TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Проектирование и разработка системы поиска и кластеризации научно-технических публикаций на основе информации из открытых Интернет-источников

УДК 004.775-047.84:001.89:002.2

Студент

Группа	ФИО	Подпись	Дата
8ПМ7И	Демидова Оксана Олеговна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Савельев Алексей Олегович	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОСГН ШБИП	Потехина Нина Васильевна	-		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко Михаил Владимирович	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

Томск – 2019 г.

Планируемые результаты обучения

Код результата	Результат обучения (выпускник должен быть готов)
Общие по направлению подготовки 09.04.04 «Программная инженерия»	
Р1	Проводить научные исследования, связанные с объектами профессиональной деятельности.
Р2	Разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах.
Р3	Составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты.
Р4	Проектировать системы с параллельной обработкой данных и высокопроизводительные системы.
Р5	Осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных.
Р6	Осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем.
Р7	Организовывать промышленное тестирование создаваемого программного обеспечения.
Профиль «Технологии больших данных»/ «Big data solutions»	
Р8	Исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях.
Р9	Понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях.
Р10	Применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

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

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

Студенту:

Группа	**ФИО**
8ПМ7И	Демидова Оксана Олеговна

Тема работы:

Проектирование и разработка системы поиска и кластеризации научно-технических публикаций на основе информации из открытых Интернет-источников	
Утверждена приказом директора (дата, номер)	№3794/с от 25.02.2019

Срок сдачи студентом выполненной работы:	6.06.2019

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Проектирование и разработка системы поиска и кластеризации научно-технических публикаций на основе данных, собранных из открытой электронной библиотеки eLibrary.

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. Изучение предметной области; 2. Обзор существующих решений 3. Разработка системы сбора данных 4. Сбор необходимых данных 5. Разработка системы хранения данных 6. Подготовка данных для кластеризации 7. Реализация метода кластеризации 8. Анализ результатов исследования 9. Расчет показателей ресурсоэффективности 10. Оценка показателей социальной ответственности
Перечень графического материала	

Консультанты по разделам выпускной квалификационной работы

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Старший преподаватель ОСГН ШБИП Потехина Нина Васильевна
Социальная ответственность	Доцент ООД ШБИП Горбенко Михаил Владимирович
Раздел на иностранном языке	Доцент ОИЯ ШБИП Диденко Анастасия Владимировна

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Обзор существующих методов кластерного анализа (Web-mining methods)

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель / консультант (при наличии):

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Савельев Алексей Олегович	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демидова Оксана Олеговна		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Уровень образования Магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения 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:	6.06.2019

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
31.03.2019	*Раздел 1. Аналитический обзор предметной области*	*25*
28.04.2019	*Раздел 2. Практическая разработка*	*30*
04.05.2019	*Раздел 3. Результаты исследования*	*15*
17.05.2019	*Раздел 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение*	*10*
23.05.2019	*Раздел 5. Социальная ответственность*	*10*

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Савельев Алексей Олегович	к.т.н.		

СОГЛАСОВАНО:

Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ7И	Демидова Оксана Олеговна

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	1. *Оклад инженера –21760;* 2. *Оклад научного руководителя – 33664;*
2. *Нормы и нормативы расходования ресурсов*	1. *Месячная норма амортизации – 3,3 %%*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	1. *Ставки налоговых отчислений во внебюджетные фонды (ст. 426 НК РФ) – 30%* 2. *Районный коэффициент по г. Томску (ст. 426 НК РФ, Постановление Правительства РФ от 13.05.92. №309) – 1,3*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	1. *Описание потенциальных потребителей;* 2. *Анализ конкурентоспособности;* 3. *Диаграмма Исикавы* 4. *SWOT-анализ.*
2. *Разработка устава научно-технического проекта*	*Формирование цели, задач и ожидаемых результатов проекта. Иерархическая структура работ.*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	1. *Планирование структуры работ проекта;* 2. *Определение трудоемкости выполнения работ;* 3. *Формирование бюджета;* 4. *Анализ рисков НТИ..*
4. *Определение ресурсной, финансовой, экономической эффективности*	*1.Выводы по разделу*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. *Оценочная карта для сравнения методов кластеризации;* 2. *Диаграмма Исикавы;* 3. *SWOT-анализ НТИ;* 4. *Оценки степени готовности научного проекта к коммерциализации;*	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ст.преподаватель ОСГН ШБИП	Потехина Нина Васильевна	-		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демидова Оксана Олеговна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ПМ7И	Демидова Оксана Олеговна

Школа	ИШИТР	Отделение школы (НОЦ)	Отделение информационных технологий
Уровень образования	Магистр	Направление/специальность	09.04.04 Программная инженерия

Исходные данные к разделу «Социальная ответственность»:	
1. Описание рабочего места (рабочей зоны, технологического процесса, механического оборудования)	*В соответствии с ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования».*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности.	*Ознакомление с:* *- ГОСТ 12.2.032-78 ССБТ* *- СанПиН 2.2.4.548–96* *- СанПиН 2.2.4/2.1.8.562–96* *- СанПиН 2.2.2/2.4.1340–03* *- ГОСТ 12.1.009-2009* *- ГОСТ 12.1.038-82 ССБТ* *- ГОСТ Р 22.3.03-94*
2. Выявление и анализ вредных факторов проектируемой производственной среды	*- Освещение* *- Микроклимат* *- Шум* *- Психофизиологические факторы: нервно-психические перегрузки*
3. Выявление и анализ опасных факторов проектируемой произведённой среды.	*- Электрический ток (источник – ПК)* *- Короткое замыкание* *- Статическое заземление (источник – ПК)*
4. Охрана окружающей среды.	*Воздействие объекта на атмосферу, гидросферу отсутствует. Воздействие на литосферу происходит при утилизации ПК, используемого для разработки, а также утилизации люминесцентных ламп освещения.*
5. Защита в чрезвычайных ситуациях.	*Возможной чрезвычайной ситуацией при разработке алгоритма является возникновение пожара на рабочем месте.*

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД	Горбенко М. В.	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демидова Оксана Олеговна		

РЕФЕРАТ

Выпускная квалификационная работа содержит 79 страниц, 15 рисунков, 20 таблиц, 23 источника,2 приложения.

Ключевые слова: Веб-майнинг, веб-скрапинг, кластеризация, алгоритмы кластеризации, анализ кластеризации.

Объектом исследования являются базы цитирований. Целью данной работы является отработка технологии сбора, анализа и кластеризации данных на основе открытых интернет-источников.

В процессе исследования были проанализированы существующие решения сбора данных из интернет-источников. Был произведен анализ подходов и инструментов web-mining. Рассмотрели методы для нахождения взаимосвязанности публикаций, и методы кластеризации, подходящие для работы с полученными данными.

В результате исследования были разработаны функции извлечения научных публикаций из открытой электронной библиотеки eLibrary. Разработали систему хранения полученных публикаций в базе данных. Произвели подготовку полученных данных к кластеризации, и реализовали кластеризацию публикаций по цитированию.

Область применения: потребителями разработанной технологии могут быть:

- наукометрические базы данных
- академические издательства
- базы цитирований
- научные организации
- интернет ресурсы, научное сообщество (университеты)

В будущем планируется проводить исследование и улучшение технологии сбора, анализа и кластеризации данных на основе цитирования публикаций.

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В данной работе используются следующие термины с соответствующими определениями:

Наукометрическая база данных – библиографическая и реферативная база данных, инструмент для отслеживания цитируемости научных публикаций.

Атрибут – переменная, связанная с классом или объектом;

URL – единый указатель ресурса;

Web Mining – это использование методов интеллектуального анализа данных для автоматического обнаружения веб-документов и услуг, извлечения информации из веб-ресурсов и выявления общих закономерностей в Интернете[5].

Контент – содержимое, информационное наполнение;

DOM – объектная модель документа;

HTML – стандартизированный язык разметки документов во Всемирной паутине.

HTTP – широко распространённый протокол передачи данных, изначально предназначенный для передачи гипертекстовых документов (то есть документов, которые могут содержать ссылки, позволяющие организовать переход к другим документам).

Оглавление

ВВЕДЕНИЕ

Отличительной чертой современности является всевозрастающий объем информации, генерируемой различными системами, средами и сообществами. Научное сообщество располагает огромным количеством материала. Распространение и увеличивающееся влияние международных баз цитирований, наряду с доступностью научных текстов, обеспечиваемой крупными академическими издательствами, с одной стороны, способствуют распространению научного знания и повышению качества результатов исследований. С другой стороны, объем научной информации, представленный в открытом доступе и перешедший в категорию больших данных, усложняет анализ, уточнение и корректировку научно-технических приоритетов на уровне государства [6].

Постоянный рост объема научной информации, представленной в электронной форме, делает процесс поиска данных все более сложным, трудоемким и неэффективным технологическим процессом. Пользователи, переходя от списка к списку документов, обязаны уточнять критерии для поиска нужной информации, и доводить свой запрос до оптимального набора слов, по которому он часто получает, во многом знакомый ему, перечень документов. Процесс поиска зацикливается, а время поиска значительно возрастает.

Целью данной работы является отработка технологии сбора, анализа и кластеризации данных на основе открытых интернет-источников.

Объектом исследования являются базы цитирования.

Предметом исследования являются системы извлечения контента из открытых интернет-источников и способы кластеризации полученных данных [2].

Актуальной является задача разработки инструментальных средств анализа больших объемов данных для автоматизации процессов классификации и оценки значимости научных текстов, выявления степени связанности и взаимного влияния перспективных направлений исследований

и визуализации структуры научной деятельности с целью поддержки принятия решений в рамках программ государственной поддержки научной, научно-технической и инновационной деятельности.

Научной новизной предлагаемого исследования является технология сбора и анализа данных для обнаружения закономерностей в большом объеме слабоструктурированной информации на примере данных научных публикаций.

Основными задачами проекта являются:

- .Разработка схемы данных сбора, хранения и предоставления агрегированной информации о научных публикациях.
- Разработка схемы данных стандарта сбора, хранения и предоставления агрегированной информации о научных публикациях.

Глава 1. Обзор предметной области

Для решения поставленных задач требуется подход, основанный на автоматизированном анализе доступного объема открытых публикаций [10].

Извлечение структурированных связных данных с веб-страниц можно представить как последовательное решение задач:

- поиска и получения целевых страниц для извлечения данных;
- распознавания участков, содержащих нужные данные;
- поиска структуры найденных данных;
- обеспечения однородности извлекаемых данных;
- объединения данных с разных источников.

Для эффективного решения задач поиска, структурирования и анализа в основном хаотично организованной информации в сети предназначено новое направление в методологии анализа данных – Web-mining.

1. Web-mining

Web-mining развивается на пересечении таких дисциплин как обнаружение знаний в базах данных, эффективный поиск информации, искусственный интеллект, машинное обучение и обработка естественных языков [11].

Задачи Web mining:

1. Поиск информации

Пользователи пользуются поисковыми ресурсами для нахождения необходимой информации. При этом обычно используют простые запросы по ключевым словам. Результатом запроса представляет собой список страниц, который был отсортирован по некому индексу релевантности, описывающему степень совпадения результата с запросом [13]. Но существующие поисковые механизмы обладают недостатками. Главным недостатком является низкая точность результата, который вызван недостаточным учетом семантических связей и контекста найденных в тексте выражений.

Индексация интересующих сегментов сети с использованием интеллектуального анализа данных, применяющего алгоритмы математической лингвистики и обработки естественных языков, является перспективным направлением Web Mining в области поиска информации.

2. Анализ структуры сегмента сети

Этот метод заключается в анализе структуры ссылок между различными веб-страницами, внутренними и внешними сайтами в выделенном сетевом сегменте. Появление этого метода было вызвано необходимостью решения задач, возникающих при анализе социальных сетей или специфических областей человеческой деятельности или знаний, например, в анализе цитирования авторов. [14]

3. Выявление знаний из веб-ресурсов

Эту задачу можно рассматривать как проблему поиска информации, описанную выше. Только здесь исследователь имеет набор веб-страниц, которые он получил в результате запроса. Далее необходимо произвести их обработку с точки зрения автоматической классификации, составления оглавлений, выявления ключевых слов и общих тем.

4. Персонализация информации

Персонализации веб-пространства - задача по созданию веб-систем, адаптирующих свои возможности (навигация, контент, баннеры и другие рекламные предложения) под пользователя на основании собранной и проанализированной информации о пользовательских предпочтениях.

Для анализа информации о пользователе следует в наименьшей степени использовать декларируемую о себе информацию, а скорее основываться на стойких шаблонах его "поведения" в сети - последовательности кликов внутри ресурса, переходах на другие под-ресурсы, периодах сетевой активности, осуществляемых покупках и т.д.

5. Поиск шаблонов в поведении пользователей

Эта задача связана с предыдущей, но ее целью является не адаптация ресурса к предпочтениям индивидуальных пользователей, а поиск

закономерностей в шаблонах взаимодействия пользователя с веб-ресурсом с целью прогнозирования его последующих действий. Анализируемые действия пользователей могут включать не только переходы по ссылкам, но и отправку форм, прокрутку страниц, добавление в избранные страницы и т.д. Найденные шаблоны используются в дальнейшем для оптимизации структуры сайта, изучения целевой аудитории и для прямого маркетинга.[21]

2. Категории Web-mining

Web-mining можно разделить на три категории:

- Анализ использования веб-ресурсов (Web Usage Mining);

Это направление основано на извлечении данных из логов веб-серверов. Целью анализа является выявление предпочтений посетителей при использовании тех или иных ресурсов сети Интернет[7].

- Извлечение веб-структур (Web Structure Mining);

Процесс обнаружения структурной информации в Интернете. Данное направление рассматривает взаимосвязи между веб-страницами, основываясь на связях между ними. Построенные модели могут быть использованы для категоризации и поиска схожих веб-ресурсов, а также для распознавания авторских сайтов.

- Извлечение веб-контента (Web Content Mining).

Поиск знаний в сети Интернет является непростой и трудоемкой задачей. Именно это направление Web-mining решает её. Оно основано на сочетании возможностей информационного поиска, машинного обучения и интеллектуального анализа данных[7].

На рисунке 1 представлена общая взаимосвязь между категориями Web Mining и задачами интеллектуального анализа данных

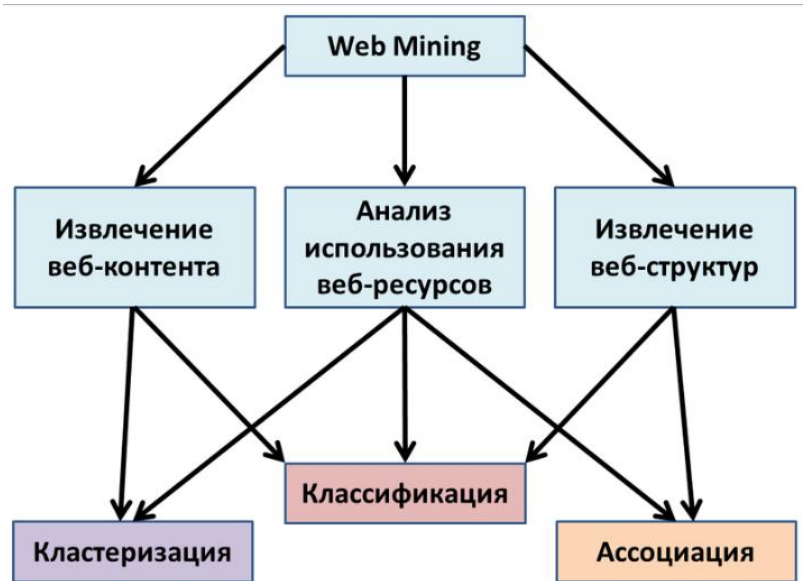


Рисунок 1 Общая взаимосвязь между категориями Web Mining и задачами интеллектуального анализа данных

В данном случае, наша работа сводиться к извлечению веб-контента (Web content mining).

3. Анализ подходов web-mining

Для извлечения данных из веб-страниц используют следующие описанные ниже техники и их сочетания [8]:

- Парсинг строк;

Можно использовать если данные отображаются с помощью, например, таблицы характеристик, когда значения параметров стандартные, а меняются только их значения. Метод не подходит для написания серьезных парсеров.

- Анализ DOM(document object model) дерева;

Используя этот подход, данные можно получить напрямую по идентификатору, имени или других атрибутов элемента дерева (таким элементом может служить параграф, таблица, блок и т.д.). Кроме того, если элемент не обозначен каким-либо идентификатором, то к нему можно добраться по некоему уникальному пути, спускаясь вниз по DOM дереву.

- Использование регулярных выражений

Этот подход лучше использовать только для извлечения данных, которые имеют строгий формат (электронные адреса, телефоны и т.д).

- XML парсинг.

Рассматривать HTML как XML данные не всегда подходящий вариант, т.к. HTML редко бывает валидным, т.е. таким, что его можно рассматривать как XML данные. Библиотеки, реализовавшие такой подход, больше времени уделяли преобразованию HTML в XML и уже потом непосредственно парсингу данных.

4. Анализ инструментов

Существуют три основных программных инструментов для извлечения данных с веб-страниц [18]:

- Библиотеки

Этот подход требует понимания процесса формирования запросов и логики работы приложения. Подходит для написания парсера к конкретному сайту. К таким инструментам относятся многочисленные библиотеки для различных языков программирования таких как JAVA, Python, PHP и т.д.

- Headless-браузеры

Данный подход позволяет обрабатывать страницу в браузере с поддержкой JavaScript, что позволяет писать свои сценарии для получения требуемой информации и даже использовать JavaScript библиотеки вроде jQuery для извлечения информации со страницы, что ускоряет разработку парсеров. К таким инструментам можно отнести PhantomJS и SlimerJS.

- 3. SaaS решения.

Данные сервисы предоставляют графический интерфейс, с помощью которого можно указать адрес страницы, указать блоки, из которых нужно извлечь информацию, а также создать ряд правил по извлечению данных. Такие сервисы не обладают той гибкостью, которую предоставляют низкоуровневые решения. Некоторые из них стоят довольно дорого, зато ими просто пользоваться.

5. Кластеризация научных публикаций

Кластеризация (или кластерный анализ) - это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма [27].

Кластеризация научных публикаций является важной проблемой в области библиометрии. Методы кластеризации регулярно применяются в библиометрической литературе для определения областей исследований или научных областей. Эти методы, например, используются для группировки публикаций в кластеры на основе их отношений в сети цитирования.

Кластеризация данных представляет собой ценный инструмент анализа данных в современных приложениях машинного обучения и интеллектуального анализа данных. Во многих случаях кластеризация используется для получения первых сведений о данных в процессе анализа и для решения ряда реальных проблем, например, как моделирование тем в интеллектуальном анализе текста.

Показано, что спектральная кластеризация (SC- Spectral clustering) является одним из наиболее эффективных алгоритмов кластеризации, благодаря своей способности решать нелинейные сепарабельные задачи. Эта эффективность может быть объяснена тем, что данные в исходном пространстве отображаются в новое вложение, где шаблоны подобных точек возникают легче. Это вложение является пространством, охватываемым собственными векторами матрицы Лапласа, которая получена из матрицы подобия графа [14].

Алгоритм данного метода работает на графовой структуре данных. Матрица данных (попарных расстояний между объектами) при этом должна быть разрежена. Спектральный метод требует задания количества кластеров.

Стоит заметить, что с увеличением их числа качество кластеризации заметно понижается.

В основе спектрального метода лежит алгоритм нормализованного разделения (normalized cuts), основанного на вычислении собственного вектора симметричного нормализованного лапласиана, которому соответствует второе с конца по величине собственное число. При разделении объектов максимизируется вес кластеров.[17]

Для того, чтобы кластеризовать публикации, необходимо определить связь между ними.

Существует три метода определения связанности публикаций по цитированию:

- Прямое цитирование (когда одна статья имеет ссылку на определенную статью);
- Отношение на основе библиографической связи (когда две работы ссылаются на общую третью работу);
- Совместное цитирование (Если хотя бы один другой документ ссылается на два общих документа, считается, что эти документы цитируются совместно.).

Глава 2. Описание алгоритмов модуля извлечения контента

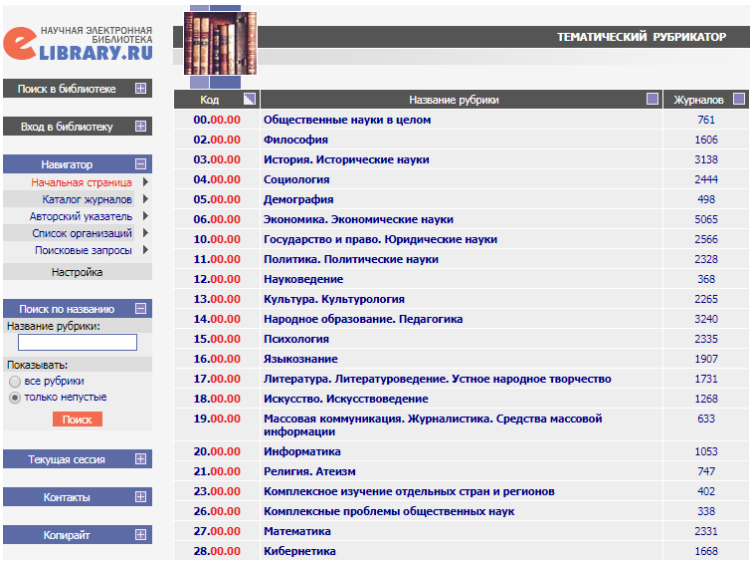
1. Описание источников первичной информации

Основными объектами системы извлечения контента из веб-ресурсов являются публикации из электронной библиотеки eLibrary. Задача будет состоять в том, чтобы выгрузить следующие данные о публикациях:

1. Название категории;
2. Название публикации;
3. Авторы;
4. Количество цитирований в системе РИНЦ;
5. Id процитированных статей.

В данной работе для анализа соберем статьи из двух категорий “Общественные науки в целом” и “Информатика”. Последовательно опишем шаги сбора данных по уровням вложенности электронной библиотеки:

1. На вход алгоритм получает адрес тематического рубрикатора. Здесь все публикации собраны по отдельным непересекающимся категориям. Главная страница тематического рубрикатора представлена на рисунке 2.



Код	Название рубрики	Журналов
00.00.00	Общественные науки в целом	761
02.00.00	Философия	1606
03.00.00	История. Исторические науки	3138
04.00.00	Социология	2444
05.00.00	Демография	498
06.00.00	Экономика. Экономические науки	5065
10.00.00	Государство и право. Юридические науки	2566
11.00.00	Политика. Политические науки	2328
12.00.00	Науковедение	368
13.00.00	Культура. Культурология	2265
14.00.00	Народное образование. Педагогика	3240
15.00.00	Психология	2335
16.00.00	Языкознание	1907
17.00.00	Литература. Литературоведение. Устное народное творчество	1731
18.00.00	Искусство. Искусствоведение	1268
19.00.00	Массовая коммуникация. Журналистика. Средства массовой информации	633
20.00.00	Информатика	1053
21.00.00	Религия. Атеизм	747
23.00.00	Комплексное изучение отдельных стран и регионов	402
26.00.00	Комплексные проблемы общественных наук	338
27.00.00	Математика	2331
28.00.00	Кибернетика	1668

Рисунок 2 Тематический рубрикатор

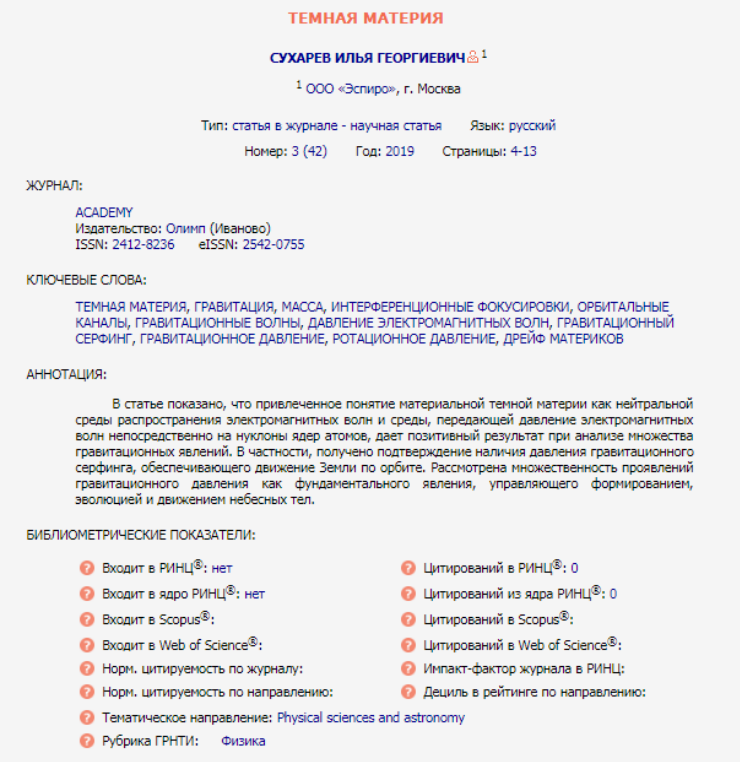
2. В каждой категории представлено конкретное количество журналов. На данном этапе нам необходимо собрать ссылки для перехода на следующий уровень, для получения данных о журналах (рис.3).

Рисунок 3 Список журналов одного из разделов тематического рубрикатора

3. На данном этапе мы видим список статей одного выпуска выбранного журнала. Чтобы получить статьи журнала (по всем выпускам), получим ссылки с боковой панели, где представлены выпуски журнала за различные года (рис.4).

Рисунок 4 Представление одного из выпусков журнала

4. На данном этапе мы собираем информацию о публикациях (рис.5).



ТЕМНАЯ МАТЕРИЯ

СУХАРЕВ ИЛЬЯ ГЕОРГИЕВИЧ [1]

[1] ООО «Эспиро», г. Москва

Тип: статья в журнале - научная статья Язык: русский

Номер: 3 (42) Год: 2019 Страницы: 4-13

ЖУРНАЛ:

ACADEMY
Издательство: Олимп (Иваново)
ISSN: 2412-8236 eISSN: 2542-0755

КЛЮЧЕВЫЕ СЛОВА:

ТЕМНАЯ МАТЕРИЯ, ГРАВИТАЦИЯ, МАССА, ИНТЕРФЕРЕНЦИОННЫЕ ФОКУСИРОВКИ, ОРБИТАЛЬНЫЕ КАНАЛЫ, ГРАВИТАЦИОННЫЕ ВОЛНЫ, ДАВЛЕНИЕ ЭЛЕКТРОМАГНИТНЫХ ВОЛН, ГРАВИТАЦИОННЫЙ СЕРФИНГ, ГРАВИТАЦИОННОЕ ДАВЛЕНИЕ, РОТАЦИОННОЕ ДАВЛЕНИЕ, ДРЕЙФ МАТЕРИКОВ

АННОТАЦИЯ:

В статье показано, что привлеченное понятие материальной темной материи как нейтральной среды распространения электромагнитных волн и среды, передающей давление электромагнитных волн непосредственно на нуклоны ядер атомов, дает позитивный результат при анализе множества гравитационных явлений. В частности, получено подтверждение наличия давления гравитационного серфинга, обеспечивающего движение Земли по орбите. Рассмотрена множественность проявлений гравитационного давления как фундаментального явления, управляющего формированием, эволюцией и движением небесных тел.

БИБЛИОМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ:

Входит в РИНЦ®: нет
Входит в ядро РИНЦ®: нет
Входит в Scopus®:
Входит в Web of Science®:
Норм. цитируемость по журналу:
Норм. цитируемость по направлению:
Тематическое направление: Physical sciences and astronomy
Рубрика ГРНТИ: Физика

Цитирований в РИНЦ®: 0
Цитирований из ядра РИНЦ®: 0
Цитирований в Scopus®:
Цитирований в Web of Science®:
Импакт-фактор журнала в РИНЦ:
Дециль в рейтинге по направлению:

Рисунок 5 Представление статьи

На данном этапе, мы можем получить количество цитирований статьи.

2. Алгоритмизация метода извлечения данных

Для реализации данной технологии будем использовать язык Python. Для извлечения контента нам необходимы http-запросы. Для отправки http-запросов существуют различные python-библиотеки, наиболее известные urllib/urllib2 и Requests. В данной работе будем использовать Requests.

В данном случае, мы не могли получить данные просто по URL. Первоначально нам было необходимо авторизоваться на сайте. Для этого необходимо было сконструировать запрос по информации из заголовков и значений JavaScript-переменных на странице.

Алгоритм получения данных состоит из 4 основных этапов, на выходе из которого мы должны получить систематизированные данные, загруженные в базу данных.

Алгоритм получения данных из электронной библиотеки eLibrary:

1. Авторизация на сайте

Авторизация на сайте была выполнения с помощью метода request.Session() (рис.2) .

```
post = "https://elibrary.ru/start_session.asp"
url = "https://elibrary.ru"

data = {"login": ' ',
        "password": '',
        "rpage": ''}

with requests.Session() as s:
    r = s.get(url, headers=headers)
    soup = BeautifulSoup(r.content, "lxml")
    r = s.post(post, data=data, headers=headers)
```

Рисунок 6 Запрос авторизации

2. Отправка http-запроса на сервер веб-ресурса;
3. Получение ответа в виде html-файла;
4. Парсинг файла.

Для работы с HTML-кодом, полученным по http-запросу, были использованы библиотеки Beautifulsoup и lxml. Приведем пример функции для получения информации о журналах и записи в соответствующую таблицу (рис.7).

```
# url - тематический раздел

def get_magazine(url):
    r = s.get(url, headers={'User-Agent': UserAgent().chrome})
    table = BeautifulSoup(r.content, "lxml").find(id = 'restab')
    row_massive = table.findAll('tr')[4:]
    for row in row_massive:
            id_mag = row.get('id')[1:]
            ISSN = row.findAll('td',align = 'center')[1].text
            name = row.find('td',align = 'left').find('b').string
            number = int(row.findAll('td',align = 'center')[2].text)
            link = 'https://elibrary.ru/' + row.find('td',align = 'left').find('a').get('href')
            cursor.execute("INSERT INTO desc_of_magazine VALUES (%s, %s, %s, %s %s)", (id_mag, ISSN, name, number,link))
```

Рисунок 7 Функция получения и записи данных о журналах

Структура данных не почти не имеет идентификаторов, поэтому, анализ данных по DOM-дереву является самым подходящим методом парсинга. Остальные функции получения данных были организованы похожим образом.

3. Выгрузка в базу данных

Результатом работы модулей изъятия контента являются HTML файлы. Общение с базой данных происходит с помощью модуля расширения cx-

Oracle. Данный модуль задействует несколько таблиц базы данных Oracle. Рассмотрим эти таблицы (рис.8)

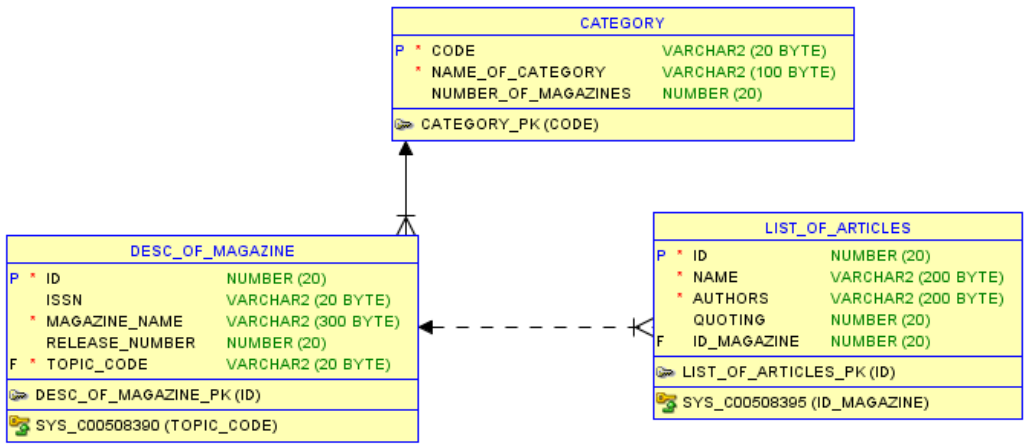


Рисунок 8 Модель используемых таблиц базы данных

Разработанный модуль использует три таблицы базы данных, в этих таблицах хранится информация о категориях журналов и статей, количество статей в журналах, информация о цитировании. Представим их более подробно. В таблице 1 представлено описание столбцов таблицы, содержащую категории журналов.

Таблица 1 Category

id	id категории
name	Название категории
number	Количество журналов в категории

В таблице 2 представлено описание столбцов таблицы, содержащую данные об отдельном журнале.

Таблица 2 desc_of_magazine

id	Id журнала
ISSN	Международный стандартный серийный номер
magazine_name	Название журнала
release_number	Количество выпусков
TOPIC_CODE	Код категории

В таблице 3 представлено описание таблицы, содержащую данные о статьях.

Таблица 3 list_of_articles

id	Id статьи
Name	Название статьи
Authors	Авторы
Citation	Id процитированных статей
code_magazine	Id журнала

Глава 3. Описание алгоритмов кластеризации

1. Предобработка данных

Чтобы кластеризовать публикации, необходимо определить связь между ними. В данной работе будем рассматривать связанность публикаций, основанную на отношениях прямого цитирования. Отношения прямого цитирования представляют особый интерес, поскольку они позволяют эффективно группировать большие наборы публикаций.

Получим связанность публикаций в виде матрицы смежности. Для определения родства N – публикаций мы используем c_{ij}, чтобы указать, ссылается ли публикация i на публикацию j ($c_{ij} = 1$) или нет ($c_{ij} = 0$).[19]

Отношения цитирования представляются в виде простого неориентированного и невзвешенного графа. В результате получим разреженную матрицу. Пример матрицы смежности представлен на рисунке 9.

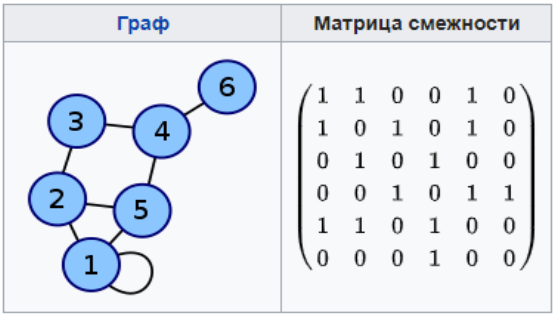


Рисунок 9 Пример матрицы смежности

2. Применение метода кластеризации

Для данной работы был выбран спектральный метод кластеризации. Данный метод часто используется для работы с графовой структурой данных. Матрица данных (попарных расстояний между объектами) при этом должна быть разрежена [20].

Для кластеризация будем использовать модуль *sklearn.cluster*. Будем использовать параметры:

- *n_clusters* - количество предполагаемых кластеров

- *affinity* - может быть одним из следующих объектов: «near_neighbors», «precomputed», «rbf» или одно из ядер, поддерживаемых sklearn.metrics.pairwise_kernels. Следует использовать только те ядра, которые дают оценки сходства (неотрицательные значения, которые увеличиваются со сходством). Это свойство не проверяется алгоритмом кластеризации.
- *n_init* - количество раз, когда алгоритм k-средних будет запускаться с разными центроидами. Окончательный результат будет наилучшим результатом последовательных прогонов n_init

В результате кластеризации лучшие результаты были получены при использовании *n_init* равным 100. Пример кластеризации представлен на рисунке 9.

```
# Cluster
sc = SpectralClustering(2, affinity='precomputed', n_init=100)
sc.fit(adj_mat)
```

Рисунок 10 Пример кластеризации

Для проверки точности кластеризации были использованы метрики:

- sklearn.metrics.adjusted_mutual_info_score(*labels_true*, *labels_pred*)
- sklearn.metrics.adjusted_rand_score(*labels_true*, *labels_pred*), где

labels_true – исходные категории статей;

labels_pred – кластер, к которому была отнесена статья после применения кластеризациию.

Проверим точность кластеризации с разным числом кластером. Точность, полученная с использованием метрики sklearn.metrics.adjusted_rand_score была получена выше, чем с sklearn.metrics.adjusted_rand_score. Представим точность кластеризации (метрика sklearn.metrics.adjusted_rand_score) в таблице 4.

Таблица 4 Точность кластеризации

Количество кластеров	**Точность**
2	0.612
3	0.591
4	0.587
5	0.450
6	0.418

Наилучший результат был получен для двух кластеров, что совпадает с исходными данными, т.к для анализа мы использовали статьи из двух категорий.

Общий вывод по разделу

В ходе выполнения работы, была произведена отработка технологии сбора, анализа и кластеризации данных на основе открытого интернет-источника**.** Были разработаны функции извлечения научных публикаций из открытой электронной библиотеки eLibrary, была произведена подготовка полученных данных к кластеризации, и реализована кластеризация публикаций по цитированию.

Для реализации функций сбора данных были проанализированы существующие решения. Был произведен анализ подходов и инструментов web-mining. Рассмотрели методы для нахождения взаимосвязанности публикаций, и методы кластеризации, подходящие для работы с получившимся результатом. Также разработали систему хранения полученных публикаций в базе данных.

В результате оценки точности кластеризации, можно сделать вывод, что данная технология может быть применима как инструментальное средство анализа данных для обнаружения закономерностей в большом объеме слабоструктурированной информации.

Глава 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Целью выполнения данного раздела является анализ и выявление потенциальной выгоды реализации технологии сбора, анализа и кластеризации публикаций.

Задачи, которые следуют из данной цели:

1. Провести анализ конкурентоспособности методов кластеризации;
2. Разработать диаграмму Исикавы;
3. Провести оценку готовности проекта к коммерциализации;
4. Составить цели проекта, определить ожидаемые результаты;
5. Спланировать работы, распределить задачи между участниками проекта;
6. Сформировать бюджет затрат;
7. Провести анализ рисков проекта.

1. Предпроектный анализ

Потенциальные потребители результатов исследования

Объем научной информации, представленный в открытом доступе и перешедший в категорию больших данных, усложняет анализ, уточнение и корректировку научно-технических приоритетов на уровне государства.

Актуальной является задача разработки инструментальных средств анализа больших объемов данных для автоматизации процессов классификации и оценки значимости научных текстов, выявления степени связанности и взаимного влияния перспективных направлений исследований и визуализации структуры научной деятельности с целью поддержки принятия решений в рамках программ государственной поддержки научной, научно-технической и инновационной деятельности.

Потребителями разработанной технологии могут быть:

- наукометрические базы данных
- академические издательства
- базы цитирований

- научные организации
- интернет ресурсы, научное сообщество (университеты)

2. Анализ конкурентных технических решений с позиции ресурсоэффективности и ресурсосбережения

Сравним выбранный подход для выявления кластеров научных публикации на основе цитирования с двумя другими распространенными методами.

Классический подход к решению задач кластеризации и классификации публикаций основывается на анализе тезауруса обрабатываемой предметной области. Но данный метод имеет существенный недостаток: таким образом нельзя производить индексирование корпусов текстов произвольных тематик. Более того, если рассматривать обработку корпусов текстов достаточно узких тематик, то в таких случаях требуются очень подробные тезаурусы, которые имеются далеко не для всех предметных областей.

Подход, основанный на извлечении ключевых выражений без априорных ограничений, носит более универсальный характер, хотя, несколько проигрывает в адекватности индексирования. Также разработка эффективного алгоритма автоматизации извлечения ключевых слов является нетривиальной задачей, так как необходимо учитывать и те случаи, когда слова, образующие термин (т.е. ключевое слово) находится не только в именительном, но и в косвенных падежах.

Кластеризация на основе прямого цитирования представляет большой интерес, так как она позволяет эффективно группировать большие наборы публикаций. Более подробное сравнение приведем в оценочной карте в таблице 1.

$Б_ф$ – разработка, представленная в диссертации (кластеризация публикаций на основе цитирования)

$Б_{к1}$ – кластеризация публикаций на основе тезауруса предметной области

$Б_{к2}$ – кластеризация публикаций на основе ключевых выражений

Таблица 5. Оценочная карта для сравнения методов кластеризации

Критерии оценки	**Вес критерия**	**Баллы**			**Конкурентоспособность**		
		$Б_ф$	$Б_{к1}$	$Б_{к2}$	$К_ф$	$К_{k1}$	$К_{k2}$
1	2	3	4	5	6	7	8
Технические критерии оценки ресурсоэффективности							
1. Удобство расчёта	0.1	5	4	3	0.5	0.4	0.3
2. Затраты компьютерных ресурсов для расчёта	0.05	4	4	4	0.20	0.20	0.20
3. Скорость получения результатов	0.1	5	4	4	0.5	0.4	0.4
4. Смысловая нагрузка критерия	0.5	5	5	4	2.5	2.5	2.0
Экономические критерии оценки эффективности							
1. Конкурентоспособность продукта	0.05	4	4	4	0.2	0.2	0.2
2. Уровень проникновения на рынок	0.1	3	5	5	0.3	0.5	0.5
3. Затраты на разработку	0.1	1	1	1	0.1	0.1	0.1
Итого	1				4.3	4.3	3.7

В результате полученной оценочной карты, разработка, представленная в диссертации $Б_ф$ и разработка конкурентов $Б_{к1}$ получили одинаковые результаты. Плюсы разработки $Б_ф$ заключаются в удобстве расчета и скорости получения результата, что позволит производить расчет по большему количеству данных. Это является основным конкурентным преимущество. Уровень проникновения на рынок низкий, что дает возможность развивать данный метод для более широкого использования.

Разработка $Б_{к2}$ показала результаты ниже $Б_ф$ и $Б_{к2}$, в главной мере это обусловлено более низкой скоростью расчета и смысловой нагрузкой критерия.

3. Диаграмма Исикавы

Диаграмма причины-следствия Исикавы – это графический метод анализа и формирования причинно-следственных связей, инструментальное

средство для систематического определения причин проблемы и последующего графического представления.

Область применения диаграммы:

- Выявление причин возникновения проблемы
- Анализ и структурирование процессов на предприятии
- Оценка причинно-следственных связей

Объектом анализа в данном случае являются ошибочные критерия оценки кластеризации.

Диаграмма Исикавы состоит из центральной вертикальной стрелки, которая представляет следствие (ошибочные критерии оценки), и подходящих к ней крупных "ребер", которые называют причинами первого порядка. К этим "ребрам" подходят стрелки поменьше, называемые причинами второго порядка, к ним - еще более мелкие - причины третьего порядка.

Были выделены факторы/группы факторов, влияющие на объект анализа (факторы - стрелки первого уровня):

- Дистрибутивы
- Методы анализа
- Оборудование
- Доступ к интернету
- Данные

Далее эти факторы были проанализированы для выявления факторов второго и третьего порядка, представленных на рисунке 1.

4. SWOT-анализ

Метод SWOT анализа — универсальная методика стратегического менеджмента. Первым шагом в проведении SWOT анализа является определение сильных и слабых сторон товара или услуги. Вторым шагом SWOT анализа является определение возможностей и угроз в будущем.

В таблице 2 представлен SWOT анализ для разработки технологии сбора, анализа и классификации, представленной в данной работе.

Таблица 6. SWOT анализ НТИ

	Сильные стороны: С1. Высокая смысловая нагрузка разрабатываемого критерия. С2. Работа с большим количеством данных С3.Широкий охват предметной области.	**Слабые стороны:** Сл1. Отсутствие прототипа научной разработки. Сл2. Узкая направленность на определенную структуру данных. Сл3. Отсутствие полного доступа к требуемым данным.
Возможности: В1. Требование улучшения существующих систем поиска в электронных библиотеках со стороны пользователей В2. Появление нового критерия оценки цитирования публикаций	Высокая смысловая нагрузка критерия позволит улучшить систему поиска в электронных библиотеках При появлении нового критерия цитирования, данный метод поможет обработать и кластеризовать данные быстрее конкурирующих разработок	При появления нового критерия оценки, можно будет расширить систему сбора данных, предназначенную для более широкой структуры данных.
Угрозы: У1. Незаинтересованность научного сообщества в новых, более полных критериях. У2. Появление новых разработок, учитывающих более широкий спектр факторов и взаимосвязей. У3. Введения дополнительных государственных требований к оценке научных работ. У4. Развитая конкуренция технологий	Высокая смысловая нагрузка критерия поможет разработать новую систему обработки данных, при появлении дополнительных требований к оценке работ.	Отсутствие доступа к требуемым данным ведет к понижению точности разработки, в следствии это ведет к уходу от предложенного продукта к продуктам конкурентов. Разработанная система не будет соответствовать требованиям

Таким образом, можно прийти к выводу, что основными рисками при дальнейшей разработке и продвижении системы являются развитая конкуренция технологий и введения дополнительных государственных требований к оценке научных работ.

5. Оценка готовности проекта к коммерциализации

Для контроля выполнения работы всегда необходимо оценивать степень готовности. Для этого необходимо заполнить специальную форму, содержащую показатели о степени проработанности проекта с позиции коммерциализации и компетенциям разработчика научного проекта. Форма для нашего проекта представлена в таблице 3.

Таблица 7. Оценка степени готовности научного проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1	Определен имеющийся научно-технический задел	5	5
2	Определены перспективные направления коммерциализации научно-технического задела	5	5
3	Определены отрасли и технологии (товары, услуги) для предложения на рынке	5	4
4	Определена товарная форма научно-технического задела для представления на рынок	4	2
5	Определены авторы и осуществлена охрана их прав	3	5
6	Проведена оценка стоимости интеллектуальной собственности	2	3

7	Проведены маркетинговые исследования рынков сбыта	5	5
8	Разработан бизнес-план коммерциализации научной разработки	2	2
9	Определены пути продвижения научной разработки на рынок	5	4
10	Разработана стратегия (форма) реализации научной разработки	3	3
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	2	3
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	1	2
13	Проработаны вопросы финансирования коммерциализации научной разработки	5	2
14	Имеется команда для коммерциализации научной разработки	3	4
15	Проработан механизм реализации научного проекта	5	4
	ИТОГО БАЛЛОВ	48	47

Перспективность данной разработки выше среднего. Для её реализаций разработчику требуются высокие знание в программировании и команда опытных программистов, которые помогут оптимизировать и ускорить все требуемые расчёты. Для оплаты труда разработчиков требуются небольшие финансовые затраты. Для коммерциализации результатов научно-технического исследования идеальным решением будет «Торговля патентными лицензиями.

6. Инициация проект

Целью данного проекта является разработка технологии сбора, анализа и кластеризации данных. Разработка моделей классификации научных публикаций на основе цитирования поможет усовершенствовать систему поиска необходимых публикаций в открытых интернет источниках и улучшить его качество. В таблице 4 представлена информация о заинтересованных сторонах проекта.

Таблица 8. Заинтересованные стороны проекта

Заинтересованные стороны проекта	**Ожидания заинтересованных сторон**
ТПУ	Качественная система поиска публикаций, которая позволит ускорить получение результатов и улучшить его качество .
Elibrary (сеть цитирования РИНЦ)	Быстрый, не затратный, несущий высокую смысловую нагрузку критерий, позволяющий сгруппировать научные работы, и в результате улучшить систему поиска схожих публикаций**.**

В таблице 9 представлена информация об иерархии целей проекта и критериях достижения целей.

Таблица 9. Цели и результат проекта

Цели проекта:	Разработка технологии сбора, анализа и кластеризации данных.
Ожидаемые результаты проекта:	Программное обеспечение, позволяющее производить анализ

	большого количества научных работ.
Критерии приемки результата проекта:	Точность результатов кластеризации.
	Требование:
Требования к результату проекта:	Низкое потребление ресурсов.
	Высокая скорость расчёта.
	Высокая смысловая нагрузка критерия.

7. Планирование управления научно-техническим проектом

Группа процессов планирования состоит из процессов, осуществляемых для определения общего содержания работ, уточнения целей и разработки последовательности действий, требуемых для достижения целей. На рисунке 2 представлена иерархическая структура работ по проекту разработки технологии сбора, анализа и классификации данных.

Таблица 10. План проекта

Название	**Длительность, дни**	**Дата начала работ**	**Дата окончания работ**	**Состав участников (ФИО ответственных исполнителей)**
Выбор темы ВКР	5	20.12.2018	25.12.2018	Демидова О.О., Савельев А.О.
Получение технического задания	4	26.12.2018	30.12.2018	Савельев А.О. Демидова О.О
Подбор материала, его анализ и обобщение	8	01.01.2019	09.01.2019	Демидова О.О.
Выбор метода выполнения работы	6	10.01.2019	17.01.2019	Демидова О.О.

Календарное планирование работ по теме	10	17.01.2019	27.01.2019	Демидова О.О.,
Первичный анализ данных	13	28.01.2019	11.02.2019	Демидова О.О.
Разработка программного продукта	40	11.02.2019	08.04.2019	Демидова
Тестирование и выявление недочётов	16	08.04.2019	24.04.2019	Демидова О.О., Савельев А.О.
Исправление найденных ошибок, доработка модели	9	24.04.2019	04.05.2019	Демидова О.О.
Составление отчёта о проделанной работе с полным анализом результатов	15	04.05.2019	20.05.2019	Демидова О.О.
Итого:	126			

Диаграмма Ганта используется для иллюстрации календарного плана работы, на котором работы по теме представляются в виде протяженных по времени отрезках. График строиться в виде таблицы с разбивкой по месяцам и декадам.

В таблице 11 представлен Календарный план-график по проведения НТИ.

Таблица 11. Календарный план-график проведения НТИ

	Вид работ	Исполнители	$T_к$, ал, дн.	Декабрь			Январь			Февраль			Март			Апрель			Май		
	Выбор темы ВКР	Руководитель, Инженер	5																		
	Получение тех. задания	Руководитель, Инжене	4																		
	Подбор материала его анализ и обобщение	Инженер	8																		
	Выбор метода выполнения работы	Инженер	6																		
	Календарное планирование работ по теме	Инженер	0																		
	Первичный анализ данных	Инженер	3																		
	Разработка программы	Инженер	40																		
	Тестирование, выявление недочётов	Руководитель Инженер	16																		
	Исправление найденных багов, доработка модели	Инженер	9																		
0	Составление отчёта о проделанной работе с полным анализом результатов	Инженер	15																		

8. Бюджет научного исследования

Для полноты и достоверности учета всех расходов сгруппируем все затраты по следующим статьям

- затраты на материалы
- затраты на амортизацию
- основная заработная плата исполнителей
- дополнительная заработная плата исполнителей темы
- отчисления во внебюджетные фонды (страховые отчисления)
- накладные расходы

Материальные затраты

В расчет взяты затраты на канцелярские товары в размере 1000 рублей.

Амортизационные отчисления

Для работы над проектом использовался ноутбук. Амортизацию рассчитаем линейным способом.

Первоначальная стоимость ПК 60000 рублей; срок полезного использования для машин офисных код 330.28.23.23 составляет 2-3 года, берем 3 года; планируется использовать ПК для написания работы в течение 6 месяцев. Тогда:

- месячная норма амортизации:

$$\text{Ан} = \frac{1}{n} * 100\% = \frac{1}{12 \times 3} \times 100\% = 2{,}8\%$$

где n - количество месяцев полезного срока эксплуатации ОС.

- ежемесячные амортизационные отчисления:

$$\text{Аг} = 60000 \times 2{,}8 = 1\,680 \text{ рублей}$$

- итоговая сумма амортизации основных средств:

$$\text{А} = 1680 \times 6 = 10080 \text{ рублей}$$

Таким образом, в материальные затраты необходимо включить сумму амортизации основных средств в сумме 10080 руб.

Заработная плата исполнителей проекта

Заработная плата рассчитывается из суммы заработной платы исполнителя и научного руководителя исходя из трудоемкости каждого этапа и занятости каждого из них на данном этапе по формуле

$$З_{зп} = З_{осн} + З_{доп}$$

где $З_{осн}$ – основная заработная плата; $З_{доп}$ – дополнительная заработная плата.

Рассчитаем основную заработную плату:

$$З_{осн} = З_{дн} \times Т_р \times (1 + К_{пр} + К_д) \times К_р$$

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

$К_{пр}$ – премиальный коэффициент (т.е. 30% от $З_{дн}$);

$К_д$ – коэффициент доплат и надбавок составляет примерно 0,2 – 0,5 (в НИИ и на промышленных предприятиях – за расширение сфер обслуживания, за профессиональное мастерство, за вредные условия: 15-20% от $З_{дн}$);

$К_р$ – районный коэффициент (для Томска 1,3);

$Т_р$ – продолжительность работ, выполняемых работником, раб. дни.

Рассчитаем среднедневную заработную плату по формуле:

$$З_{дн} = \frac{З_м \times М}{F_д}$$

$З_м$ – оклад работника за месяц, 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 персонала, раб. дн.

Таблица 12 – Баланс рабочего времени (для 6-дневной недели)

Показатели рабочего времени	Дни
Календарные дни	365
Нерабочие дни (праздники/выходные)	66
Потери рабочего времени (отпуск/невыходы по болезни)	56
Действительный годовой фонд рабочего времени	243

Для расчета основной заработной платы инженера берем оклад, равный окладу 21760 руб. Для расчета основной заработной платы руководителя в расчет возьмем оклад, равный 33664 руб.

Таблица 13 – Расчет основной заработной платы

Исполнители	$З_{дн}$, руб.	$К_{пр}$	$К_{д}$	$К_{р}$	$Т_{р}$	$З_{осн}$, руб.
Инженер	931	0	0	1,3	103	124 661
Научный руководитель	1440	0,3	0,2	1,3	21	58 968

В дополнительную заработную плату входят суммы выплат, предусмотренные трудовым кодексом, например, оплата ежегодных и дополнительных отпусков, оплата времени, связанного с выполнением государственных и общественных обязанностей и т.д. запланируем дополнительную заработную плату в размере 15 % от основной заработной платы исполнителей,

В таблице 14 представлен расчет затрат на заработную плану исполнителей.

Таблица 14 – Затраты на заработную плату без отчислений

Исполнители	$З_{осн}$, руб.	$З_{доп}$, руб.	$З_{зп}$, руб.
Инженер	124 661	18 699	143 360
Научный руководитель	58 968	8 845	67 813
Итого	183 629	27 544	211 173

Отчисления во внебюджетные фонды (страховые отчисления)
Общие тарифы страховых взносов в 2019 году в ИФНС:

- 22% — на пенсионное страхование;
- 2,9% — страхование по временной нетрудоспособности;
- 5,1% — медицинское страхование.

Величина отчислений во внебюджетные фонды определяется исходя из формулы:

$$З_{внеб} = k_{внеб} * (З_{осн} + З_{доп}),$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.).

Таким образом, с учетом результатов расчета зарплат на заработную плату величина отчислений во внебюджетные фонды составляет:

$$З_{внеб} = 0{,}3 * 307044 = 63352 \text{ руб}$$

Накладные расходы

При выполнении проекта могут возникнуть косвенные издержки – накладные расходы, возникающие дополнительно к основным затратам, например, на консультационные услуги, оплату коммунальных услуг, расходы на услуги связи (телефон, интернет) и так далее.

Их величина определяется по следующей формуле:

$$З_{накл} = (\text{сумма статей расходов}) \cdot k_{нр};$$

где $k_{нр}$ – коэффициент, учитывающий накладные расходы.

Величину коэффициента накладных расходов можно взять в размере 16%.

$$З_{накл} = (1\,000 + 10\,080 + 211173 + 63352) \cdot 0{,}16 = 45\,697 \text{ рублей.}$$

Формирование бюджета

После выполнения всех расчетов по статьям можно определить плановую общую себестоимость проекта:

Таблица 15 – Бюджет затрат

Наименование	Сумма, руб.	в %
Затраты на материалы	1000	0,30%
Затраты на амортизацию	10080	3,04%
Затраты на основную заработную плату	183629	55,43%
Затраты на дополнительную заработную плату	27544	8,31%
Страховые взносы	63352	19,12%
Накладные расходы	45697	13,79%
Общий бюджет	331302	100,00%

Исходя из расчета бюджета затрат следует, что наибольшая его часть приходится на основную заработную плату исполнителей (55,43 %). Также необходимо отметить, что расходы на страховые взносы (19,12 %) составляют немаловажную часть расходов. Затраты на амортизацию и материалы составляют небольшую долю (суммарно 3,34 %). Это связано с отсутствие необходимости использования дорогостоящего оборудования и материалов.

9. Реестр рисков проекта

Риски проекта включают в себя возможные неопределенные события, которые могут возникнуть в проекте и вызвать последствия, которые повлекут нежелательные эффекты.

В результате можно сказать, что риск отсутствия соединения с сетью интернет является самым опасным для нашего проекта.

С позиций ресурсной и финансовой ресурсоэффективности можно сделать вывод о том, что научно – техническое решение, разрабатываемое нами, является достаточно эффективным. В результате проведенной работы были выявлены потенциальные потребители, которые могут быть

заинтересованы в данной разработке, так как она поможет улучшить качественные характеристики систем поиска и кластеризации наукометрических баз данных.

В результате составления оценочной карты сравнили разработку, представленную в данной работе, с конкурирующими разработками. Составление диаграммы Исикавы помогло выяснить возможные варианты проблем, на которые следует обратить внимание.

Выяснили сильные и слабые стороны НТИ с помощью SWOT анализа.

Был оценен бюджет научного исследования, величина которого составила 331302 руб. Разработка не требует больших финансовых затрат. Низкое потребление ресурсов и высокая скорость расчета позволят конкурировать с другими технологиями кластеризации.

Результат данной работы может помочь в принятиях решений в рамках программ государственной поддержки научной, научно-технической и инновационной деятельности. Так как поможет выявлять научные тенденции в следствия правильного анализа данных.

Глава 5. Социальная ответственность

Введение

Целью данной магистерской работы является разработка инструментальных средств анализа больших объемов данных для автоматизации процессов классификации и оценки значимости научных текстов, выявления степени связанности и взаимного влияния перспективных направлений исследований.

Основными задачами являются:

- Программная реализация прототипа системы извлечения знаний из большого объема слабоструктурированных данных
- Разработка модели классификации научных публикаций и методики их использования для различных областей знаний.

Следовательно, работу можно классифицировать как работу разработчика программного обеспечения.

С появлением компьютеров произошли серьезные изменения в условиях производственной деятельности работников умственного труда. Их труд стал более интенсивным, напряженным, требующим значительных затрат умственной, эмоциональной и физической энергии.

Обеспечение безопасной жизнедеятельности человека в значительной степени зависит от правильной оценки опасных, вредных производственных факторов. Одинаковые по тяжести изменения в организме человека могут быть вызваны различными причинами. Это могут быть какие-либо факторы производственной среды, чрезмерная физическая и умственная нагрузка, нервно-эмоциональное напряжение, а также разное сочетание этих причин.

В данном разделе рассмотрены вопросы безопасной жизнедеятельности на этапе разработки технологии для сбора, анализа и кластеризации данных. Так как все работы на данном этапе велись в аудитории за компьютером, необходимо рассмотреть вредные факторы,

связанные с этим видом работ, воздействие на окружающую среду и возможные чрезвычайные ситуации.

1. Правовые и организационные вопросы обеспечения безопасности

Организационные мероприятия при компоновке рабочей зоны

При организации рабочего места необходимо учитывать требования безопасности, промышленной санитарии, эргономики, технической эстетики. Несоблюдение требований безопасности приводит к тому, что при работе за компьютером сотрудник может ощущать дискомфорт: возникают головные боли и резь в глазах, появляются усталость и раздражительность. У некоторых людей нарушается сон, аппетит, ухудшается зрение, начинают болеть руки, шея, поясница и тому подобное. При ненормированной работе возможно нервное истощение[10].

При организации работы на ПЭВМ должны выполняться следующие условия:

- рабочее место с персональным компьютером (ПК) должно располагаться по отношению к оконным проемам так, чтобы свет падал сбоку, предпочтительнее слева;
- нужно избегать расположения рабочего места в углах комнаты или лицом к стене (расстояние от ПК до стены должно быть не менее 1 м), экраном и лицом к окну;
- ПК желательно устанавливать так, чтобы, подняв глаза от экрана, можно было увидеть самый удаленный предмет в комнате, так как перевод взгляда на дальнее расстояние – один из самых эффективных способов разгрузки зрительной системы при работе на ПК;
- при наличии нескольких компьютеров расстояние между экраном одного монитора и задней стенкой другого должно быть не менее 2 м, а расстояние между боковыми стенками соседних мониторов – не менее 1,2 м [12];
- окна в помещениях с ПЭВМ должны быть оборудованы

регулируемыми устройствами (жалюзи, занавески, внешние козырьки и т.д.);

- монитор, клавиатура и корпус компьютера должны находиться прямо перед оператором; высота рабочего стола с клавиатурой должна составлять 680–800 мм над уровнем стола; а высота экрана (над полом) 900-1280 см;
- монитор должен находиться от оператора на расстоянии 60–70 см на 20 градусов ниже уровня глаз [11];
- пространство для ног должно быть: высотой не менее 600 мм, шириной не менее 500 мм, глубиной не менее 450 мм. Должна быть предусмотрена подставка для ног работающего шириной не менее 300 мм с регулировкой угла наклона 0-20 градусов;
- рабочее кресло должно иметь мягкое сиденье и спинку, с регулировкой сиденья по высоте, с удобной опорой для поясницы;
- Следовать руководству.
- Положение тела пользователя относительно монитора должно соответствовать направлению просмотра под прямым углом или под углом 75 градусов [13].

Правильная поза и положение рук оператора являются весьма важными для исключения нарушений в опорно-двигательном аппарате и возникновения синдрома постоянных нагрузок.

Согласно СанПиНу 2.2.2.542-96 при 8-ми часовой рабочей смене на ВДТ и ПЭВМ перерывы в работе должны составлять от 10 до 20 минут каждые два часа работы.

2. Особенности законодательного регулирования проектных решений

При работе с персональным компьютером очень важную роль играет соблюдение правильного режима труда и отдыха.

Вид трудовой деятельности при проведении исследований в лаборатории за компьютером входит в группу В – творческая работа в режиме с диалогом ПЭВМ. Категория тяжести и напряженности работы с

ПЭВМВ определяется в зависимости от суммарного времени непосредственной работы с ПЭВМ за рабочую смену, но не более 6 ч за смену. В табл. 16 представлены сведения о регламентированных перерывах, которые необходимо делать при работе на компьютере, в зависимости от продолжительности рабочей смены, видов и категорий трудовой деятельности с ВДТ (видеодисплейный терминал) и ПЭВМ [10].

Таблица 16 Время регламентированных перерывов при работе на компьютере

Категория работы с ВДТ или ПЭВМ	Уровень нагрузки за рабочую смену при видах работы с ВДТ	Суммарное время регламентированных перерывов, мин
	Группа В, часов	При 8-часовой смене
I	до 2,0	30
II	до 4,0	50
III	до 6,0	70

Время перерывов дано при соблюдении указанных Санитарных правил и норм. При несоответствии фактических условий труда требованиям Санитарных правил и норм время регламентированных перерывов следует увеличить на 30%.

Эффективность перерывов повышается при сочетании с производственной гимнастикой или организации специального помещения для отдыха персонала с удобной мягкой мебелью, аквариумом, зеленой зоной и т.п.

В данном разделе дипломной работы приведены: оценка условий труда на рабочем месте, анализ вредных и опасных факторов труда, разработка мер защиты от них. Темой выпускной работы является «Проектирование и разработка системы поиска и кластеризации научно-технических публикаций на основе информации из открытых Интернет-источников». Объектом исследования выступает рабочее место, оборудование, помещение, в котором

находится это рабочее место. Все исследования производились во время работы над дипломным проектом в помещении, где выполнялась эта работа. Лаборатория, в которой проводилось исследование и разработка алгоритмов находится в научно-технической библиотеке Томского Политехнического университета (НТБ ТПУ).

Полностью безопасных и безвредных производств не бывает, поэтому с целью уменьшения воздействия различных неблагоприятных факторов прибегают к такой дисциплине как охрана труда.

Основным оборудованием для выполнения исследований является компьютер. При этом, опасным для разработчика фактором является высокое напряжение в электрической сети и как следствие, опасность поражения электрическим током. Напряжение в сети составляет 220В при частоте 50Гц, что является смертельно опасным в случае поражения работающего электрическим током.

К вредным производственным факторам, при работе с компьютером следует отнести:

1. повышенный уровень электромагнитных излучений, основными источниками которых является монитор компьютера;
2. отклонение показателей микроклимата
3. повышенный уровень шума, источниками которого являются вентиляторы внутри системного блока и блока питания компьютера, накопители на жестких и магнитных дисках, светильники люминесцентных ламп и др.
4. повышенный уровень ионизирующих излучений, источником которых является дисплей монитора компьютера
5. недостаточная освещённость рабочей зоны [1]

3. Повышенный уровень электромагнитных излучений

Как любые электрические приборы, видеотерминалы (ВДТ) и системные блоки производят электромагнитное излучение, воздействие которого на человека зависит от напряжённостей электрического и

магнитного полей, потока энергии, частоты колебаний, размера облучаемого тела.

Нарушения в организме человека при воздействии электромагнитных полей незначительных напряженностей носят обратимый характер. При воздействии полей, имеющих напряженность выше предельно допустимого уровня, развиваются нарушения со стороны нервной, сердечно-сосудистой систем, органов пищеварения и некоторых биологических показателей крови.

Большая часть электромагнитных излучений происходит не от экрана монитора, а от видеокабеля и системного блока. В портативных компьютерах практически всё электромагнитное излучение идет от системного блока, располагающегося под клавиатурой. Современные машины выпускаются заводом-изготовителем со специальной металлической защитой внутри системного блока для уменьшения фона электромагнитного излучения.

Напряженность электромагнитного поля на расстоянии 50см вокруг ВДТ по электрической составляющей должна быть не более:

- В диапазоне частот 5 Гц2 кГц 25 В/м;
- В диапазоне частот 2 кГц400кГц2,5 В/м.

Плотность магнитного потока должна быть не более:

- В диапазоне частот 5 Гц2 кГц 250 нТл;
- В диапазоне частот 2 кГц400кГц 25 нТл.

Возможные способы защиты от ЭМП:

Основной способ увеличение расстояния от источника, экран видеомонитора должен находиться на расстоянии не менее 50 см от пользователя;

Применение приэкранных фильтров, специальных экранов и других средств индивидуальной защиты, прошедших испытание в аккредитованных лабораториях и имеющих соответствующий гигиенический сертификат.

4. Отклонение показателей микроклимата

Проанализируем микроклимат в помещении, где находится рабочее место. Воздух рабочей зоны (микроклимат) производственных помещений

определяют следующие параметры: температура, относительная влажность, скорость движения воздуха. Параметры микроклимата оказывают непосредственное влияние на тепловое самочувствие человека и его работоспособность.

Например, понижение температуры и повышение скорости воздуха может привести к переохлаждению организма.

При повышении температуры воздуха возникают обратные явления. Исследователями установлено, что при температуре воздуха более 30 ^{0}C работоспособность человека начинает падать.

Для человека определены максимальные температуры в зависимости от длительности их воздействия и используемых средств защиты. Предельная температура вдыхаемого воздуха, при которой человек в состоянии дышать в течение нескольких минут без специальных средств защиты, около 116 ^{0}C.

Существенное значение имеет равномерность температуры. Вертикальный градиент ее не должен выходить за пределы 5 ^{0}C/метр.

Переносимость человеком температуры, как и его теплоощущение, в значительной мере зависит от влажности окружающего воздуха. Чем больше относительная влажность, тем меньше испаряется пота в единицу времени и тем быстрее наступает перегрев тела. Особенно неблагоприятное воздействие на тепловое самочувствие человека оказывает высокая влажность при t_{oc} > 30 ^{0}C, так как при этом почти все выделяемая теплота отдается в окружающую среду при испарении пота.

Недостаточная влажность воздуха также может оказаться неблагоприятной для человека вследствие интенсивного испарения влаги со слизистых оболочек, их пересыхания и растрескивания, а затем и загрязнения болезнетворными микроорганизмами. Поэтому при длительном пребывании людей в закрытых помещениях рекомендуется ограничиваться относительной влажностью в пределах 3070 %.

Вместе с потом организм теряет значительное количество минеральных солей (до 1%, в том числе 0,4 0,6 NaCl). При неблагоприятных условиях

потеря жидкости может достигать 810 л за смену и в ней до 60 г поваренной соли (всего в организме около 140 г NaCl). Потеря соли лишает кровь способности удерживать воду и приводит к нарушению деятельности сердечно-сосудистой системы.

Атмосферное давление оказывает существенное влияние на процесс дыхания и самочувствие человека. При работе в условиях избыточного давления снижаются показатели вентиляции легких за счет некоторого урежения частоты дыхания и пульса. Длительное пребывание при избыточном давлении приводит к токсическому действию некоторых газов, входящих в состав вдыхаемого воздуха. Оно проявляется в нарушении координации движений, возбуждении или угнетении, галлюцинациях, ослаблении памяти, расстройстве зрения и слуха [2].

Оптимальные значения характеристик микроклимата приведены в таблицах 15 и 16.

По степени физической тяжести работа инженера-программиста относится к лёгкой физической работе категории I а, с энергозатратами организма до 120 Дж/с, т.к. работа проводилась сидя, не требуя систематического физического напряжения.

Таблица 17 – Оптимальные значения характеристик микроклимата

Период года	Категория работ по уровню энергозатрат, Вт	Температура воздуха, °С	Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	Ia (до 139)	22 - 24	21 - 25	60 - 40	0,1
Теплый	Ia (до 139)	23 - 25	22 - 26	60 - 40	0,1

Таблица 18 –Допустимые значения характеристик микроклимата

Период года	Категория работ по уровню энергозатрат, Вт	Температура воздуха, °С	Температура поверхностей°С	Относит. влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	Iа(до 139)	20,0-25,0	19,0-26,0	15-75	0,1
Теплый	Iа (до 139)	21,0-28,0	20,0-29,0	15-75	0,1-0,2

Параметры микроклимата в помещении, где находится рабочее место, регулируются системой центрального отопления и приточно-вытяжной вентиляцией, и имеют следующие значения: влажность 40%, скорость движения воздуха 0,1 м/с, температура летом 20-25°С, зимой 20-22°С, что соответствует требованиям, представленным в таблице 3.

К мероприятиям по оздоровлению воздушной среды в производственном помещении относятся: правильная организация вентиляции и кондиционирования воздуха, отопление помещений. Вентиляция может осуществляться естественным и механическим путём. В рабочем помещении должны подаваться следующие объёмы наружного воздуха: при объёме помещения до 20м^3 на человека – не менее 30м^3 в час на человека; при объёме помещения более 40м^3 на человека и отсутствии выделения вредных веществ допускается естественная вентиляция.

В аудитории отсутствует принудительная вентиляция. Имеется лишь естественная, т.е. воздух поступает и удаляется через щели, окна, двери. Основной недостаток такой вентиляции в том, что приточный воздух поступает в помещение без предварительной очистки и нагревания. Естественная вентиляция допускается при условии, что на одного работающего приходится более 40м^3 объема воздуха в помещении. Поскольку в помещении не выполняется требование к объему воздуха на

одного работающего (объём на одного человека — 28,88м^3), то наличие принудительной вентиляции просто необходимо.

В зимнее время в помещении необходимо предусмотреть систему отопления. Она должна обеспечивать достаточное, постоянное и равномерное нагревание воздуха. В помещениях с повышенными требованиями к чистоте воздуха должно использоваться водяное отопление. В рассматриваемой аудитории используется водяное отопление со встроенными нагревательными элементами и стояками.

5. Недостаточная освещённость рабочей зоны

Недостаточное освещение влияет на функционирование зрительного аппарата, то есть определяет зрительную работоспособность, на психику человека, его эмоциональное состояние, вызывает усталость центральной нервной системы, возникающей в результате прилагаемых усилий для опознания четких или сомнительных сигналов. [4]

Для оптимизации условий труда имеет большое значение освещение рабочих мест. Задачи организации освещённости рабочих мест следующие: обеспечение различаемости рассматриваемых предметов, уменьшение напряжения и утомляемости органов зрения. Производственное освещение должно быть равномерным и устойчивым, иметь правильное направление светового потока, исключать слепящее действие света и образование резких теней.

Среди качественных показателей световой среды очень важным является коэффициент пульсации освещенности (Кп). Требования к коэффициенту пульсации освещенности наиболее жесткие для рабочих мест с ПЭВМ — не более 5%. [4] Оптимальная яркость экрана дисплея составляет 75–100 кд/м2. При такой яркости экрана и яркости поверхности стола в пределах 100–150 кд/м2 обеспечивается продуктивность работы зрительного аппарата на уровне 80–90 %, сохраняется постоянство размера зрачка на допустимом уровне 3–4 мм.

Таблица 19 Нормируемые показатели естественного, искусственного и совмещенного освещения в соответствии с СанПиН 2.2.1/2.1.1.1278-03

Помещения	Рабочая поверхность и плоскость нормирования КЕО и освещенности и высота плоскости над полом, м	Естественное освещение		Совмещенное освещение		Искусственное освещение				
		КЕО е.н, %		КЕО е.н, %		Освещенность, лк				
		При верхнем или комбинированном освещении	При боковом освещении	При верхнем или комбинированном освещении	При боковом освещении	При комбинированном освещении		При общем освещении	Показатель дискомфорта, М, не более	Коэффициент пульсации освещенности, К_п, %, не более
						всего	от общего			
1	2	3	4	5	6	7	8	9	10	11
Помещения для работы с дисплеями и видеотерминалами, залы ЭВМ	Г – 0,8 Экран монитора : В – 1,2	3,5 -	1,2 -	2,1 -	0,7 -	500 -	300 -	400 200	15 -	10 -
Кабинеты, рабочие комнаты	Г – 0,8	3,0	1,0	1,8	0,6	400	200	300	40	15

Местное освещение не должно создавать бликов на поверхности экрана и увеличивать освещенность экрана ПЭВМ более 300 лк. Следует ограничивать прямую и отраженную блесткость от любых источников освещения.

В лаборатории, где проводятся исследования, используется смешанное освещение, т.е. сочетание естественного и искусственного освещения.

Естественным освещением является освещение через окна. Искусственное освещение используется при недостаточном естественном освещении. В данном помещении используется общее искусственное освещение.

Лаборатория освещается 3 светильниками, в каждом из которых установлено 4 люминесцентных лампы типа ЛБ-40. Светильники расположены равномерно по всей площади потолка в ряд, создавая при этом равномерное освещение рабочих мест. Световой поток каждой из ламп в помещении свидетельствует о соблюдении норм освещенности.

6. Повышенный уровень шума на рабочем месте

Одним из важнейших параметров, наносящим большой ущерб для здоровья и резко снижающим производительность труда, является шум.

Шум может создаваться работающим оборудованием, установками кондиционирования воздуха, преобразователями напряжения, работающими осветительными приборами дневного света, а также проникать извне.

В результате исследований установлено, что шум и вибрация ухудшают условия труда, оказывают вредное воздействие на организм человека. Действие шума различно: он затрудняет разборчивость речи, вызывает снижение работоспособности, повышает утомляемость, вызывает необратимые изменения в органах слуха человека. Шум воздействует не только на органы слуха, но и на весь организм человека через центральную нервную систему. Ослабляется внимание, ухудшается память, снижается реакция, увеличивается число ошибок при работе.

Производственные помещения, в которых для работы используются ПЭВМ, не должны граничить с помещениями, в которых уровень шума и вибрации превышают нормируемые значения. При выполнении основной работы на ПЭВМ уровень шума на рабочем месте не должен превышать 50 дБ. Допустимые уровни звукового давления в помещениях для персонала, осуществляющего эксплуатацию ПЭВМ при разных значениях частот, приведены в таблице 6.3. [5]

Таблица 20 Допустимые уровни звукового давления на рабочих местах расчетчиков, программистов вычислительных машин

Уровни звукового давления, дБ, в октавных полосах со среднегеометрическими частотами, Гц								Уровни звука и эквивалентные уровни звука, дБ А
63	125	250	500	1000	2000	4000	8000	
71	61	54	49	45	42	40	38	50

По субъективным ощущениям шумовая обстановка на рабочем месте соответствует норме.

7. Электробезопасность

Статическое электричество возникает в результате сложных процессов, связанных с перераспределением электронов и ионов при соприкосновении двух поверхностей неоднородных жидких или твердых веществ, на которых образуется двойной электрический слой. При механическом разделении поверхностей происходит разделение зарядов этого двойного электрического слоя. При этом между разделенными поверхностями, несущими электрический заряд, образуется разность потенциалов и возникает электрическое поле.

В помещении разрядные токи статического электричества чаще всего возникают при прикосновении пользователей к любому из элементов ЭВМ. Такие разряды опасности для человека не представляют, однако, кроме неприятных ощущений, они могут привести к выходу из строя ЭВМ.

Для снижения величин возникающих зарядов статического электричества в помещении покрытие полов выполнено из однослойного линолеума.

При работе с электроприборами очень важно соблюдать технику безопасности.

Под техникой безопасности понимается система организационных мероприятий и технических средств, направленная на предотвращения воздействия на работника вредных и опасных производственных факторов.

Электрические установки представляют для человека большую потенциальную опасность, которая усугубляется тем, что органы чувств человека не могут на расстоянии обнаружить наличие электрического напряжения на оборудовании.

В зависимости от условий в помещении опасность поражения человека электрическим током увеличивается или уменьшается. Не следует работать с компьютером в условиях повышенной влажности (относительная влажность

воздуха длительно превышает 75%), высокой температуры (более 35°С), наличии токопроводящей пыли, токопроводящих полов и возможности одновременного соприкосновения к имеющим соединение с землей металлическим элементам и металлическим корпусом электрооборудования. Таким образом, работа может проводиться только в помещениях без повышенной опасности, при этом существует опасность электропоражения:

1) при непосредственном прикосновении к токоведущим частям во время ремонта ПЭВМ;

2) при прикосновении к нетоковедущим частям, оказавшимся под напряжением (в случае нарушения изоляции токоведущих частей ПЭВМ);

3) при соприкосновении с полом, стенами, оказавшимися под напряжением;

4) имеется опасность короткого замыкания в высоковольтных блоках: блоке питания и блоке дисплейной развёртки [6].

Лаборатория, в которой проводились работы, по опасности электропоражения относится к помещениям без повышенной опасности, то есть отсутствуют условия, создающие повышенную опасность.

В помещении используются приборы, потребляющие напряжение 220В переменного тока с частотой 50Гц. Это напряжение опасно для жизни, поэтому обязательны следующие меры предосторожности:

1)перед началом работы нужно убедиться, что выключатели и розетка закреплены и не имеют оголённых токоведущих частей;

2)при обнаружении неисправности оборудования и приборов необходимо не делая никаких самостоятельных исправлений сообщить ответственному за оборудование;

3) запрещается загромождать рабочее место лишними предметами. При возникновении несчастного случая следует немедленно освободить пострадавшего от действия электрического тока и, вызвав врача, оказать ему необходимую помощь.

8. Экологическая безопасность

Вследствие развития научно-технического прогресса, постоянно увеличивается возможность воздействия на окружающую среду, создаются предпосылки для возникновения экологических кризисов. В то же время прогресс расширяет возможности устранения создаваемых человеком ухудшений природной среды.

Под окружающей нас средой понимается совокупность «чистой» природы и среды созданной человеком.

Защита окружающей среды - это комплексная проблема, требующая усилий всего человечества. Наиболее активной формой защиты окружающей среды от вредного воздействия выбросов промышленных предприятий является полный переход к безотходным и малоотходным технологиям и производствам. Это потребует решения целого комплекса сложных технологических, конструкторских и организационных задач, основанных на использовании новейших научно-технических достижений [7].

9. Загрязнение атмосферного воздуха

Во время проведения исследований выбросы вредных веществ в атмосферу не осуществляются. Загрязнение атмосферного воздуха может возникнуть в случае возникновения пожара в учебном корпусе, в этом случае дым и газы от пожара будут являться антропогенным загрязнением атмосферного воздуха.

10. Отходы

Основные виды загрязнения литосферы – твердые бытовые и промышленные отходы.

При проведении исследований, образовывались различные твердые отходы. К ним можно отнести: бумагу, батарейки, лампочки, использованные картриджи, отходы от продуктов питания и личной гигиены, отходы от канцелярских принадлежностей и т.д. [7]

Защита почвенного покрова и недр от твердых отходов реализуется за счет сбора, сортирования и утилизации отходов и их организованного захоронения.

11. Безопасность в чрезвычайных ситуациях

Мероприятия по пожарной профилактике разделяются на организационные, технические, эксплуатационные и режимные.

Организационные мероприятия предусматривают правильную эксплуатацию оборудования, правильное содержание зданий и территорий, противопожарный инструктаж рабочих и служащих, обучение производственного персонала правилам противопожарной безопасности, издание инструкций, плакатов, наличие плана эвакуации.

К техническим мероприятиям относятся: соблюдение противопожарных правил, норм при проектировании зданий, при устройстве электропроводов и оборудования, отопления, вентиляции, освещения, правильное размещение оборудования.

К режимным относятся установление правил организации работ и соблюдение противопожарных мер [6].

12. Оценка пожарной безопасности помещения

Согласно нормам технологического проектирования, в зависимости от характеристики используемых в производстве веществ и их количества, по пожарной и взрывной опасности помещения подразделяются на категории А, Б, В, Г, Д.

Наличие в лаборатории деревянных изделий (столы, шкафы), электропроводов напряжением 220В, а также применение электронагревательных приборов с открытыми нагревательными элементами – паяльниками дает право отнести помещение по степени пожаро и взрывобезопасности к категории В. Категория помещения «В»: помещения, в которых горючие и трудногорючие жидкости, твердые горючие и трудногорючие вещества и материалы (в том числе пыли и волокна), вещества и материалы, находящиеся в помещении, способны при

взаимодействии с водой, кислородом воздуха или друг с другом гореть, при условии, что помещения, в которых они имеются в наличии или обращаются, не относятся к категориям А или Б.

Необходимо предусмотреть ряд профилактических мероприятий технического, эксплуатационного, организационного плана.

В качестве возможных причин пожара можно указать следующие:

1) наличие горючей пыли (некоторые осевшие частицы пыли способны к самовозгоранию);

2) короткие замыкания;

3) опасная перегрузка сетей, которая ведет за собой сильный нагрев токоведущих частей и загорание изоляции;

4) нередко пожары происходят при пуске оборудования после ремонта [9].

Для предупреждения пожаров от коротких замыканий и перегрузок необходимы правильный выбор, монтаж и соблюдение установленного режима эксплуатации электрических сетей, дисплеев и других электрических средств автоматизации.

Следовательно, необходимо предусмотреть ряд профилактических мероприятий технического, эксплуатационного, организационного плана.

Причиной возгорания может быть:

1)неисправность токоведущих частей установок;

2)работа с открытой электроаппаратурой;

3)короткие замыкания в блоке питания или высоковольтном блоке дисплейной развертки;

4)несоблюдение правил пожарной безопасности;

5)наличие горючих компонентов: документы, двери, столы, изоляция кабелей и т.п.

Для предупреждения возникновения пожара необходимо соблюдать следующие правила пожарной безопасности:

1) исключение образования горючей среды (герметизация оборудования, контроль воздушной среды, рабочая и аварийная вентиляция);

2) применение при строительстве и отделке зданий несгораемых или трудно сгораемых материалов.

Необходимо в аудитории проводить следующие пожарно-профилактические мероприятия:

1) организационные мероприятия, касающиеся технического процесса с учетом пожарной безопасности объекта;

2) эксплуатационные мероприятия, рассматривающие эксплуатацию имеющегося оборудования;

3) технические и конструктивные, связанные с правильным размещением и монтажом электрооборудования и отопительных приборов.

Организационные мероприятия:

1) противопожарный инструктаж обслуживающего персонала;
2) обучение персонала правилам техники безопасности;
3) издание инструкций, плакатов, планов эвакуации.

Эксплуатационные мероприятия:

1) соблюдение эксплуатационных норм оборудования;
2) обеспечение свободного подхода к оборудованию;
3) содержание в исправности изоляции токоведущих проводников.

Технические мероприятия:

1) соблюдение противопожарных мероприятий при устройстве электропроводок, оборудования, систем отопления, вентиляции и освещения. В лаборатории имеется углекислотный огнетушитель типа ОУ–2, установлен рубильник, обесточивающий всю аудиторию, на двери аудитории приведен план эвакуации в случае пожара, и на досягаемом расстоянии находится пожарный щит (2 этаж НТБ). Если возгорание произошло в электроустановке, для его устранения должны использоваться углекислотные огнетушители типа ОУ–2.

2) профилактический осмотр, ремонт и испытание оборудования.

Кроме устранения самого очага пожара, нужно своевременно организовать эвакуацию людей [6].

Список используемых источников

1. H. Small Paradigms, citations, and maps of science: A personal history [Электронный ресурс]. – Режим доступа https://onlinelibrary.wiley.com/doi/full/10.1002/asi.10225,
2. Loet Leydesdorff, Stephen Carley, Ismael Rafols Global maps of science based on the new Web-of-Science categories [Электронный ресурс]. – Режим доступа https://link.springer.com/content/pdf/10.1007%2Fs11192-012-0784-8.pdf,
3. Handbook of Quantitative Science and Technology Research. Chapter 10. [Электронный ресурс]. – Режим доступа https://link.springer.com/content/pdf/10.1007%2F1-4020-2755-9.pdf
4. H. Small Interpreting maps of science using citation context sentiments: a preliminary investigation [Электронный ресурс]. – Режим доступа https://link.springer.com/content/pdf/10.1007%2Fs11192-011-0349-2.pdf
5. R. Klavans, K. W. Boyack Is there a Convergent Structure of Science? A Comparison of Maps using the ISI and Scopus Databases [Электронный ресурс].https://pdfs.semanticscholar.org/6c89/a7baef42f4f409f1a2c743530c980c40ec33.pdf,
6. Johan Bollen, Herbert Van de Sompel, Aric Hagberg, Luis Bettencourt, Ryan Chute, Marko A. Rodriguez, Lyudmila Balakireva Clickstream Data Yields High-Resolution Maps of Science [Электронный ресурс]. – Режим доступа http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004803
7. А.В. Соколов, С.А. Шашнов, М.Н. Коцемир, А.Ю. Гребенюк Определение приоритетовнаучно-технологического сотрудничества стран БРИКС [Электронный ресурс]. – Режим доступа https://iorj.hse.ru/data/2017/12/06/1161559079/А.В.%20Соколов,%20С.А.%20Шашнов,%20М.Н.%20Коцемир,%20А.Ю.%20Гребенюк.pdf
8. Николаев В.С., Оныкий Б.Н., Соколина К.А., Ушмаров И.А. Агентное сканирование мировых интернет-ресурсов по естественнонаучным

и технологическим направлениям // системы высокой доступности, том: 10, номер: 2, 2014, с. 50-53.

9. Артамонов А.А., Леонов Д.В., Оныкий Б.Н., Проничева Л.В Мультиагентная информационно-аналитическая система по естественно-научным и технологическим направлениям // системы высокой доступности, том: 10, номер: 2, 2014, с. 45-49.

10. Будзко В.И., Леонов Д.В., Николаев В.С., Оныкий Б.Н., Соколина К.А. Развитие информационно-аналитической поддержки научно-технической деятельности в национальном исследовательском ядерном университете "мифи" // системы высокой доступности, том: 7, номер: 4, 2011, с. 4-17.

11. Beyer, M. Gartner Says Solving “Big Data” Challenge Involves More Than Just Managing Volumes of Data, International Journal of Communications, Network and System Sciences,Vol.10 No.3, March 31, 2017

12. Emilio Ferraraa , Pasquale De Meob , Giacomo Fiumarac , Robert Baumgartner, Web Data Extraction, Applications and Techniques: A Survey // Knowledge-Based Systems, July 2014, C. 301-323

13. de S Sirisuriya , A Comparative Study on Web Scraping, Proceedings of 8th International Research Conference, KDU, Published November 2015, C.135-140

14. Abdelhakim Herrouz , Chabane Khentout , Mahieddine Djoudi , Overview of Web Content Mining Tools, The International Journal of Engineering And Science (IJES), Выпуск 6, June 2013, C. 106-110.

15. Березкин Д.В., Метод автоматизированного извлечения знаний из слабоструктурированных источников и его применение для создания корпоративных информационных систем, Труды 3 молодых ученых, аспирантов и студентов «Информатика и системы управления». Москва: Изд. МГТУ им. Н. Э. Баумана. 2007. С. 315-320.

16. Janssens F, Leta J, Glänzel W, De Moor B. Towards mapping library and information science. Inform Process Manag. 2006;42(6)

17. Boyack KW, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? J Am Soc Inf Sci]

18. Jarneving B. Bibliographic coupling and its application to research-front and other core documents. J Infometr. 2007;

19. Small H, Griffith BC. The structure of scientific literatures I: Identifying and graphing specialties. Sci Stud. 1974;

20. 7. Janssens F, Glänzel W, De Moor B. A hybrid mapping of information science. Scientometrics. 2008;75(3):607–631. 10.1007/s11192-007-2002-7

21. 8. Small H. Update on science mapping: Creating large document spaces. Scientometrics. 1997;38(2);

22. 9. Waltman L, Van Eck NJ, Noyons ECM. A unified approach to mapping and clustering of bibliometric networks. J Infometr. 2010;

23. 10. Waltman L, Van Eck NJ. A new methodology for constructing a publication-level classification system of science. J Am Soc Inf Sci Tec. 2012;

Список публикаций и научных достижений

- Очное участие в XV Международной научно-практической конференции студентов конференции МСИТ 04-07 декабря 2017 г.;
- Публикация “Разработка методики применения метода главных компонент к кардиосигналу” О. Н. Вылегжанин, О. О. Демидова ; Молодежь и современные информационные технологии : сборник трудов XV Международной научно-практической конференции студентов, аспирантов и молодых учёных, 04-07 декабря 2017 г., г. Томск. — Томск : Изд-во ТПУ, 2017. — [С. 37-38].;
- Публикация “Сравнительный анализ техник извлечения данных из веб-страниц при решении задачи кластеризации научных публикаций” А.О. Савельев, О.О. Демидова, сборник трудов XIV Международной научно-практической конференции «Электронные средства и системы управления», 28-30 ноября 2018 г., г.Томск. — Томск : Изд-во ТУСУР, 2018. — [С. 116-120]
- Участие в летней школе “Intelligent Vehicles and Electric Vehicles“, Китай г. Пекин 2018г.;
- "Вторая международная школа-конференция по гетерогенным вычислительным инфраструктурам", в рамках симпозиума NEC' 2017, г. Будва, Черногория.
- Участие в академическом обмене с Техническим университетом г. Мюнхен (Германия) с 01.10.2018 г. по 31.03.2019 г.;
- Стипендия ПЛЮС для академического обмена;
- Стипендия Правительства РФ (1 семестр 2018-2019 учебного года).

ПРИЛОЖЕНИЕ А

Раздел 1
Обзор МЕТОДОВ WEB-MINING
WEB-MINING methods

Студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демидова Оксана Олеговна		

Консультант ОИТ ИШИТР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Соколова Вероника Валерьевна	к.т.н.		

Консультант – лингвист ОИЯ ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ ШБИП	Диденко Анастасия Владимировна	к.ф.н		

Introduction

The characteristic feature of modernity is an increasing amount of information. The scientific community has a huge amount of material. On the one hand, propagation and increasing influence of citation bases promote spreading scientific knowledge and improving the quality of research results. On the other hand, huge amount of scientific information that is presented in the public domain complicates the analysis, clarification and adjustment of scientific and technical priorities at the state level. The continuous increase of scientific information that is presented in the electronic form makes the data searching process more complicated, time consuming and inefficient technological process.

A user has to refine search criteria every time, according to which they often receive a familiar list of documents. The search process loops, and the search time increases significantly.

The purpose of this work is developing the technology for collecting, analyzing and clustering data based on the open Internet sources. The object of research is the database of citations. The subject of the research is content extraction systems from open Internet sources and the methods of clustering the obtained data.

The main tasks of the project are development of the data scheme of the standard for the collection, storage and provision of aggregated information on scientific publications.

Overview of the subject area

Solving these tasks requires an approach based on an automated analysis of the available pool of open publications. Extracting structured and related data from web pages is reduced to the sequential solution of the five tasks:

- search and retrieving of targeted pages for data retrieval;
- recognition of areas containing the necessary data;
- search of the structure of the found data;
- providing the uniformity of the extracted data;
- combining data from different sources.

Web-mining is the new direction in methodology of data analysis. It is intended for effectively solving the problems of searching, structuring and analyzing chaotically organized information in the network.

1 Web-mining

Web mining develops at the intersection of such disciplines as the discovery of knowledge in databases, effective information retrieval, artificial intelligence, machine learning and processing of natural languages.

The tasks of web-mining are:

- Searching information;
- Identification of knowledge from web resources;
- Personalization of information;
- Search for patterns of user behavior.

Users often use searching recourse for searching necessary information through simple requests by key words. Results of searching are lists of pages that are sorted by some index of relevance that describes a level of matching with a result of request. However, existing search engines have disadvantages. The main one is the low accuracy of the result, caused by the insufficient consideration of the semantic links and the context of the expressions found in the text. Indexing network segments using data mining that have mathematical linguistic and natural

language processing algorithms are the promising direction for web mining in the field of information retrieval.

The method of searching information consisted in analyzing the structure of links between different web pages, internal and external sites in a selected network segment. The appearance of this method was caused by the need to solve problems that arose from the social networks analysis or specific areas of human activity or knowledge, for example, in the analysis of author's citation.

The task of identification of knowledge from web resources intersects with the information search problem that already described. But a researcher already has a set of web pages that are obtained as a result of the request. Furthermore it is required to do their processing from the point of view of automatic classification, drawing up tables of contents, revealing keywords and general topics. The received knowledge can be presented in the form of trees, describing the structure of documents or in the form of logical and semantic expressions. Solving some of these problems is offered by Text Mining, a technology for automatically extracting knowledge in large volumes of textual material, based on a combination of linguistic, semantic, statistical, and machine learning techniques. Personalization of web space is the task of creating web systems that adapt their capabilities (navigation, content, banners and other promotional offers) for the user based on the collected and analyzed information about user preferences.

For analyzing information about a user, the declared information about a someone should be used to the least extent, but rather be based on persistent patterns of them "behavior" in the network - a sequence of clicks within a resource, switching to other sub-resources, periods of network activity, purchases and etc.

The task of searching for patterns of user behavior is related to the previous one, but its goal is not adapting a resource to the preferences of individual users, but looking for patterns in patterns of user interaction with a web resource in order to predict them subsequent actions. Analyzed user actions may include not only referrals, but also sending the forms, scrolling pages, favorite pages, etc. Found

templates are used in the future to optimize the structure of the site, research the target audience and for direct marketing.

2 Web mining categories

Web mining can be generally divided into three categories like Web content mining, Web structure mining and Web usage mining.

Web content mining is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. Techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.

Web structure mining is the process of analyzing the nodes and a connection structure of a website through using of graph theory. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected.

Web usage mining is the process of extracting patterns and information from server logs to gain insight into user activity including where the users are from, how many clicked what item on the site and the types of activities being done on the site.

Figure 1 shows the general relationship between Web Mining categories and data mining tasks.

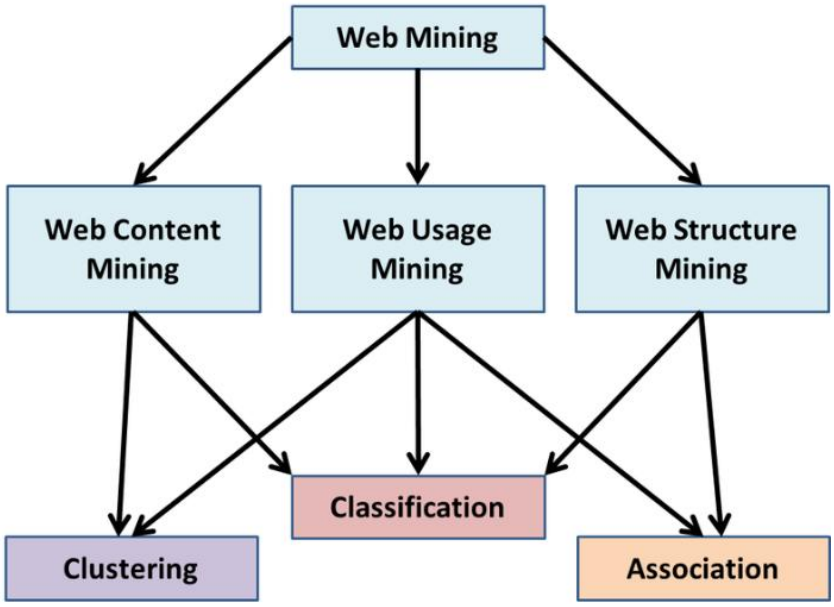


Figure 12 Three categories of Web mining

In this case, our task is to extract web content (Web content mining).

3 Analysis of web-mining approaches

There are some approaches of extracting data from web-sources as parsing strings, DOM analysis (document object model), Regular Expressions and XML parsing.

We can use the approach of parsing strings if data are displayed using, for example, a table of characteristics, when the values of the parameters are standard. The method is not suitable for writing serious parsers.

Using the DOM analysis approach, data can be obtained directly from the identifier, name, or other attributes of a tree element (such an element can be a paragraph, table, block, etc.). In addition, if an element is not identified by any identifier, then it can be reached via a unique path going down the DOM tree.

The approach of regular expressions is used only for extracting data that have a strict format (email addresses, phone numbers, etc.).

We should not consider HTML as XML data. It is not always the suitable option, since HTML data are rarely valid, i.e. so that it can be viewed as XML data. Libraries that implemented this approach have devoted more time to converting HTML to XML and only then directly to data parsing.

4 Analysis of web-mining tools

Web scraping, web crawling, and any other form of web data extraction can be complicated. Between obtaining the correct page source, to parsing the source correctly and obtaining data in a usable form, there is a lot of work to be done. Different users have very different needs, and there are tools out there for all of them like libraries, headless browsers and SaaS solutions.

The approach of libraries requires understanding of the query generation process and application logic. It is suitable for writing a parser to a specific site. Such tools include numerous libraries for various programming languages such as JAVA, Python, PHP.

The approach of headless browsers allows processing the page in a browser with JavaScript support. This approach allows writing scripts to get the required information and even use JavaScript libraries like jQuery to extract information from the page. These tools include PhantomJS and SlimerJS.

SaaS solutions provide a graphical interface where person can specify the page address, specify the blocks where person want to extract information, and create a number of rules for extracting data. Such services do not have the flexibility in deference with low-level solutions.

Conclusion

We have reviewed the basic concepts and approaches of the web mining. In this paper, we explore data in the category of the Web content mining, since the necessary information that we need to extract and analyze, is in the form of unstructured text data.

For extracting data from web page we choose the DOM tree analysis method, since many elements do not have a specific identifier. Also, we decide to use the

tool for extracting data from web pages with using programming language libraries.

ПРИЛОЖЕНИЕ В

Примеры некоторых функций для извлечения данных (рис.11,12,13)

```
def set_articles(FILENAME, TOPIC_NAME, num):
        for i,row in enumerate(reader):
            if i>0 and i<num:
                        el_massiv = []

                        id.append({'id': int(row[0])})
                        field.append({'field': 2})

                        name_article.append({'name_article':row[2]})

                        if row[6] != 'None':
                            citation = row[6].split(',')
                            el_massiv = []

                            if len(citation)>1: # ищем id процитированных статей

                                for n,elem in enumerate(citation):
                                    if n == 0:
                                        el = elem[2:-1]
                                        el_massiv.append(el)

                                    elif n == (len(citation)-1):
                                        el = elem[2:-2]
                                        el_massiv.append(el)

                                    else:
                                        el = elem[2:-1]
                                        el_massiv.append(el)
                                cit_el.append({'link_cictation' : el_massiv })

                            else:
                                el = row[6][2:-2]
                                el_massiv.append(el)
                                cit_el.append({'link_cictation' : el_massiv })

                        else:
                            cit_el.append({'link_cictation': 0 })

                       # print(cit_el)
```

Рисунок 13

```
def match(cit_el, i):
    print(i)
    for j, elem_id in enumerate(massiv_id):
        print(j)
        if cit_el == elem_id:

            row.append(i)
            col.append(j)

            row.append(j)
            col.append(i)

            massiv_cit[i][j] = 1
            massiv_cit[j][i] = 1
```

Рисунок 14

```
massiv_cit = df['link_cictation']

for i,elem in enumerate(massiv_id):
    id_cit = elem

    if massiv_cit[i] != '0':
        citation = massiv_cit[i].split(',')

        if len(citation)>1:
              for n,elem in enumerate(citation):
                        if n == 0:
                            cit_el = elem[2:-1]

                        elif n == (len(citation)-1):
                            cit_el = elem[2:-2]

                        else:
                            cit_el = elem[2:-1]

        else:
            cit_el = massiv_cit[i][6][2:-2]

        match(cit_el, i)
```

Рисунок 15