показывает, какая часть светового потока ламп попадает на рабочую поверхность. Он зависит от индекса помещения i, типа светильника, высоты светильников над рабочей поверхностью h и коэффициентов отражения стен $\rho_С$ и потолка $\rho_П$.

Индекс помещения определяется по формуле:

$$i = \frac{S}{h * (A + B)} \quad (47)$$

где

h - допустимая высота подвеса светильников с люминесцентными лампами;

А – ширина;

В – длина.

Помещение имеет ширину А = 3 м, длину В = 6 м, допустимая высота подвеса светильников с люминесцентными лампами h = 1,8 м. Требуется создать освещение Е = 200 лк. Коэффициент отражения светлых стен $\rho_С$ = 30%, светлого потолка $\rho_П$ = 50%. Коэффициент запаса $К_З = 1{,}5$, коэффициент неравномерности Z = 1,1.

Определено, что в помещении можно разместить 3 ряда светильников с люминесцентными лампами, в каждом из которых можно разместить 2 светильника. С учетом того, что в каждом светильнике типа ШОД 2-40 установлено 2 люминесцентные лампы, общее число люминесцентных ламп в помещении N = 6.

Тогда индекс помещения будет равен:

$$i = \frac{S}{h*(A+B)} = \frac{3*6}{1{,}8*(3+6)} = \frac{18}{16{,}2} = 1,(11);$$

В соответствии со значениями коэффициентов отражения стен $\rho_С$ = 30%, потолка $\rho_П$ = 50% и индекса помещения $i = 1.11$ коэффициенты использования светового потока светильников с люминесцентными лампами $\eta = 35$ %.

Тогда световой поток равен:

$$Ф = \frac{Е_Н * S * К_З * Z * 100}{N * \eta} = \frac{200 * 18 * 1{,}5 * 1{,}1 * 100}{6 * 35} = 2828{,}57 \text{ Лм.}$$

Определяем потребный световой поток ламп в каждом из рядов:

Выбираем ближайшую стандартную лампу ЛХБ 40 Вт с потоком 2700 Лм. Делаем проверку выполнения условия:

$$-10\% \leq \frac{Ф_{л.станд} - Ф_{л.расч}}{Ф_{с.станд}} * 100\% \leq 20\% \qquad (48)$$

Получаем:

$$-10\% \leq \frac{2700 - 2828{,}57}{2700} * 100\% \leq 20\% = -10\% \leq -4{,}76\% \leq 20\%$$

Определяем электрическую мощность осветительной установки:

$$P = 6 * 40 = 240 \text{ Вт.}$$

6.1.2 Анализ выявленных опасных факторов при разработке программного продукта

6.1.2.1 Опасность поражения электрическим током

ПК является электрооборудование, а значит при работе на ПК есть вероятность поражения электрическим токам.

Действие электрического тока на живую ткань носит разносторонний характер. Проходя через тело человека, электрический ток производит термическое, электролитическое, механическое и биологическое воздействие. Термическое действие тока проявляется в ожогах отдельных участков тела, нагреве и повреждении кровеносных сосудов; электролитическое – в разложении органической жидкости, в том числе крови, что вызывает нарушение ее состава, а также ткани в целом; механическое - в расслоении, разрыве тканей организма; биологическое - в раздражении и возбуждении живых тканей организма, а также в нарушении внутренних биологических процессов. Например, взаимодействуя с биотоками организма, внешний ток может нарушить нормальный характер их воздействия на ткани и вызвать непроизвольные сокращения мышц [66].

Главным и определяющим фактором воздействия электрического тока на тело человека является сила тока. Предельно допустимая неощутимая величина переменного тока 0,3 мА. Эта величина считается условно безопасной [67].

Человек начинает ощущать воздействие проходящего через него тока при увеличении силы тока до 0,6 – 1,6 мА для переменного тока частотой 50 Гц и 5 – 7 мА для постоянного тока. Эти значения называются пороговыми

ощутимыми токами. Для переменного тока характер ощущения проявляется в виде пощипывания, легкого дрожания пальцев, для постоянного тока – в виде зуда, ощущения нагрева [67].

При увеличении силы тока до 10 мА для переменного тока и 50 мА постоянного тока возникает второе пороговое значение – неотпускающий или удерживающий ток. При этом происходит судорожное сокращение мышц и человек не в состоянии самостоятельно освободиться от действия тока (разжать пальцы и отпустить токопровод, за который он взялся) [67].

При значениях переменного тока 100 мА, а постоянного 300 мА его воздействие передается непосредственно на мышцу сердца. При длительности воздействия 0,5 сек может наступить остановка или фибрилляция сердца. Это третье пороговое значение токов – фибрилляционный ток [67].

Согласно ГОСТ 12.1.038-82 на рабочем месте программиста допускается уровень тока равный предельно допустимой неощутимой величине.

К поражению электрическим током при работе с ПК приводит наличии оголенных участков на кабеле, нарушении изоляции распределительных устройств и от токоведущих частей компьютера, при работе во влажной одежде или со влажными руками [61].

Для обеспечения безопасной эксплуатации электроустановок применяются различные способы защиты.

Для защиты от прямого прикосновения применяют следующие меры:

- основная изоляция токоведущих частей;
- размещение токоведущих частей вне зоны досягаемости (например, внутри системного блока).

Защиту от косвенного прикосновения обеспечивают с помощью:

- защитного заземления;
- автоматического отключения питания при скачке тока.

6.1.2.2 Пожароопасность

Во время работы за ПК риск пожара повышен, так как ПК является электрической установкой. Правильно рассчитанная, выполненная и эксплуатируемая электрическая установка не представляет пожарной опасности. Причинами, нарушающими нормальную работу установки, могут быть короткое замыкание, перегрузка проводов сети, возникновение больших переходных сопротивлений и прочее.

Пожарная и взрывная безопасность – это система организационных мероприятий и технических средств, направленная на профилактику и ликвидацию пожаров и взрывов на производстве [68].

Методы противодействия пожару на предприятии делятся на уменьшающие вероятность возникновения пожара (профилактические) и направленные непосредственно на защиту и спасение людей от огня [68].

Выделенные профилактические действия [68]:

- определение и оборудование места для курения;
- определение порядка обесточивания электрооборудования в случае пожара и по окончании рабочего дня;
- определение порядка осмотра и закрытия помещений после окончания работы;
- определение порядка и срока прохождения противопожарного инструктажа и занятий по пожарно-техническому минимуму, а также назначение ответственных за их проведение.
- организация деятельности подразделений пожарной охраны.

Выделенные защитно-спасательные действия [68]:

- разработка планов эвакуации;
- регламентирование действий при обнаружении пожара;
- устройство эвакуационных путей;
- устройство систем обнаружения пожара, оповещения и управления эвакуацией людей при пожаре;

- применение первичных средств пожаротушения;
- применение автоматических установок пожаротушения;

6.2 Экологическая безопасность

В данном подразделе рассматривается характер воздействия проектируемого решения на окружающую среду.

Сам программный продукт ни во время его проектирования, ни во время его разработки, ни во время эксплуатации не приносит никакого вреда окружающей среде. Однако главное используемое средство разработки – непосредственно ПК – может принести вред окружающей среде.

В основном весь вред окружающей среде, приносимый ПК, крутится вокруг утилизации непригодного для работы компьютера или его неисправных составляющих: монитора, клавиатуры, компьютерной мыши, материнской платы, жесткого диска, процессора, блока питания, оперативной памяти или системы охлаждения.

При неправильной утилизации старого ПК или его неисправных составляющих происходит следующее [69]:

5. Загрязнение атмосферы
    - выбросы в атмосферу углекислого газа и образование тепла при сожжении ПК;
6. Загрязнение литосферы
    - Старые компьютеры, ноутбуки во время перегнивания в течение долгих десятков лет выделяют загрязняющие окружающую среду токсические вещества, в том числе образование твердых отходов, относящихся к IV классу опасности (системный блок компьютера, клавиатура, манипулятор "мышь")
    - В любой электронной технике, в том числе и в компьютере, содержатся редкие и ценные материалы, не использование которых повторно является неоправданной тратой.

- Также в состав ПК входят некоторые элементы не подверженные разложению.

7. Загрязнение гидросферы:
    - Выделение во время перегнивания загрязняющие окружающую среду токсические вещества, в том числе жидких отходов, относящихся к IV классу опасности (системный блок компьютера, клавиатура, манипулятор "мышь")

Согласно ГОСТ Р 55102-2012 [70] персональный компьютер содержит следующие компоненты, которые должны быть раздельно собраны при выводе его из эксплуатации, представленные в таблице 5.

Таблица 24 – Перечень основных компонентов некоторых типов электротехнического и электронного оборудования

Компоненты	**Тип оборудования**	
	Системный блок и клавиатура персонального компьютера	**Монитор персонального компьютера**
Металлы	✓	
Двигатель	✓	
Пластик	✓	✓
ЭЛТ		✓
ЖК-экран		✓
Электропровода	✓	
Трансформатор	✓	
Печатные платы	✓	✓
ХИТ	✓	✓
Внешние электропровода	✓	✓

В качестве верного варианта утилизации старого ПК является передача старой техники к отходам электронной промышленности. Переработка такого рода отходов осуществляется разделением на однородные компоненты, химическим выделением пригодных для дальнейшего использования компонентов и направлением их для дальнейшего использования: кремний алюминий, золото, серебро, редкие металлы [71].

Пластмассовые части ПК утилизируются при высокотемпературном нагреве без доступа воздуха. Для опасных отходов используют теплоту сжигания. Отходы, которые не подлежат переработке, утилизации и вторичному использованию, подлежат захоронению на полигонах или в почве [71].

Также во время разработки программного продукта каждый из участников проекта в процессе жизнедеятельности наносит вред окружающей среде:

8. Загрязнение гидросферы:
    - Образование сточных вод, как последствие употребления воды. Бытовые сточные воды характеризуются присутствием минеральных и органических загрязняющих веществ. Особую форму загрязнителей представляют микроорганизмы, в т. ч. болезнетворные.

Для снижения вреда окружающей среде производится очистка сточных вод в системах канализации. Для это применяются отстаивание и фильтрация. Возможна дополнительная очистка с использованием озонаторов и ультрафиолета.

Разработка и эксплуатация системы видеомониторинга и идентификации пользователей происходит в офисном помещении. Офис является источником следующих видов отходов:

- Твердые отходы бумага, канцелярские принадлежности, комплектующие;

- Люминесцентные лампы.

Бумага может быть переработана и использована в качестве вторсырья. Для сбора макулатуры в России существуют специальные пункты приема. Некоторые из них предоставляют услугу вывоза макулатуры.

Отдельного внимания заслуживает вопрос утилизации люминесцентных ламп. Они покрыты люминесцентным веществом, имеют стеклянную оболочку и электроды. Внутри таких ламп находится инертный газ с парами ртути. В случае повреждения корпуса лампы, пары ртути попадают в атмосферу, загрязняя ее. Поэтому после окончания срока службы люминесцентные лампы необходимо сдавать на специальные предприятия по утилизации, имеющие специальную лицензию на данный вид деятельности. В Томске к таким предприятиям относятся Экотом, Полигон, Утилизация.

6.3 Безопасность в чрезвычайных ситуациях.

В данном подразделе проводится краткий анализ возможных чрезвычайных ситуаций, которые могут возникнуть при разработке, производстве или эксплуатации проектируемого решения.

6.3.1 Перечень возможных ЧС при разработке и эксплуатации проектируемого решения

Чрезвычайные ситуации могут быть техногенного, природного, биологического, социального или экологического характера [72].

При работе на рабочем месте, оснащённым ПК, можно выделить следующие техногенные чрезвычайные ситуации: взрывы, пожары, обрушение помещений, аварии на системах жизнеобеспечения [72].

Также можно выделить следующие природные чрезвычайные ситуации: наводнения, ураганы, бури, природные пожары [72].

Выделяют следующие чрезвычайные ситуации биологического характера: эпидемии, пандемии [72].

Выделяют следующие чрезвычайные ситуации социального характера: война, терроризм [72].

И наконец при работе за ПК выделяю следующие экологические чрезвычайные ситуации: разрушение озонового слоя, кислотные дожди [72].

Самой вероятной из перечисленных чрезвычайной ситуацией является ситуация техногенного характера – пожар. Причинами пожара при работе с ПК могут быть следующие:

- Не соблюдение правил пожарной безопасности;
- Не исправность проводки;
- Не исправность или неправильное использования электроприборов, в том числе ПК;
- Не исправность или неправильное использование устройств искусственного освещения.

В качестве превентивных мер нужно выполнять следующие действия [67]:

- определение и оборудование места для курения;
- определение порядка обесточивания электрооборудования в случае пожара и по окончании рабочего дня;
- определение порядка осмотра и закрытия помещений после окончания работы;
- определение порядка и срока прохождения противопожарного инструктажа и занятий по пожарно-техническому минимуму, а также назначение ответственных за их проведение.
- организация деятельности подразделений пожарной охраны.

6.3.2 Разработка действий в результате возникшей ЧС и мер по ликвидации её последствий.

Главное правило при любой чрезвычайной ситуации – не паниковать. Паника не даст рассуждать логически и приведет к неверным действиям, которые не только не способствуют спасателям помочь вам, а только помешает им это сделать.

В остальном при пожаре нужно действовать следующим образом [73]:

1. Необходимо определить для себя стоит ли выходить из здания или нет. Для этого нужно убедиться, что за дверью нет пожара, приложив свою руку к двери или к металлической ручке. Если они горячие, то ни в коем случае нельзя открывать дверь.
2. Нельзя входить туда, где большая концентрация дыма и видимость менее 10 метров.

Если дым и пламя позволяют выйти из помещения (здания) наружу:

1. Необходимо покинуть здание, используя основные и запасные пути эвакуации.
2. Попутно нужно отключать от сети электрооборудование.
3. Передвигаться нужно на четвереньках, так как вредные продукты горения скапливаются на уровне нашего роста и выше, закрывая при этом рот и нос подручными средствами защиты.
4. По пути за собой необходимо плотно закрывать двери.
5. Покинув опасное помещение, ни в коем случаи не нужно возвращаться назад.
6. Сообщите о себе должностным лицам.
7. Если на момент выхода из помещения не видно спасательных служб, необходимо вызвать их по телефону.

Если дым и пламя не позволяет выйти из здания:

1. Нужно покрыть свое тело полностью мокрым покрывалом (тканью).
2. Проверить существует ли возможность выйти на крышу или спуститься по пожарной лестнице.
3. Если возможности эвакуироваться нет, то для защиты от тепла и дыма необходимо надёжно загерметизировать комнату, в которой вы остались замкнутыми:
    - плотно закрыть входную дверь, заткнуть щели двери изнутри помещения, используя при этом любую ткань;

- закрыть окна, форточки, вентиляционные отверстия; ни в коем случае не нужно открывать и разбивать окна, так как нарушится герметичность помещения, что приведёт к увеличению температуры и площади пожара;
- перед закрытием окна можно вывесить за него большой и яркий кусок ткани, или иной приметный предмет, который сможет просигнализировать спасателям о вашем присутствии в этой комнате;
- если есть вода, нужно постоянно смачивать дверь и пол.

4. Если помещение наполнилось дымом, передвигаться нужно на четвереньках, прикрыв рот и нос влажной тряпкой (носовым платком, рукавом от рубашки), в сторону окна и находитесь возле окна.
5. Если в наличии присутствует рабочий телефон, то обязательно нужно позвонить в спасательные службы и сообщить свое местоположение.

Для самостоятельного тушения пожара нужно воспользоваться ручными углекислотными огнетушителями (типа ОУ-2, ОУ-5) или пожарным краном внутреннего противопожарного водопровода. Использование перечисленных средств возможно для тушения начальных возгораний различных веществ и материалов, за исключением веществ, горение которых происходит без доступа воздуха. Огнетушители должны постоянно содержаться в исправном состоянии и быть готовыми к действию. Категорически запрещается тушить возгорания в помещениях при помощи химических пенных огнетушителей (типа ОХП-10) [67].

6.4 Правовые и организационные вопросы обеспечения безопасности

Рабочее место, оснащенное ПК, должно быть организовано в соответствии с требованиями документов:

- ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования» [74]

- СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы».

Данные документы содержат требования к организации непосредственной рабочей зоны. Также присутствуют требования организации помещений, где данная рабочая зона непосредственно располагается.

Соблюдение требований в вышеуказанных документах позволит обеспечить подходящие условия для работы с ПК и минимизировать влияние вредных и опасных факторов. Далее представлены основные общие требования к организации рабочих мест пользователей электронно-вычислительных машин, согласно СанПиН 2.2.2/2.4.1340-03 [59]:

- При размещении рабочего места, расстояние между рабочим столом и дисплеем, должно быть не менее 2,0 м, а расстояние между боковыми поверхностями дисплея – не менее 1,2 м.
- Дисплей должен находиться от глаз пользователя на расстоянии 60 - 70 см, но не ближе 50 см с учетом размеров алфавитно-цифровых знаков и символов.
- Дисплей должен включать возможность регулировки яркости и насыщенности изображения.
- Площадь на рабочее место пользователя ПК с дисплеем на базе электронно-лучевой трубки (ЭЛТ) должна составлять не менее 6 $м^2$, с дисплеем на базе плоских дискретных экранов (жидкокристаллические, плазменные) - 4,5 $м^2$.
- Конструкция рабочего стола должна обеспечивать оптимальное размещение на рабочей поверхности используемого оборудования. Поверхность рабочего стола должна иметь коэффициент отражения 0,5 -0,7.

- Для внутренней отделки интерьера помещений, где расположены ПК, должны использоваться диффузно отражающие материалы с коэффициентом отражения для потолка - 0,7-0,8; для стен - 0,5-0,6; для пола - 0,3-0,5.
- Конструкция рабочего стула должна обеспечивать поддержание рациональной рабочей позы при работе с компьютером. Рабочий стул должен быть подъемно-поворотным, регулируемым по высоте и углам наклона сиденья и спинки.
- Поверхность сиденья, спинки и других элементов рабочего стула должна быть полумягкой, с нескользящим, слабо электризующимся и воздухопроницаемым покрытием, обеспечивающим легкую очистку от загрязнений

Помимо перечисленных требований должны выполняться все требования, упомянутые в разделе «производственная безопасность».

6.5 Заключение главы социальная ответственность

В течении разработки раздела «Социальная безопасность» были рассмотрены следующие важные аспекты деятельности, как: производственная безопасность, экологическая безопасность, безопасность в чрезвычайных ситуациях, и правовые и организационные вопросы обеспечения безопасности.

Таким образом по каждому рассмотренному пункту были созданы точные инструкции для обеспечения комфортной и безопасной деятельности студента, работающего над диссертационной работай в условиях общежития.

Заключение

Данная диссертационная работа посвящена решению задачи идентификации пользователей-экспертов социальной сети в заданной предметной области.

Была создана следующая методика, позволяющей обработать и интерпретировать данные ранжирования узлов социального графа при помощи нескольких эффективных способов для идентификации пользователей-экспертов социальной сети:

1. Взять выборку данных, содержащую показатели Боргатти, информационной энтропии, и коммуникационной эффективности всех пользователей социальной сети, кто связан с заданной предметной областью;
2. Выбирать первые 20 значимых узла взятой выборки данных и передать их обученному классификатору;
3. Получить результат классификатора, передать полную выборку данных подобранному классификатором алгоритму кластерного анализа;
4. Получить результаты кластеризации, перевести их в доступную для понимания человеком форму, выдать результат.

В качестве практической части работы было разработано программное обеспечение для идентификации пользователей-экспертов в социальной сети. Разработанное программное обеспечение соответствует всем предъявленным ему требованиям, а именно:

- Наличие средств для загрузки данных;
- Возможность кластеризации данных алгоритмом DBScan;
- Возможность кластеризации данных алгоритмом Fast Affinity Propagation;
- Возможность автоматической кластеризации данных, организованной по выведенной методике;

- Наличие средств определения качества выполненной кластеризации;
- Наличие средств визуализации результатов кластеризации.

Также был проведен анализ качества обучения классификатора дерева решений, применяемого в функции автоматической кластеризации для определения наиболее эффективного метода кластеризации в зависимости от исходных данных. Анализ показал, что классификатор обучен приемлемо.

Еще был проведен анализ качества применения алгоритмов кластеризации DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети. Анализ показал, что применения алгоритмов DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети путем автоматической кластеризации является верным решением, так как показатели говорят о качественной кластеризации данных.

Список публикаций

1. Виноградова Д. А. Исследование применимости алгоритмов кластеризации Affinity Propagation, Dbscan к решению задачи поиска пользователей-экспертов в социальных сетях // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых (Томск, 3–7 декабря 2018 г.) / Томский политехнический университет. – Томск: Изд-во ТПУ, 2019. – Сек.1 – с.160-162

Список используемых источников

1. Borgatti S.P. Identifying sets of key players in a social network // Computational & Mathematical Organization Theory. – 2006. – Vol. 12, № 1. – P. 21-34
2. Arulselvan A., Commander C., Elefteriadou L., Pardalos P., Detecting critical nodes in sparse graphs // ComputOper Res. – 2009. – № 36. – P. 2193–2200
3. Lalou M., Tahraoui M.A, Kheddouci H. The Critical Node Detection Problem in networks: A survey (Review) // Computer Science Review. – 2018. – №. 28. – P. 92–117
4. Shetty J., Adibi J. Discovering important nodes through graph entropy the case of enron email database // 3rd International Workshop on Link Discovery. – 2015. – P. 74-81
5. Borgatti S.P., Carley K.M., Krackhardt D. On the robustness of centrality measures under conditions of imperfect data // Social Networks. – 2006. – Vol. 28, № 2. – P.124-176
6. Ortiz-Arroyo D. Discovering Sets of Key Players in Social Networks // Computational Social Networks Analysis. – 2010. – P. 27-47
7. Krebs V.E. Uncloaking terrorist networks // Volume. – 2002. – Vol. 7, № 4. – P. 1-10
8. Batagelj V., Mrvar A.: Pajek datasets. – 2006, URL: http://vlado.fmf.uni-lj.si/pub/networks/data/ (дата обращения: 30.05.2019)
9. UCINET software datasets [Электронный ресурс] URL: https://sites.google.com/site/ucinetsoftware/datasets/ (дата обращения: 22.05.2019)
10. Сравнение способов идентификации пользователей социальных сетей, являющихся экспертами в заданной предметной области / Е.Е. Лунева, А.А. Ефремов, Е.А. Кочегурова, П.И. Банокин, В.С. Замятина

// Системы управления и информационные технологии. – 2017. – №4(70). – С. 63-68

11. Ефремов А. А. Использование процедуры ранжирования Кендалла–Уэя для идентификации ключевых игроков социального графа / А. А. Ефремов [и др.] // Доклады ТУСУР. – 2018. – Т. 21, № 1. – С. 80–85. DOI: 10.21293/1818-0442-2018-21-1-80-85
12. Лунева Е.Е., Замятина В.С., Банокин П.И., Ефремов А.А. Оценка влияния пользователя социальной сети в данной области знаний // Физический журнал: Серия конференций. - 2017. –Тол. 803, № 1. - С.1-6.
13. Handbook of Graph Theory / edited by J.L. Gross, J. Yellen, P. Zhang. – 2nd edition. Boca Raton: CRC Press, 2014. – 1610 p.
14. Ray S.S. Graph Theory with Algorithms and its Ap-plications: in Applied Science and Technology / S.S. Ray. – New Dehli: Springer India, 2013. – 214 p.
15. Svami M. Graphs, Networks and Algorithms / M. Svami, K. Thulasiraman – Monreal: Wiley, 1984. – 455 p.
16. Keener J.P. The Perron–Frobenius Theorem and the Ranking of Football Teams / J.P. Keener // SIAM Review. – 1993. – Vol. 35(1). – P. 80-93
17. Машинное обучение [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Машинное_обучение (дата обращения: 27.05.2019)
18. Машинное обучение [Электронный ресурс] Распознавание, URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение#.D0.A1.D0.BF.D0.B5.D1.86.D0.B8.D1.84.D0.B8.D1.87.D0.B5.D1.81.D0.BA.D0.B8.D0.B5_.D0.BF.D1.80.D0.B8.D0.BA.D0.BB.D0.B0.D0.B4.D0.BD.D1.8B.D0.B5_.D0.B7.D0.B0.D0.B4.D0.B0.D1.87.D0.B8 (дата обращения: 27.05.2019)

19. Машинное обучение для людей [Электронный ресурс] Вастрик.ру, URL: https://vas3k.ru/blog/machine_learning/#scroll90 (дата обращения: 27.05.2019)
20. Обучение с учителем [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Обучение_с_учителем (дата обращения: 27.05.2019)
21. Обучение с учителем [Электронный ресурс] Распознавание, URL: http://www.machinelearning.ru/wiki/index.php?title=Обучение_с_учителем (дата обращения: 27.05.2019)
22. Обучение без учителя [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Обучение_без_учителя (дата обращения: 27.05.2019)
23. Обучение без учителя [Электронный ресурс] Распознавание, URL: http://www.machinelearning.ru/wiki/index.php?title=Обучение_без_учителя (дата обращения: 27.05.2019)
24. Кутуков Д. С. Применение методов кластеризации для обработки новостного потока [Текст] // Технические науки: проблемы и перспективы: материалы Междунар. науч. конф. (г. Санкт-Петербург, март 2011 г.). – СПб.: Реноме, 2011. – С. 77-83. – URL https://moluch.ru/conf/tech/archive/2/207/ (дата обращения: 27.03.2018).
25. Кластерный анализ [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Кластерный_анализ (дата обращения: 27.03.2018)
26. Сенько О. В. Методы кластерного анализа проектирование данных на плоскость, метод главных компонент // лекции по курс «Математические основы теории прогнозирования» / лекция 11 – URL: http://www.machinelearning.ru/wiki/images/2/2f/MOTP14_11.pdf (дата обращения: 27.03.2018)

27. Алгоритмы семейства FOREL [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Алгоритмы_семейства_FOREL (дата обращения: 27.03.2018)
28. Алгоритм ФорЭл [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=Алгоритм_ФОРЭЛЬ (дата обращения: 27.03.2018)
29. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования, 2007 / URL: http://www.ccas.ru/voron/download/Clustering.pdf (дата обращения: 27.03.2018)
30. Обзор алгоритмов кластеризации данных [Электронный ресурс] habrahabr, URL: https://habrahabr.ru/post/101338/ (дата обращения: 27.03.2018)
31. Алгоритм кластеризации, основанный на минимальном остовном дереве [Электронный ресурс], URL: http://algowiki-project.org/ru/Участник:Guryanovak/Алгоритм_кластеризации,_основанный_на_минимальном_остовном_дереве (дата обращения: 27.03.2018)
32. Остовное дерево [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Остовное_дерево (дата обращения: 27.03.2018)
33. Минимальное остовное дерево [Электронный ресурс] Wikipedia, URL: https://ru.wikipedia.org/wiki/Минимальное_остовное_дерево (дата обращения: 27.03.2018)
34. Affinity propagation [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/Affinity_propagation (дата обращения: 27.03.2018)

35. AFFINITY PROPAGATION FAQ [Электронный ресурс], URL: http://www.psi.toronto.edu/affinitypropagation/faq.html (дата обращения: 27.03.2018)
36. Интересные алгоритмы кластеризации, часть первая: Affinity propagation [Электронный ресурс] habrahabr, URL: https://habrahabr.ru/post/321216/ (дата обращения: 27.03.2018)
37. Jia Y, Wang J, Zhang C, Hua XS (2008) Finding image exemplars using fast sparse affinity propagation. In: Proceedings of the 16th ACM international conference on multimedia (MM ’08). ACM, New York, USA, pp 639–642
38. Jia Y, Wang J, Zhang C, Hua XS (2008) Finding image exemplars using fast sparse affinity propagation. In: Proceedings of the 16th ACM international conference on multimedia (MM ’08). ACM, New York, USA, pp 639–642
39. Yasuhiro Fujiwara, Go Irie, Tomoe Kitahara Fast Algorithm for Affinity Propagation // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence P 2238 – 2243
40. DBSCAN [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/ DBSCAN (дата обращения: 27.03.2018)
41. Плотностный алгоритм кластеризации (DBSCAN) [Электронный ресурс], URL: http://algowiki-project.org/ru/Участник:Igor.orpanen/Плотностный_алгоритм_кластеризации_(DBSCAN) #PDBSCAN (дата обращения: 27.03.2018)
42. Интересные алгоритмы кластеризации, часть вторая: DBSCAN [Электронный ресурс] habrahabr, URL: https://habrahabr.ru/post/322034/ (дата обращения: 27.03.2018)
43. Королев С. Open Machine Learning Course / Topic 7. Unsupervised learning: PCA and clustering [Электронный ресурс] URL: https://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/ju

pyter_english/topic07_unsupervised/topic7_pca_clustering.ipynb?flush_cache=true (дата обращения: 10.03.2019)

44. Meila M. 2007. Comparing clusterings – an information based distance // Journal of Multivariate Analysis, 98:873– 895.
45. Rosenberg A. Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure // Department of Computer Science Columbia University New York, NY 10027, сс. 410–420
46. 2.3.9.5. Silhouette Coefficient¶ [Электронный ресурс] Scikit learn, URL: https://scikit-learn.org/stable/modules/clustering.html (дата обращения: 10.03.2019)
47. Decision Trees in C# [Электронный ресурс] César Souza, URL: http://crsouza.com/2012/01/04/decision-trees-in-c/ (дата обращения: 25.05.2019)
48. Воронцов К. В. Лекции по логическим алгоритмам классификации, 2007 / URL: http://www.ccas.ru/voron/download/LogicAlgs.pdf (дата обращения: 27.05.2019)
49. Confusion matrix [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/Confusion_matrix (дата обращения: 1.06.2018)
50. Sensitivity and specificity [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/Sensitivity_and_specificity (дата обращения: 1.06.2018)
51. Positive and negative predictive values [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values (дата обращения: 1.06.2018)
52. Accuracy and precision [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/Accuracy_and_precision (дата обращения: 1.06.2018)

53. Boughorbel S. Jarray F. El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric // PLoS ONE 12(6): e0177678 – 2017 - URL: https://doi.org/10.1371/journal. pone.0177678
54. Matthews correlation coefficient [Электронный ресурс] Wikipedia, URL: https://en.wikipedia.org/wiki/Matthews_correlation_coefficient (дата обращения: 1.06.2018)
55. Accord.NET Framework [Электронный ресурс], URL: http://accord-framework.net/ (дата обращения: 25.05.2019)
56. Nzy3d-api [Электронный ресурс] GitHub, URL: https://github.com/jzy3d/nzy3d-api (дата обращения: 25.05.2019)
57. Леоненков А. Самоучитель UML // Глава4. Диаграмма вариантов использования (use case diagram) [Электронный ресурс] Библиотека электронной литературы в формате fb2, URL: https://litresp.ru/chitat/ru/%D0%9B/leonenkov-aleksandr/samouchitelj-uml/4 (дата обращения: 31.05.2019)
58. Производственная безопасность это [Электронный ресурс] / center-yfc URL: https://center-yf.ru/data/Menedzheru/proizvodstvennaya-bezopasnost-eto.php, свободный, Яз. Русс. (дата обращения 20.05.2019)
59. СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы». Яз. Рус. (дата обращения 20.05.2019)
60. СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных помещений». Яз. Рус. (дата обращения 20.05.2019)
61. ГОСТ 12.1.038–82 «Система стандартов безопасности труда. Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов». Яз. Рус. (дата обращения 20.05.2019)
62. Федеральный закон от 22.07.2008 N 123-ФЗ (ред. От 13.07.2015) «Технический регламент о требованиях пожарной безопасности»

[Электронный ресурс] / КонсультантПлюс. URL: http://www.consultant.ru/document/cons_doc_LAW_78699/, свободный. Яз. Рус. (дата обращения 20.05.2019)

63. Ефремова О. С. Требования охраны труда при работе на персональных электронно-вычислительных машинах. – 2-е изд., перераб. и доп. – М.: Издательство «Альфа-Пресс», 2008. Яз. Рус. (дата обращения 20.05.2019)
64. Назаренко О. Б. Безопасность жизнедеятельности: учебное пособие / О. Б. Назаренко, Ю. А. Амелькович; Томский политехнический университет. – 3-е изд., перераб. и доп. – Томск: Изд-во ТПУ, 2013. Яз. Рус. (дата обращения 20.05.2019)
65. СНиП 23-05-95. «Естественное и искусственное освещение». Яз. Рус. (дата обращения 3.06.2019)
66. Поражение электрическим током [Электронный ресурс] / grandars URL: http://www.grandars.ru/shkola/bezopasnost-zhiznedeyatelnosti/porazhenie-elektricheskim-tokom.html, свободный, Яз. Русс. (дата обращения 20.05.2019)
67. О.А. Калиничева Основы электробезопасности в электроэнергетике / Северный (Арктический) федеральный университет имени М.В.Ломоносова. – Учеб. Пос. – Архангельск «С(А)ФУ», 2015 – 126 с
68. Ю.В. Анищенко, М.В. Гуляев, М.Э.Гусельников Методические указания к выполнению лабораторной работы по курсу «Безопасность жизнедеятельности» для студентов очного и заочного обучения всех направлений и специальностей. – Томск: Изд-во ТПУ, 2009 – 46 с
69. Куда деть старый компьютер: утилизировать, сдать на запчасти или продать [Электронный ресурс] / rcycle.net URL:

https://rcycle.net/tehnika/bytovaya-i-orgtehnika/utilizacija-staryh-kompyuterov, свободный, Яз. Русс. (дата обращения 20.05.2019)

70. ГОСТ Р 55102-2012 «Ресурсосбережение. Обращение с отходами. Руководство по безопасному сбору, хранению, транспортированию и разборке отработавшего электротехнического и электронного оборудования, за исключением ртутьсодержащих устройств и приборов» Яз. Рус. (дата обращения 3.06.2019)
71. Утилизация отходов компьютерной техники и компьютеров [Электронный ресурс] / vtorothody.ru URL: https://vtorothody.ru/utilizatsiya/kompyuternoj-tehniki-i-kompyuterov.html, свободный, Яз. Русс. (дата обращения 20.05.2019)
72. ГОСТ Р 22.9.28-2015. Национальный стандарт Российской Федерации. Система стандартов безопасности труда. Безопасность в чрезвычайных ситуациях. Инструмент аварийно-спасательный. Классификация, Яз. Русс. (дата обращения 20.05.2019)
73. Действия при пожаре в квартире или здании [Электронный ресурс] / Сибирский федеральный университет URL: http://my.sfu-kras.ru/safety/fire-building, свободный, Яз. Русс. (дата обращения 20.05.2019)
74. ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования» (дата обращения 3.06.2019)

Приложение А

Таблица А.1 - Сравнение алгоритмов кластерного анализа, не требующих явного задания количества кластеров

Название алгоритма	**Сложность алгоритма**	**Форма получающихся кластеров**	**Результат алгоритма**	**Чувствительность алгоритма**	**Специфики алгоритма**	**Пригодность для использования в работе**
FOREL	$O(n^2)$	Сфера или полусфера с заданным радиусом	Матрица принадлежности к кластерам	Неустойчив (зависимость от выбора начального объекта)	+Сходимость алгоритма -Плохая применимость алгоритма при плохой разделимости выборки на кластеры -Необходимость априорных знаний о ширине (диаметре) кластеров -Плохая применимость алгоритма при необходимости разбить выборку на кластеры разной плотности	Непригоден

Продолжение таблицы А.1

Алгоритм выделения связных компонент	Зависит от конкретной модификации алгоритма	Произвольная	Древовидная структура кластеров	Неустойчив (зависимость от выбора R)	+Простота реализации - Сложный подбор параметра R, иногда подбор невозможен, из-за чего алгоритм сложен в использование -Плохая применимость алгоритма при плохой разделимости выборки на кластеры	Непригоден
Алгоритм минимального покрывающего дерева	O(n2 log n)	Произвольная	Древовидная структура кластеров	Устойчив	+Более точное разделение, чем в алгоритме выделения связных компонент -Плохая применимость алгоритма при плохой разделимости выборки на кластеры	Пригоден

Продолжение таблицы А.1

Fast Affinity propagation	Меньше или равно $O(n^2T)$	Произвольная	Центры кластеров, матрица принадлежности	Устойчив	+Хорошая применимость алгоритма при плохой разделимости выборки на кластеры +Выделены центры кластеров	Пригоден
DBSCAN	$O(n^2)$	Произвольная	Матрица принадлежности к кластерам или шуму	Устойчив	-Алгоритм отбрасывает шум +Хорошая применимость алгоритма при необходимости разбить выборку на кластеры разной плотности	Пригоден, при условии исключения удаления шума

Пояснение о пригодности алгоритма DBSCAN:

DBSCAN имеет важный минус – в процессе выполнения DBSCAN происходит удаление шума, что означает исключение из выборки отделяющихся объектов. Однако, подобные объекты могут быть самыми важными пользователями-экспертами социальной сети, которые имеет авторитет в заданной области на порядок выше остальных. Если исключить из DBSCAN процесс удаления шума, то данный алгоритм станет одним из лучших для решения поставленной задачи.

Приложение Б

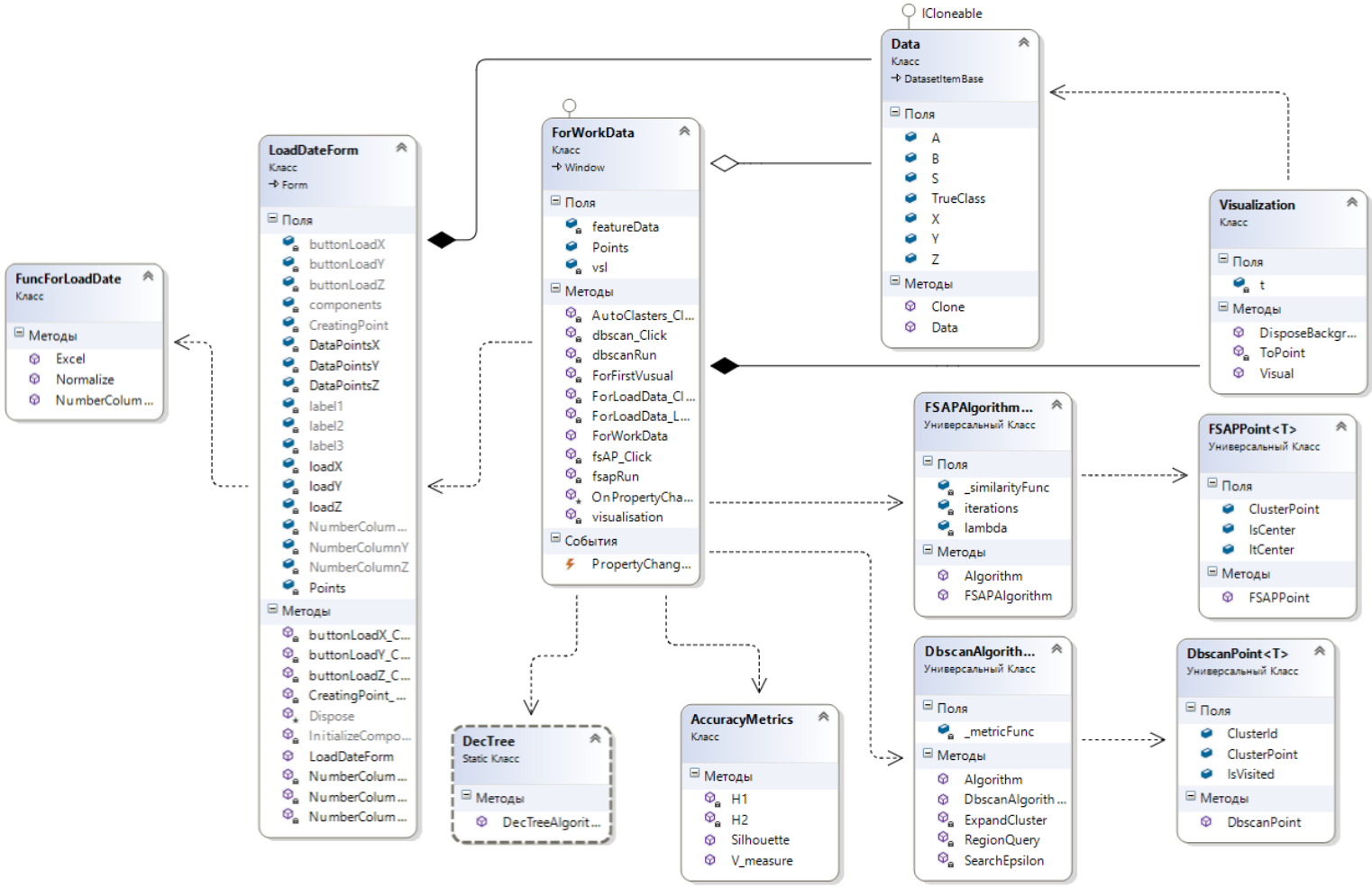

Рисунок Б.1 – Диаграмма классов

Приложение В

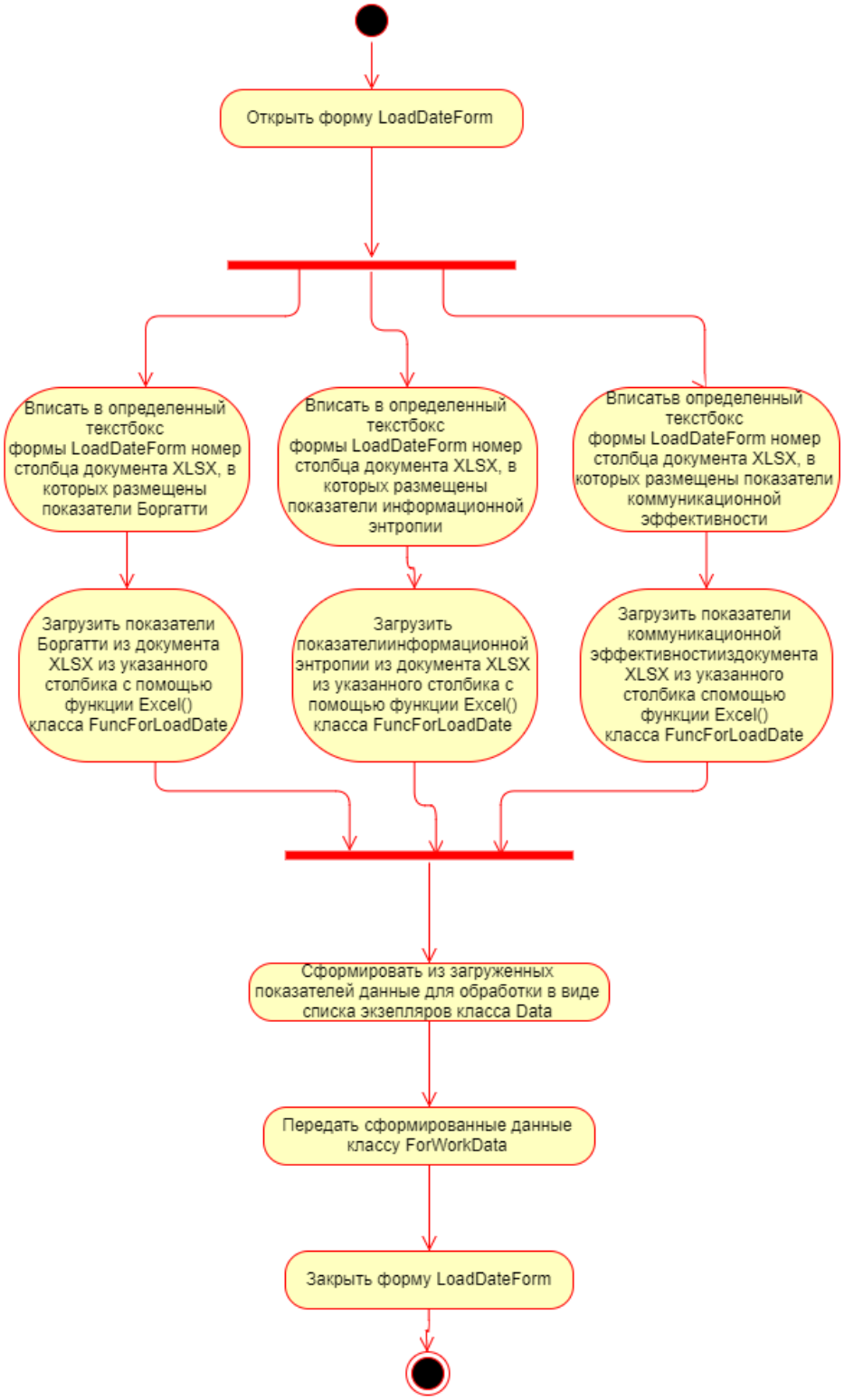

Рисунок В.1 – Диаграмма деятельности Функция загрузки и обработки данных показателей Боргатти, информационной энтропии, и коммуникационной эффективности из документа формата XLSX

Приложение Г

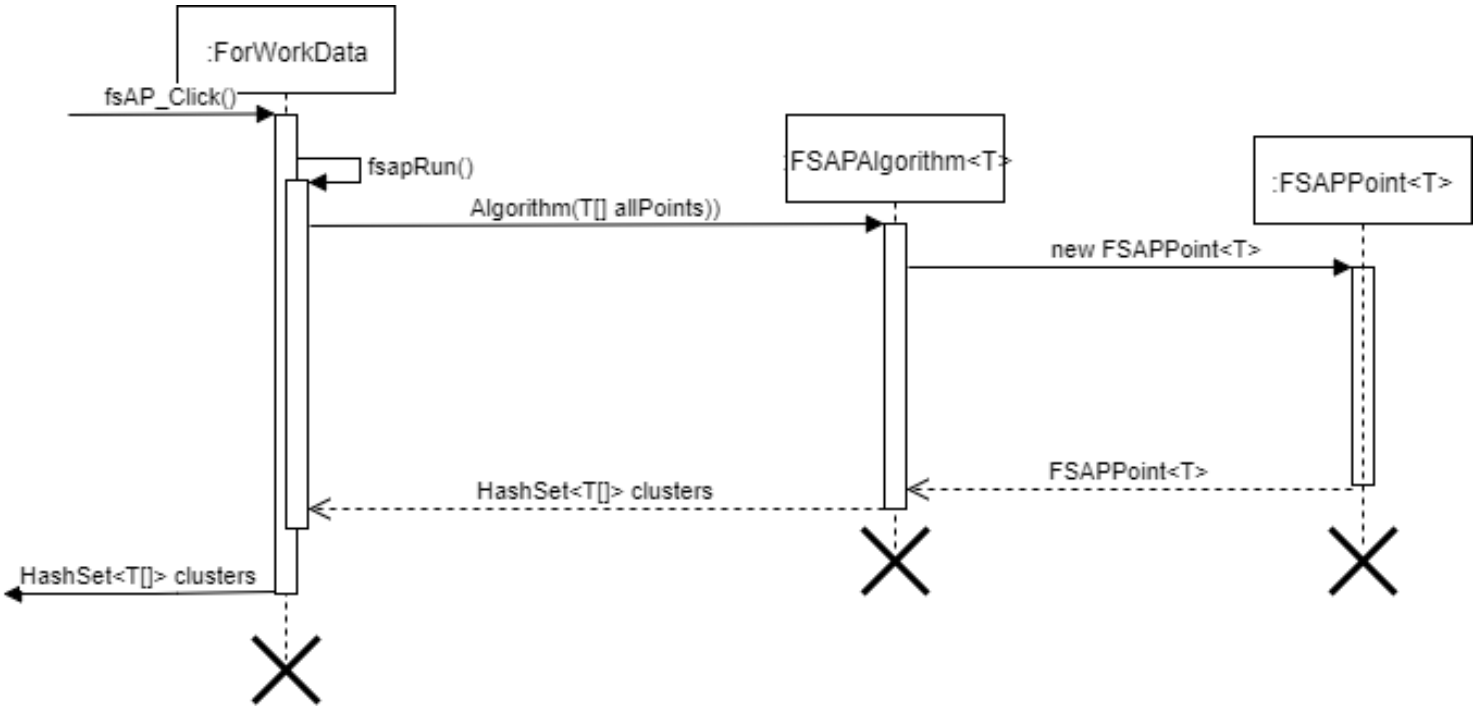

Рисунок Г.1 – Диаграмма последовательности для функции кластеризации данных алгоритмом Fast Affinity propagation

Приложение Д

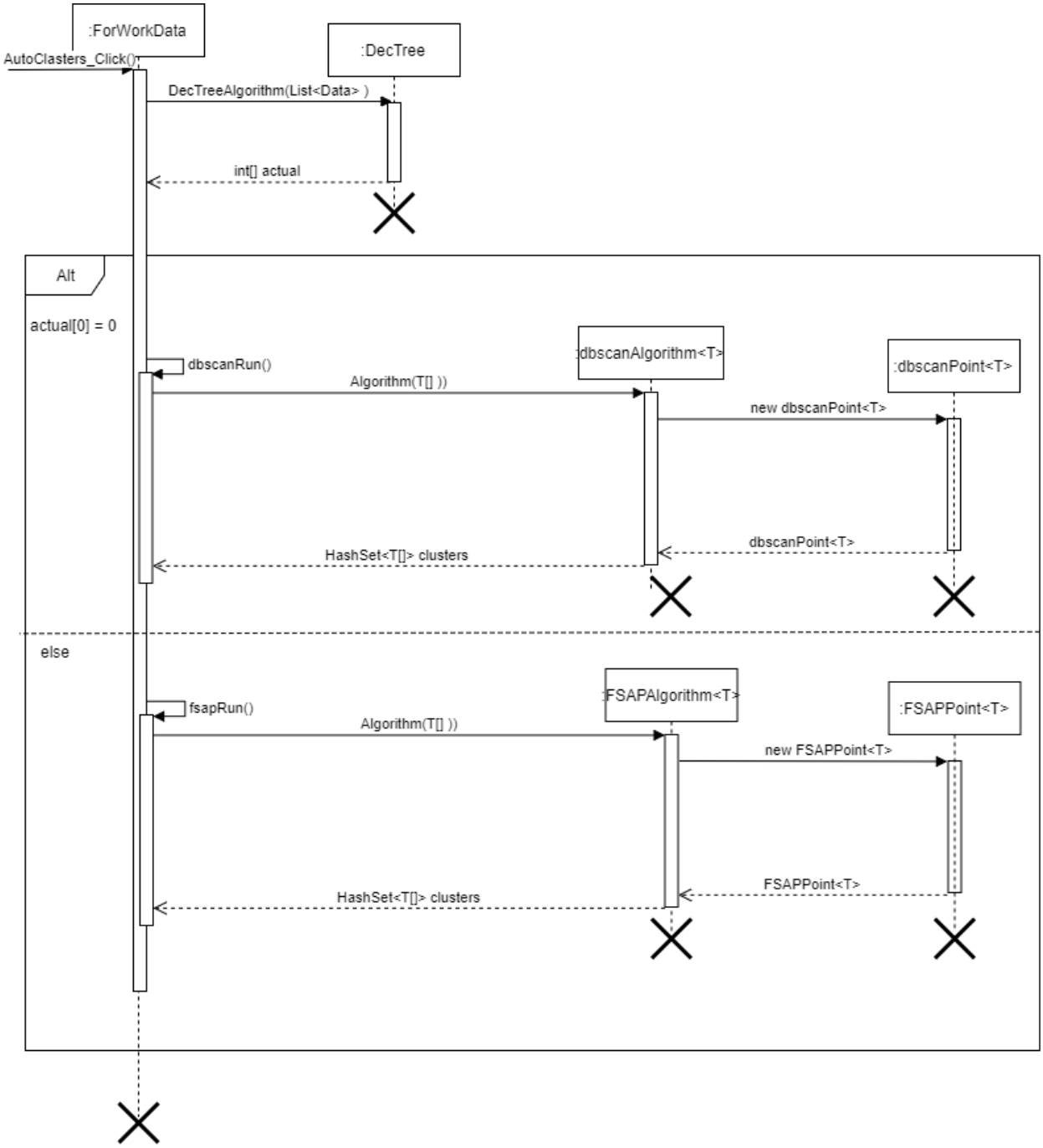

Рисунок Д.1 – Диаграмма последовательности функции автоматической кластеризации

Приложение Е

Принятые сокращения:

FAP – Fast Affinity propagation

Таблица Е.1 - 3. Анализ качества применения алгоритмов кластеризации DBScan и Fast Affinity Propagation для решения задачи идентификации пользователей-экспертов социальной сети

Название выборки	**Алгоритм кластеризации**	**Время выполнения, (сек)**	**Силуэт выборки**	**V-мера**	**Алгоритм подходящий для кластеризации**	**Алгоритм, выбранный классификатором**
3П500Ч	DBScan	0,071	1	0,9176	DBScan	DBScan
	FAP	55,515	0,9984	0,6795		
5П200Ч	DBScan	0,039	1	0,8538	DBScan	DBScan
	FAP	3,840	0,9955	0,7057		
10П300Ч	DBScan	0,031	1	0,7933	DBScan	FAP
	FAP	12,388	0,9975	0,7160		
10П200Ч	DBScan	0,013	1	0,7774	DBScan	FAP
	FAP	4,373	0,9956	0,6981		
20П200Ч	DBScan	0,012	0	0	FAP	FAP
	FAP	4,694	0,9966	0,6481		

Приложение Ж

Таблица Ж.1 - диаграмма Ганта

Код работы	Вид работ	$T_к$, раб., дн.	Исполнители	Продолжительность выполнения работ													
				февр.			март			апрель			май			июнь	
				1	2	3	1	2	3	1	2	3	1	2	3	1	2
1	Определение целей и задач	4	НР														
2	Составление и утверждение технического задания	9	НР														
			И														
3	Разработка календарного плана	7	НР														
			И														
4	Подбор и изучение материалов по теме	24	И														
5	Обсуждение литературы	1	НР														
			И														
6	Процесс реализации приложения	33	И														
7	Оформление расчетно-пояснительной записки	15	И														

Продолжение таблицы Ж.1

8	Оформление графического материала	2	НР															
			И															
9	Подведение итогов	2	НР															
			И															

Приложение 3

Принятые сокращения:

ВНР - Вероятность наступления, оценивается по пятибалльной шкале (1-5);

ВР - Влияние риска (1-5), оценивается по пятибалльной шкале;

УР - Уровень риска, оценивается из суммы ВНР и ВР;

Определение уровней риска:

Н – Низкий уровень риска, показатель 0-3;

С – Средний уровень риска, показатель 4-6;

В – Высокий уровень риска, показатель от 7-10;

Таблица Б.1 – Реестр рисков

№	Риск	Потенциальное воздействие	ВНР	ВР	УР	Способы смягчения риска	Условия наступления
1	Объем работ оценен не верно	Не выполнение поставленных сроков	2	3	С	1. Полное и глубокое изучение тематики проекта перед установлением сроков; 2. Возможность изменения сроков проекта.	1. Поверхностное понимание тематики проекта; 2. Сложность тематики проекта.
2	Алгоритмы кластеризации реализованы некорректно	Приложение работает некорректно	1	5	С	Глубокое изучение алгоритмов кластеризации	Поверхностное представление, как работают алгоритмы кластеризации во время их разработки
3	Функция загрузки данных не реализована или работает некорректно	Приложение не пригодно к использованию	1	5	С	Подбор ответственных исполнителей проекта	Невнимательность или безолаберность исполнителей проекта

Приложение И

Introduction

Chapter1. Analysis of the possibility of applying machine learning methods to the task of identifying users-experts of a social network in a certain area

Студент:

Группа	**ФИО**	**Подпись**	**Дата**
8ВМ71	Виноградова Дарья Андреевна		

Руководитель ВКР:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент	Лунева Е. Е.	к.т.н.		

Консультант – лингвист отделения ИЯ:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Старший преподаватель	Куркан Н. В.			

Introduction

The modern world can not be imagined without social networks. The high level of involvement of the modern person in the process of virtual communication leads to a huge accumulation of various data on social networks, the analysis of which can reveal valuable information of various kinds.

Information obtained from social networks may be decisive in addressing anti-terrorism issues, political analytics, forecasting the reputational risks of a company or a subject, assessing the demand for a product or service, and monitoring public opinion.

Social networks have a significant impact on people's opinions. Social network users often express their own opinions on current issues, events, products or services in areas that interest them. However, it can be noted that the opinion of some users in certain areas has a high impact on the opinion of other users.

Information about influential users (hereinafter referred to as user-expert) in a certain area can be useful in solving various practical problems. An example of solving a marketing task: a cosmetics production company can get a high-quality advertisement from a user-expert of a social network in the beauty area.

At present, it is impossible to definitely name the only best method for finding users-experts of a social network in a given area. This work is aimed at creating a methodology that allows to process and interpret the ranking data of the nodes of a social graph using several effective methods. The practical part of the work is connected with the development of the corresponding algorithmic base and software for the identification of users-experts.

Chapter 1.Analysis of the possibility of applying machine learning methods to the task of identifying users-experts of a social network in a certain area

Before starting to create a methodology that allows processing and interpreting the ranking data of social graph nodes for identifying users-expert , or to designe and to develope software for identifying users-expert, it is necessary to review and analyze literature and other sources of information on the identification topic users-expert in the social network.

This chapter reviews the task of identifying users-expert in a social network, considers the possibility of using machine learning methods to solve it, and also defines the goals and objectives of this dissertation.

1. Overview of the task of identifying users- expert in the social network

Many literary sources are devoted to the problem of identifying significant nodes in social networks [1-3]. To distinguish such nodes, use different terms such as: an influential node, a key player, a critical node, a vital node, etc. [3].

In [1], this problem was designated as the problem of finding key players (KPP, Key player problem) and was divided into two classes: Positive (POS) and Negative (NEG).

The KPP NEG class is aimed to find such a node (or group of nodes) of a social graph, the removal of which will make the graph as fragmented as possible. Approaches to solving this problem can be applied to solving problems in the field of medicine; For example, in epidemics, it is the task of identifying a risk group whose vaccination will minimize the spread of the disease. Also, these approaches are used to solve for anti-terrorist tasks, in particular, when identifying participants whose exclusion from the terrorist group will maximally separate the remaining members. Later this class of tasks was assigned the term CNP (Critical Node Problem) [2] and an analytical description was given.

The second class of KPP POS tasks is aimed at identifying a node (or group of nodes) that maximally distributes information over the network. Solving this

problem is necessary, for example, to solve the problem of introducing informers into a terrorist network. These informants will disseminate incorrect information as quickly as possible [1].

The users of social networks, who are experts in a certain area, set the topics for discussion, express their opinion on certain questions on a given topic and current events. At the same time, other users listen to the opinion of experts and often change their views on the topics read out in the this area under discussion. Thus, the opinions of users-expert are distributed on the social network due to various actions of readers, such as commenting, repost, mention, “like”. So the task of identifying users-expert in social networks has some differences from the CNP task and, obviously, belongs to the class of KPP POS tasks.

Analysis of the literature [1, 4-6] revealed the following ways to solve the KPP POS problem, which are considered the most effective:

• a method based on calculating the Borgatti indicator;

• calculation of the information entropy of the graph vertices;

• calculation of communication efficiency.

However, the application of the proposed methods of solving the problem of KPP POS to the task of identifying users-experts of social networks, being leaders of public opinion or considered as experts in a certain area, is associated with the following difficulties:

- • In the literature, the solution to this problem is tested on undirected and / or unweighted graphs [1, 4, 7–9]. The use of such graphs significantly reduces their ability to adequately represent the real group of users of social networks, because it does not allow to take into account additional information about the relationship between users [10, 11].
- • There is a difficulty in choosing the optimal way to solve this problem. This difficulty is related to the fact that the assessment of the effectiveness of solutions is hampered by the lack of actual data, and may also depend on the specific characteristics of the analyzed social graphs.

Thus, to obtain a high-quality result of solving the problem of identifying users-expert in social networks, it is necessary to use all the above-mentioned methods for solving the KPP POS problem. After that, the obtained disparate results for each method must be combined and interpreted to obtain a single solution.

2.Overview of effective ways to solve the KPP POS problem as applied to the task of identifying users-expert of the social network

As mentioned earlier, the following methods are considered the most effective ways to solve the KPP POS problem:

• a method based on calculating the Borgatti indicator;

• calculation of the information entropy of the graph vertices;

• calculation of communication efficiency.

Each of the above methods uses the social graph as input data.

2.1. The method of constructing a social graph in solving the problem of identifying users-expert

The input data for searching users-expert in social network in a certain area is quantitative information that reflects the interest of social network users in expert opinion. Such information includes:

- the number of reposts - r,
- comments - c,
- mentions of an expert user - m,
- reposts with comments - x.

Obviously, the interest of user *j* by publications of user *i* can be expressed as a function:

$$f\left(r_{ij}, c_{ij}, m_{ij}, x_{ij}\right). \quad (1)$$

The analytical description of this function was proposed in [12].

Social graph is defined as a weighted oriented graph [10]:

$$G = (V, A, w) \quad (2)$$

Where

$V = \{v_i\}$ – the set of nodes of the graph ($i = 1, 2, ..., N$),

$A = \{(v_i, v_j)\} \subset V \times V$ – a set of directed ribs (arcs),

$w: A \to (0,1]$ – the weight function, which assigns a certain value $w(a_{ij}) \in (0,1]$. to each arc. $a_{ij} = (v_i, v_j) \in A$.

Then, on the basis of data selected from social network for a certain area, can build a social graph, the nodes of which are connected by edges according to the following principle: there is a directed edge (arc) a_{ij} aij connecting the nodes i and j, if the j-th user commented on publications i- he user, or made them a repost, or a repost with a comment, or mentioned in his comments user i. In this case, the following function is valid [10]:

$$w(a_{ij}) = f(r_{ij}, c_{ij}, m_{ij}, x_{ij}) \quad (3)$$

2.2.Overview of selected methods for solving the problem of KPP POS

2.2.1. The method based on the calculation of the indicator Borgatti

The principal difference of the method, that is proposed by Borgatti in [1], from other selected methods, is that the author proposes to specify the number of key players that should be found in the social graph in advance. For this, a subset of potential key players $K \subset V$ with the required number of elements is selected from the set of nodes V and the indicator C_K is determined using the following formula [1]:

$$C_K = \sum_{v_j \in V \setminus K} \max_{v_i \in K} \left(\frac{1}{d_{ij}} \right), \quad (4)$$

Where

K – he set into which some tested combination of nodes fell,

V – the set of all nodes of the graph,

d_{ij} – the shortest distance from the node v_i to the node v_j, while from the set K the node is selected, from which the shortest distance to the node v_j is minimal.

After calculating the values C_K for all possible sets of *K*, can choose a combination of nodes that make up the desired set of key players: the greater the value of C_K for the checked combination of nodes, the more influence users in this group can have on their readers. For the correct application of the Borgatti method as the weight of the arc between the nodes of the social graph, use the value calculated by the formula:

$$w(a_{ij}) = \frac{1}{f(r_{ij}, c_{ij}, m_{ij}, x_{ij})}. \quad (5)$$

2.2.2. Calculation of the information entropy of the graph vertices

The method based on the calculation of the informational entropy index implies a gradual elimination from the social graph of nodes and arcs incident to them. At the same time, the entropy of the social graph is calculated by the following formula [4]:

$$H(G,P) = \sum_{i=1}^{|V|} p_i \cdot \log_2\left(\frac{1}{p_i}\right) \quad (6),$$

Where

$V = \{v_i\}$ – the set of nodes of the graph G,

$P = \{p_i\}$ – the probability distribution density on the set V, specifying the probability that the node v_i represents the most significant user-expert.

The value p_i depends on the interest of social network users by the *i*-th user messages, expressed as the ratio of the number of reposts, comments, etc., generated by the *i*-th user messages to the entire volume of social network messages. Based on this, the values p_i, $i \in [1, N]$ can be calculated as follows:

$$p_i = \frac{\sum_{v_j \in V^{[i]}} w_{ij}}{\sum_{v_k \in V} \sum_{v_l \in V^{[k]}} w_{kl}}, \quad (7)$$

Where

$V^{[s]}$ – the set of a finite number of nodes of arcs outgoing from a node v_s,

$w_{yz} = f(r_{yz}, c_{yz}, m_{yz}, x_{yz})$ – arc's weight a_{yz}.

2.2.3. Calculation of communication efficiency

In graph theory, the problem of ranking players according to the results of a circular tournament is known [13, 14, 15]. To solve this problem, a complete directed graph without loops is compiled with the number of vertices *n* equal to the number of players. Moreover, any two vertices are connected by an arc, and the arc from the i-th vertex to the j-th one means the victory of the i-th player over the j-th. In the adjacency matrix of *A* tournament, the element is equal to one $a_{ij} = 1$ if there is an arc from the *i*-th vertex to the *j*-th; else the element is equal to zero $a_{ij} = 0$.

The sum of the elements of the *i*-th line $s_i^{\langle 1 \rangle} = \sum_{j=1}^{n} a_{ij}$ corresponds to the number of victories by the *i*-th player during the tournament, and is a measure of the strength of the i-th player. A vector $\mathbf{S}^{\langle 1 \rangle}$ composed of elements $s_i^{\langle 1 \rangle}$ is called a first-order force vector [16].

In the trivial case, all elements of the vector $\mathbf{S}^{\langle 1 \rangle}$ are different, and the ranking of the players is not difficult. However, if at the end of the tournament several players scored an equal number of victories (scored the equal number of points), unequivocal ranking of players may be difficult [13, 14].

The following procedure involves calculating the iterated forces of the *k*-th order of the players using one of the following formulas:

$$\mathbf{S}^{\langle k \rangle} = \mathbf{A} \cdot \mathbf{S}^{\langle k-1 \rangle} \quad (8)$$

$$\mathbf{S}^{\langle k \rangle} = \mathbf{A}^{k} \cdot \mathbf{S}^{\langle 1 \rangle} \quad (9)$$

Where

$k \geq 2$ - iteration number;

$\mathbf{A}$ - adjacency matrix;

$\mathbf{S}^{\langle j \rangle}$ - *j*-th order iterated vector.

3.Overview and analysis of machine learning methods in relation to the task of identifying users-experts of the social network.

In order to identify users-expert in a certain area, it is necessary to combine and interpret the solution data of the KPP POS task of each social network user who is connected with a given certain area. The volume of the user base in social networks is very large today. It means that these solution data to the KPP POS problem of all users united by interest in one certain area will also be numerous, and it very difficult to analyze them. Today, machine learning is often used to analyze large aggregations of data.

Machine learning is an extensive subsection of artificial intelligence that studies methods for constructing algorithms capable of learning [17]. The goal of machine learning is to predict results from input data.

After analyzing the sources [17-19], can distinguish two main types of machine learning:

- Supervised learning — a method of machine learning when a machine forcibly learned with examples provided to it. Examples include inputs (input data of objects) and expected outputs (expected results). The machine needs to establish some dependency between the inputs and the expected outputs. Further, the machine, using the derived dependence, will select the output values for the new inputs supplied to it [20, 21].
- Unsupervised learning - a method of machine learning, when the machine learns itself to perform the task. Usually, this method is suitable for solving a problem when only inputs are known, and it is required to establish internal interrelations, dependencies, regularities existing between objects [22, 23].

All other types of machine learning is presented in [17-19] are particular cases of the selected types of machine learning, or their mixing.

Classical problems solved by supervised learning:

- • Classification - the task of distributing objects to a pre-selected class.

- • Regression - the task of predicting the desired value.

Classical problems solved by unsupervised learning:

- • Clustering - the task of dividing objects into previously unknown groups.
- • Generalization - the task of reducing the data dimension.
- • Association - the task of identifying sequences, building dependencies of objects from each other.

The task of identifying users-expert assumes that all users of a social network associated with a certain area need to be divided into two previously known groups: writers (users-expert) and readers (other users). Thus, the task of identifying users-expert is related to the task of classification, which is solved by supervised learning. However, data classification requires a learning sample (examples of inputs with expected outputs for them), the creation of which will take a lot of time and effort.

To solve the problem of identifying users-expert, the nodes of the social graph can be ranked not only by classification. If for the solution of this problem use clustering that solved by unsupervised learning then internal interrelations, dependencies and patterns will be established between the objects, on the basis of which the users in question will be divided into groups. Using this method will allow to obtain additional information about the groups of readers who are under the influence of users-expert, and information about potential hidden leaders and experts.

Thus, to maximize the usefulness of solving the problem of identifying users-expert in a social network, it is necessary to use clustering analysis to solve it.

4. Goal and tasks of the work

The given overview and analysis of literature and other sources of information on the topic of identification of users-expert in a social network showed that to solve a given task, the following steps are necessary:

- Data for each user of a social network who associated with a certain area obtain. These data intended for solving the KPP POS problem and

represent Borgatti indicator, information entropy indicator and indicator of communication efficiency.

- To make a ranking of the obtained indicators using clustering analysis.

The goal of this work is creating a methodology that allows to process and to interpret the ranking data of the nodes of the social graph programmatically to identify users-expert.

To achieve this goal it is necessary to solve the following tasks:

- • Develop a methodology for processing the ranking data of social graph nodes for the task of identifying users-expert of a social network.
- • Design software for processing ranking data of the social graph nodes for identifying users-expert of the social network according to the developed methodology.
- • Implement this software.