МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное

учреждение высшего образования

«Московский авиационный институт

(национальный исследовательский университет)» (МАИ)

УТВЕРЖДАЮ

Заведующий кафедрой

«Теория вероятностей и компьютерное моделирование»

д.ф.-м.н., профессор

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.И. Кибзун

«04» января 2022 г.

ОТЧЕТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:

Сформулированная на данный момент тема ВКР

(заключительный)

|  |  |  |
| --- | --- | --- |
| Научный руководитель  к.ф.-м.н., доцент |  | фио |
| Исполнитель  магистр группы 8О- |  | фио |

Москва 2022

# РЕФЕРАТ

Отчет 15 с., 7 рис., 1 табл., 5 источн., 1 прил.

Объектом исследования являются ………….

Цель работы – ………………….

В результате работы разработана …………….. Дальнейшее исследование может включать в себя исследование …………………..

Оформляем НИР в соответствии с ГОСТ 7.32-2017

**СОДЕРЖАНИЕ**

[РЕФЕРАТ 2](#_Toc86178564)

[ВВЕДЕНИЕ 4](#_Toc86178565)

[Основная часть отчета о НИР 5](#_Toc86178566)

[ЗАКЛЮЧЕНИЕ 11](#_Toc86178567)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 12](#_Toc86178568)

[ПРИЛОЖЕНИЕ 1 13](#_Toc86178570)

# ВВЕДЕНИЕ

В данной научно-исследовательской работе проводится исследование возможных подходов к решению задачи объединения поисковых запросов в зависимости от принадлежности к определенной тематике.

Появление данной задачи – прямое следствие желания понимать пользователя любой поисковой системы, его мотивы, выявлять закономерности в его действиях, стремление улучшать алгоритмы поиска и ранжирования, алгоритмы рекомендательных систем. В истории запросов пользователя содержится бесценная информация как о нем, так и о качестве системы, которой он пользуется. Но еще больше информации содержится в совокупности запросов, объединенных одной тематикой.

Запросом в задаче является последовательность символов, которую пользователь вводит в поисковую строку, чтобы найти интересующую его информацию. Запрос характеризуется также рядом атрибутов, таких как, например, время, потраченное на запрос, множество документов, открытых пользователем, различные метки запроса и многое другое.

Множество последовательных запросов, совершенных пользователем в течение времени с момента активации поисковой системы до прерывания взаимодействия с ней, будем называть сессией.

Под тематикой в задаче будет пониматься информационная потребность пользователя.

Результаты данной работы будут использованы в компьютерной справочной правовой системе. Потенциальные приложения результатов зависят от степени экстраполяции предложенного решения: его применение в рамках конкретной сессии или его применение на множестве всех сессий.

При применении построенных алгоритмов и моделей в рамках одной сессии можно выделить следующие перспективные задачи, для решения которых могут быть использованы результаты группировки запросов по потребностям:

* Аналитика статистики поисковых запросов, определенных в качестве потребности
* Модификация рекомендательных алгоритмов для тематически смежных запросов путем добавления дополнительной информации о текущей потребности пользователя
* Формализация определения кликов вне выдачи с учетом группировки запросов по потребностям

В случае экстраполяции решения на множество всех сессий могут возникать следующие задачи:

* Расширение кластеров запросов с учетом информации о принадлежности запросов различным потребностям
* Формализация и сегментация потребностей
* Выявление потребностей, имеющих недостаточную представленность документами
* Статистика по потребностям, выявление проблемных сегментов потребностей

При решении данной задачи основной акцент будет сделан на построение моделей, основанных на извлечении и обработке лингвистических признаков. Имеющиеся атрибуты запросов могут быть использованы при построении базовых моделей, а также при оценке качества работы алгоритмов.

Таким образом, в данной постановке задачи требуется с использованием методов машинного обучения объединить запросы пользователей в рамках одной или нескольких сессий в группы потребностей.

# Основная часть отчета о НИР

В данной работе производится первичный анализ области исследования, определяются возможные подходы к решению задачи и основные этапы работы, приводятся примеры разработок в сходных областях и их возможные модификации в терминах текущей задачи, проводится анализ предлагаемых методов решения.

**Анализ области исследования**

Исходя из задач, при решении которых могут быть использованы результаты данной работы, могут принципиально отличаться как алгоритмы решения, так и подходы к нему в целом. Решения могут быть эвристическими, могут включать в себя построение более сложных алгоритмов, в том числе с использованием моделей машинного обучения. От выбора подхода к решению зависят существование базовых решений, набор используемых атрибутов запроса, определение методов извлечения эвристик и построения правил, способы оценки параметров алгоритма, необходимость в наличии разметки данных, методы оценки качества и многое другое.

Таким образом, при выборе той или иной области применения результатов возникают некоторые ограничения и допущения при решении задачи.

Предварительный анализ уровней экстраполяции решения представлен в таблице 1.

**Таблица 1 – Анализ степени экстраполяции**

|  |  |  |
| --- | --- | --- |
| Критерий | Уровень экстраполяции | |
| В рамках одной сессии | На всем множестве сессий |
| Выделяемые потребности | Потребности сильно привязаны к конкретным пользователю или сессии; детализация потребностей | Обобщенные для всех пользователей потребности; проблемы с градациями крупных потребностей |
| Решаемые задачи | Узкие точечные задачи | Глобальные задачи |
| Наличие базового решения | Есть | Нет |
| Разметка данных | Может быть получена: отсутствие необходимости размечать данные вне одной сессии; необходимость учета элемента субъективности экспертов разметки  Возможность обучения с учителем | Затруднительна или невозможна: сложно выбрать репрезентативное количество данных; большие человеческие затраты; неконтролируемый рост погрешности разметки при увеличении количества данных; риски упущения важных примеров или повторения паттернов, не требующих усиленного внимания  Необходимость обучения без учителя |
| Вычислительные затраты | Минимальные вычислительные затраты при расчетах результатов;  Необходимость многократного повторения вычислений для каждой сессии | Большие вычислительные затраты при расчетах результатов;  Относительно редкая необходимость пересчета результатов |
| Оценка качества | Ручная проверка незатруднительна; существует возможность проверки качества на размеченных данных | Ручная проверка невозможна из-за отсутствия ограничений множества запросов; возможна проверка качества на размеченных на нескольких сессиях данных, требует оценки репрезентативности выборки |
| Обобщение результатов для решения задач в рамках противоположной экстраполяции | Возможность настройки дополнительных правил объединения потребностей на сессиях в кластеры более общих потребностей, сложность интерпретации результатов обобщения | Адекватная модель должна достаточно хорошо работать и в рамках одной сессии, выделяемые потребности могут быть более общими |
| Дополнительно | Потребность не будет корректно обработана, если она решалась в рамках нескольких сессий;  Необходимость проведения расчетов на всем множестве сессий и запросов при использовании вероятностных показателей и моделей | - |

Таким образом, при экстраполяции в рамках одной сессии решение может опираться на результаты базовой модели, становится доступным использование моделей обучения по аналогии, а также упрощается проверка результатов и оценка качества моделей. На множестве всех сессий необходимо вводить дополнительные метрики качества, не основанные на разметке данных, превалирует использование эвристических алгоритмов и моделей кластеризации, результирующие решения возможно также распространить на одну сессию.

**Обработка данных**

При автоматизированном решении любой задачи необходима грамотная обработка данных: форматы данных и извлекаемые из них признаки должны быть формализованы каждый раз при решении задачи.

В данной работе могут быть использованы две основных сущности для выделения значимых признаков:

1. Текст запроса
2. Атрибуты запроса

В работе на тему построения графа запросов [1] предлагается использование в качестве признаков выделенные из атрибутов запросов показатели среднего времени по всем сессиям, сумма обратных величин разницы времени между двумя запросами, количество сессий, в которых пара запросов появлялась вместе, средняя длина сессий, среднее количество кликов, средние позиции запросов в сессии. Также в [1] упоминается возможность использования информации о документах поисковой выдачи, в особенности об открытых пользователем документах.

Авторы [2] предлагают использование показателя разницы времени между последовательными запросами в совокупности с лингвистическими и семантическими признаками, выделенными из теста запросов, а также мера схожести открытых веб-страниц по запросам. В представленном в [2] алгоритме используются такие меры близости текстов запросов как индекс Жаккара, косинусовая мера, расстояние перестановки слов запросов.

В [3] при решении задачи сегментации слов запросов использовались только векторные представления слов word2vec, Glove и fasttext. Стоит отметить, что при решении поставленной в [3] задачи векторные представления обусловлены на слова, в то время как при решении задачи группировки запросов по потребностям необходимо обуславливаться на весь запрос в целом. Это возможно несколькими способами:

* Усреднение вектора по всем словам запроса
* Формирование вектора путем применения механизма внимания [4]
* Использование специальных фреймворков (sent2vec, universal-sentence-encoder)

Таким образом, можно выделить достаточно большое количество признаков схожести и различия запросов.

При обработке данных важно учитывать вероятное наличие ошибок в запросах, опечаток, неаккуратность при замене одного запроса другим и другие нюансы реальных данных. Также необходимо удалить все стоп-слова, а токены привести к определенной унифицированной форме, в качестве которых могут выступать стемминг, лематизация, n-граммы, разбиение по морфемам и многое другое.

Дополнительной задачей при обработке данных в исследовательской работе выступит специфика поисковой системы. Документы системы носят правовой характер, содержат множество юридических терминов, обладают рядом стилистических особенностей изложения материала. Также стоит отметить, что в правовой области существует достаточное количество сленга, который может требовать отдельной обработки.

**Разметка запросов**

Поставленная задача может решаться несколькими способами, такими как кластеризация запросов, классификация на предмет принадлежности к одной потребности, построение языковой модели, способной дать интерпретацию запросам в терминах потребности. В случае применения классификации строго необходима разметка данных, в то время как при построении языковой модели наличие разметки не всегда обязательно. Например, в [5] при решении задачи сегментации запросов на сложные семантически обусловленные термы авторы предлагают использовать модель, обученную на парах одинаковых запросов.

В работе [6] при решении сходной с [5] задачи строится языковая модель, однако интересны описанные в статье подходы к разметке данных. Достаточно серьезной проблемой при разметке текстовых данных является наличие для одних и тех же примеров различных интерпретаций, в ряде случаев, противоречащих друг другу. Интерпретации зависят от конкретного асессора – человека, размечающего данные. По этой причине в случае большого разброса интерпретаций построенная модель может давать достаточно неточные оценки и быть несостоятельной.

Авторы [6] предлагают проводить разметку данных несколькими независимыми асессорами, а итоговую разметку получать методом голосования оценок по каждой позиции. Первый подход к разметке, описанный в [6], подразумевает добавление в коллекцию токенов специального символа, обозначающего место разделения. Например, таким символом может выступать “|”. В терминах задачи группировки запросов по потребностям при разметке также могут понадобиться дополнительные специальные токены, предназначенные для разделения запросов между собой или запросов и дополнительных текстов.

Далее необходимо провести голосование простым большинством по поводу каждой новой спорной позиции, где может находиться токен “|”. В случае, если большинство асессоров считают, что на рассматриваемой позиции необходим специальный токен, то в итоговой разметке данный токен остается на соответствующей позиции.

Другой способ разметки, описанный в [6], - тэги BIO. Данный способ разметки используется в задаче распознавания именованных сущностей, где символ B обозначает начальный токен именованной сущности, I – внутренний, а O – за пределами именованной сущности. В интерпретации для задачи выделения запросов одной потребности достаточно ставить в соответствие каждому запросу тэги B или I, что будет обозначать соответственно начало и продолжение цепочки запросов одной потребности.

Очевидным недостатком методов разметки, предложенных в [6], является невозможность группировать непоследовательные запросы, принадлежащие одной потребности. Данная проблема может быть решена бинарной классификацией пар запросов [3], где каждой паре запросов ставится метка принадлежности одной потребности. Такой подход к разметке позволяет решать задачу не только с помощью построения языковой модели с классификатором, но и более простыми методами машинного обучения для задачи классификации с использованием не только информации о лексике и семантике запросов.

Таким образом, в таблице 2 приведена краткая характеристика рассмотренных подходов к разметке данных.

**Таблица 2 – Подходы к разметке данных**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Разметка данных | | |
| Специальные разделяющие токены | BIO-tagging | Бинарная разметка |
| Достоинства | Простота интерпретации;  Простота при разметке;  Возможность приведения нескольких разметок к эталонной;  Отсутствие противоречий после приведения к эталонной разметке; | Простота при разметке;  Возможность обозначения выбросов тэгом O;  Возможность приведения нескольких разметок к эталонной;  Отсутствие противоречий после приведения к эталонной разметке; | Простота интерпретации;  Возможность разметки непоследовательных запросов;  Возможность использования различных методов машинного обучения;  Допускается принадлежность одного запроса разным потребностям; |
| Недостатки | Работа только с цепочками последовательных запросов;  Необходимость использования языковых моделей;  Отсутствие возможности принадлежности одного запроса разным потребностям; | Работа только с цепочками последовательных запросов;  Необходимость использования языковых моделей;  Отсутствие возможности принадлежности одного запроса разным потребностям; | Сложность при разметке длинных цепочек запросов;  Высокая вероятность возникновения противоречий;  Не рекомендуется приведение к эталонной разметке по причине высоких рисков возникновения противоречий; |

**Подходы к решению**

Рассматриваемая в данной работе задача чаще всего решается на основе атрибутов запросов с использованием вероятностных оценок и эвристических правил [1, 2], таких как подбор порогов степени схожести, описание сценариев обработки различных данных и многое другое. Использование лингвистики в таких решениях сводится к минимуму и ограничивается оценками степени схожести пар запросов, а использование языковых моделей (статистическая модель машинного перевода [5], рекуррентная нейронная сеть Encoder-Decoder [6]) происходит в смежных задачах и не редко требует разметки, в том числе для возможности классификации векторных представлений запросов [3].

Данные подходы к решению могут быть адаптированы под задачу данной исследовательской работы, однако приоритетными направлениями при решении выступают лингвистически и семантически обусловленные модели, а также методы обучения без учителя (кластеризация).

Одним из эмпирических подходов к решению может выступать модификация метода, описанного в [2]. Авторы данного алгоритма предлагают в том числе использовать информацию о степени URL открытых по запросам сайтов. В интерпретации на решаемую в данной исследовательской работе задачу можно использовать подобную информацию об открываемых документах. Для расчета показателей степени схожести могут выступать:

* Названия документов
* Сниппеты – краткие текстовые изложения сути документа
* Доминантные интерпретации документов – выделенные экспертами интерпретации цели поиска по совокупности запроса и документа

Также одной из возможных модификаций может быть использование информации о схожести выделенных по текстам запросов аспекты [7] – объекты запроса, главные предметы интереса пользователя. Главной проблемой при выделении аспектов может стать необходимость точного задания их количества, а, следовательно, дополнительных затрат на подбор данного параметра.

В работе [8] на тему выявления намерений пользователя при задании запросов рассматриваются такие подходы к обучению без учителя, как PLSA, k-Means, самоорганизующаяся карта Кохонена, CFS, дерево Чау-Лю, метод ассоциативных правил. Некоторые модели требуют точного задания параметров количества кластеров, что будет проблематично в связи с невозможностью оценить количество потребностей пользователей, однако также существуют некоторые модификации подобных моделей, позволяющие решать задачу подбора параметров автоматически на основе обучающей выборки.

Отдельно в [8] на рассмотрение выносится графическая модель с древовидной структурой, использующая только информацию из текста запроса и идентифицирующая новые слова с помощью WordNet. Стоит оговориться, что WordNet предназначена только для английского языка, однако для русского языка построена похожая база данных.

Также в этой [8] работе приведен обзор решение задачи в терминах классификации.

Отдельного внимания заслуживает метод кластеризации DBSCAN, так как он позволяет идентифицировать выбросы, которые, в силу человеческого фактора, имеют место быть в реальных данных.

Также задача данной работы может решаться в терминах тематического моделирования – разделе задач машинного обучения, направленных на изучение способов построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. При решении задачи выделения по запросам потребностей пользователей можно использовать наработки в области тематического моделирования как для выделения тематик на основе только текстов запросов, так и для открываемых пользователем документов, которым в соответствие будут ставиться конкретные запросы пользователей. Второй возможный подход может позволить решить данную задачу на множестве всех сессий с дальнейшим выделение подкластеров запросов внутри каждой определенной тематики.

В области тематического моделирования наиболее известна модель LDA, для которой существуют модификации (Hierarchical Dirichlet process), позволяющие автоматизировать подбор количества тематик.

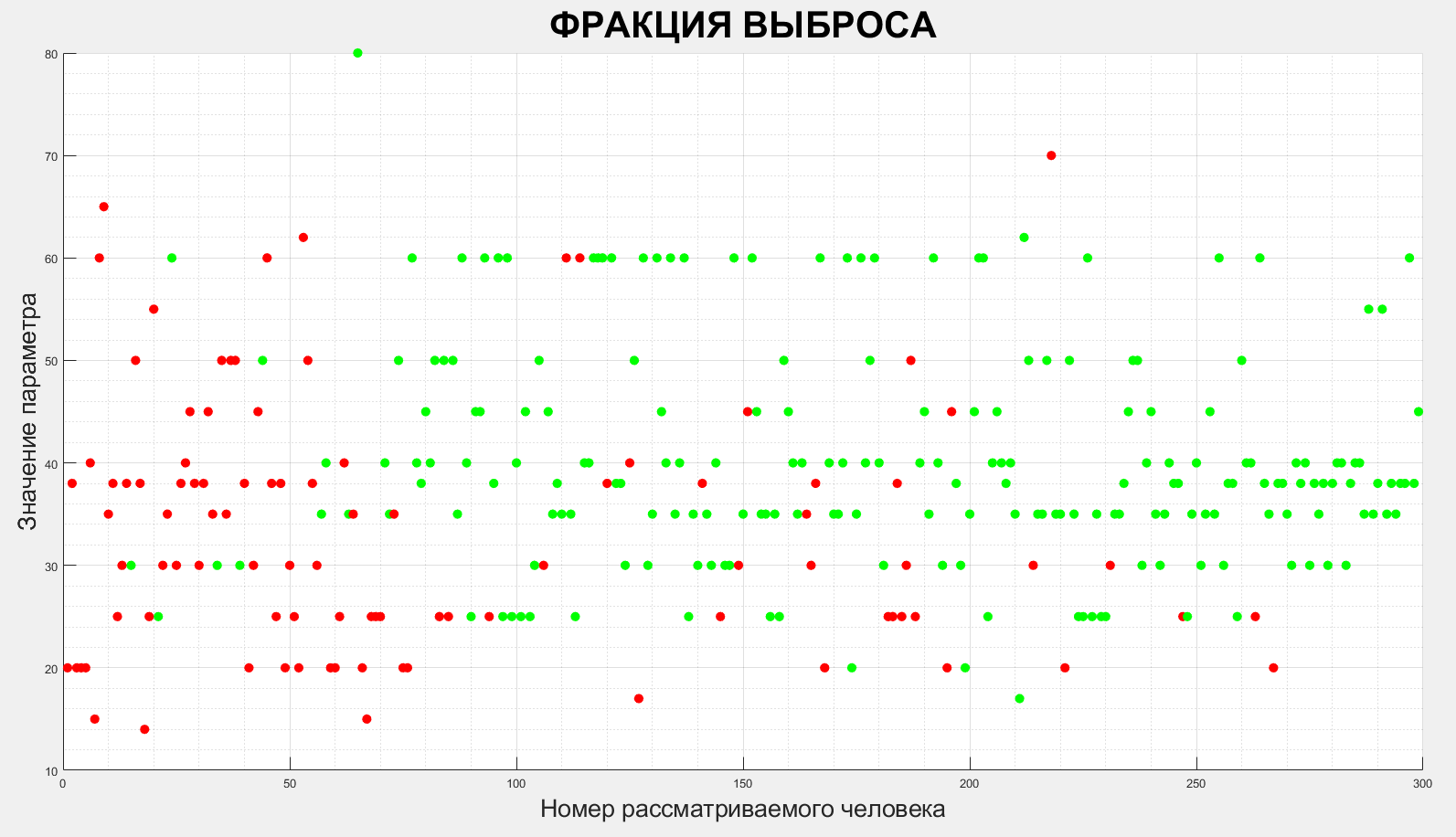
В области языкового моделирования существует достаточно большое количество различных фреймворков и предобученных моделей, которые могут быть применены для решения поставленной задачи. Фреймворки sematch и semantic-text-similarity предоставляют функционал для работы с задачами семантической схожести текстов. В них можно найти предобученные модели, определяющие степень семантической схожести текстов. Таким образом, данные фреймворки могут выступать в качестве базовых моделей при объединении запросов в группы на основе степени семантической схожести. Также существуют различные модификации предобученной модели BERT [9], которая на этапе своего обучения в том числе училась классифицировать пары предложений на предмет их последовательности. Однако при использовании таких моделей и техник существует проблема недообученности моделей под спецификацию правовых текстов. В этом случае необходимо доучивать модели на размеченных корпусах текстов, так как модели необходимо понимать отношения между предложениями специфической области.

Таким образом, предобученные модели могут выступать в качестве базовых решений, а их дальнейшая модификация возможна только в условиях наличия разметки данных.

Еще одним возможным подходом к решению задачи с помощью языковых моделей может быть обучение собственных моделей, заточенных исключительно под особенности правовой лексики

**Оценка качества**

…



*Рисунок 1- Сопоставление результатов анализа «Фракция выброса» с наступлением летального исхода*

# ЗАКЛЮЧЕНИЕ

Таким образом, в ходе данной работы было произведено исследование …

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Димитрия Князь. Анализ основных алгоритмов кластеризации многомерных данных. – М.: LAP Lambert Academic Publishing, 2014. – 64 с.

The Query-flow Graph: Model and Applications

Segmenting User Sessions in Search Engine Query Logs Leveraging Word Embeddings

Towards Semantic Query Segmentation

Attention Is All You Need

Unsupervised Query Segmentation Using Monolingual Word Alignment Method

Query Segmentation via RNNs Encoder-Decoder Framework

An Unsupervised Neural Attention Model for Aspect Extraction

Unsupervised Identification of the User’s Query Intent in Web Search

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Изображение выглядит как текст

Автоматически созданное описание

Примеры на стр 23-24 Госта

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата Подпись

# ПРИЛОЖЕНИЕ

Код программы, реализующей алгоритм «случайный лес»:

import numpy as np #linear algebra