Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

РАЗРАБОТКА АЛГОРИТМОВ РАБОТЫ С ФОРМАЛЬНОЙ МОДЕЛЬЮ ДИАЛОГОВ, ПРЕДСТАВЛЕННЫХ В ВИДЕ ГРАФОВ

Автор: Савон Юлия Константиновна _____

Направление подготовки: 01.03.02 Прикладная математика и информатика

Квалификация: Бакалавр

Руководитель ВКР: Ульянцев В.И., доцент, к.т.н. _____

Санкт-Петербург, 2020 г.

Обучающийся Савон Юлия Константиновна
Группа М3437 Факультет ИТиП

Направленность (профиль), специализация
Математические модели и алгоритмы в разработке программного обеспечения

Консультанты:

а) Ступаков И.М., канд. тех. наук, доцент _____

ВКР принята «____» _____ 20___ г.

Оригинальность ВКР ____%

ВКР выполнена с оценкой _____

Дата защиты «25» июня 2020 г.

Секретарь ГЭК Павлова О.Н. _____

Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

УТВЕРЖДАЮ

Руководитель ОП

проф., д.т.н. Парфенов В.Г. _____

«____» _____ 20__ г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Обучающийся Савон Юлия Константиновна

Группа М3437 **Факультет** ИТиП

Квалификация: Бакалавр

Направление подготовки (специальность): 01.03.02 Прикладная математика и информатика

Направленность (профиль): Математические модели и алгоритмы в разработке программного обеспечения

Тема ВКР: Разработка алгоритмов работы с формальной моделью диалогов, представленных в виде графов

Руководитель Ульянцев В.И., доцент, к.т.н., доцент факультета информационных технологий и программирования Университета ИТМО

2 Срок сдачи студентом законченной работы до: «10» июня 2020 г.

3 Техническое задание и исходные данные к работе

Требуется провести исследование и разработать набор алгоритмов для выявления отвлечений в графовой модели для телефонной диалоговой системы. Алгоритм принимает набор диалогов, в размере нескольких тысяч. Предварительно выстроенный граф и кластеризацию для фраз оператора.

На выходе ожидается получить набор отвлечений и перестроенный граф. В качестве метрики качества будет использоваться сравнение с уже существующими графами, которые создавались вручную.

4 Содержание выпускной квалификационной работы (перечень подлежащих разработке вопросов)

Пояснительная записка должна описывать предметную область диалогов представленных в виде графов. Так же формулировать цель и задачу выделения отвлечений, содержать описание алгоритмов их поиска. Должны быть описаны сложности и методы их разрешения, если они возникали. Кроме того должны быть приведены примеры работы алгоритмов и сравнение с существующими решениями. Кроме того пояснительная записка должна содержать описания задач из смежных областей и их то, как эти задачи связны с задачей решаемой в работе.

5 Перечень графического материала (с указанием обязательного материала)

Графические материалы и чертежи работой не предусмотрены

6 Исходные материалы и пособия

а) Среда разработки Visial Studio Code;

б) ГОСТ 7.32–2001 «Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления».

7 Дата выдачи задания «01» сентября 2019 г.

Руководитель ВКР _____

Задание принял к исполнению _____«01» сентября 2019 г.

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Обучающийся: Савон Юлия Константиновна

Наименование темы ВКР: Разработка алгоритмов работы с формальной моделью диалогов, представленных в виде графов

Наименование организации, где выполнена ВКР: Университет ИТМО

ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

1 Цель исследования: Разработа алгоритма выделяющего отвлечения в диалоге, представленном в виде графа.

2 Задачи, решаемые в ВКР:

а) Разработать алгоритмы выделения отвлечий;

б) Реализовать описанные алгоритмы;

в) Перестроить граф в соответствии с используемой моделью в компании;

г) Проанализировать результаты работы алгоритмов;

д) Интегрировать разработки в инфраструктуру компании.

3 Число источников, использованных при составлении обзора: 0

4 Полное число источников, использованных в работе: 0

5 В том числе источников по годам:

Отечественных			**Иностранных**		
Последние 5 лет	От 5 до 10 лет	Более 10 лет	Последние 5 лет	От 5 до 10 лет	Более 10 лет
0	0	0	0	0	0

6 Использование информационных ресурсов Internet: нет

7 Использование современных пакетов компьютерных программ и технологий:

Пакеты компьютерных программ и технологий	**Раздел работы**
Пакет `tabularx` для чуть более продвинутых таблиц	??, Приложения ??, ??
Пакет `biblatex` и программное средство `biber`	Список использованных источников

8 Краткая характеристика полученных результатов

9 Гранты, полученные при выполнении работы

10 Наличие публикаций и выступлений на конференциях по теме выпускной работы

По теме этой работы был сделан доклад на Конгрессе Молодых Ученых.

Обучающийся Савон Ю.К. _____

Руководитель Ульянцев В.И. _____

«____» _____ 20__ г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 5
1. Обзор 7
1.1. Описание предметной области 7
1.2. Анализ 7
1.3. Постановка задачи 8
1.4. Выводы по первой главе 8
2. Теоретическая часть 9
2.1. Модель предлагаемого решения 9
2.2. Соответствие поставленным требованиям 10
3. Реализация 11
3.1. Подход с большим количеством рёбер входящих в вершину . 11
3.2. Подход с поиском циклов 11
3.3. Улучение кластеризации 11
4. Полученные результаты 13
4.1. Используемые данные 13
5. Заключение 14
5.1. Улучение кластеризации 14
5.2. Алгоритм поиска отвлечений по рёбрам 14

ВВЕДЕНИЕ

Эта работа посвящена исследованию в области диалоговых систем.

Диалоговая система – алгоритм, который умеет принимать участие в диалоге на естественном языке и использует правила общения между людьми.

В качестве примера диалоговых систем можно привести:

— чат-боты

— голосовые помощники

— автоответчики в колл-центрах

Такие диалоговые системы могут быть как довольно простыми (например чат-бот отвечающий на заранее известный набор команд), так и сложными (например бот, отвечающий на вопрос на естественном языке и в качестве ответа возвращающий некоторую информацию из базы знаний).

В последнее время набрали популярность технологии распознования и генерации речи, которые позволили создавать диалоговые системы, ведущие телефонный разговор. Данная работа заточена под алгоритмы для телефонных звонков.

Такие звонки с одной стороны должны быть не отличимы от звонков человека, с другой они должны придерживаться некоторого сценария.

Сценарий диалога с оператором — некоторый алгоритм, предоставленный человеку, который звонит по заданному набору телефонных номеров. Целью сценария обычно является получить или донести до клиента информацию.

Несмотря на то, что под сценарием диалога мы понимаем некоторый алгоритм, необходимо понимать что для человека и диалоговой системы это принципиально разные сущности. Между скриптом[1] и алгоритмом, с точки зрения набора действий для машины есть большая разница.

Для человека это скорее список вопросов которые он должен задать и информация о возможных продуктах и услугах. Кроме того человек может помнить некоторые факты и выдавать их дополнительно зави-

[1]Здесь и далее в тексте **скрипт** и **сценарий** будут использоваться как синонимы

симости от контекста. он сам умеет обрабатывать ситуации такие как отвлечение от основных вопросов или переспрашивание.

Для графа же, любую реакцию надо прописывать, все возможные данные хранить и обновлять. Кроме того есть требование поддерживать этот скрипт доступным для восприятия человеком (например лингвистом), поскольку возникает необходимость в ручном анализе и редактировании.

Телефонная диалоговая система — программа, которая используя сценарий умеет проводить диалог с клиентом, интерпретировать и записывать информацию полученную от клиента, а так же состояния завершенного разговора. Кроме того робот умеет отвечать на заранее прописанный в скрипте набор вопросов и возвращаться обратно к диалогу.

На данный момент существуют графы для диалогов, которые создаются вручную. Но писать их долго, а продумывать все важные случаи реакций сложно и трудоёмко.

Кроме того хочется иметь возможность усложнять вариативность диалогов. В связи с этим, ставится глобальная задача по представлению набора диалогов в виде графа.

Поскольку некоторые из вопросы или дискуссии могут возникнуть в любом месте диалога. И количество таких случаев достаточно велико. То было решено отделить их в отдельные подграфы и сделать возможность переходить в них при некоторых условиях из каждой вершины. Вдальнейшем мы будем называть такие случаи **отвлечениями**.

В работе будут рассмотрены различные алгоритмы поиска таких отвлечений. А так же рассмотренны варианты улучшения для кластеризации, использующие данные о контексте.

ГЛАВА 1. ОБЗОР

1.1. Описание предметной области

Изначальная общая задача стоит следующая: есть диалоги, необходимо восстановить граф и найти отвлечения.

На данный момент кроме восстанавливаемой нами модели есть граф который составляется вручную. В этом графе вершинами являются кластеры фраз диалоговой системы, а рёбрами фразы человека. То есть при ответах человека происходит переход в разные новые вершины.

Особенностями данной структуры, которые важно отметить, является то, что помимо обычных переходов существуют так же скрытые переходы, которые по умолчанию могут встретиться в любом месте. В качестве примера можно привести вопрос: ”А какую компанию вы представляете?”. При звонке люди могут задать этот вопрос не сразу.

Для того, чтобы не рассматривать каждый такой случай при переходе из каждой вершины, выделяются **отвлечения**.

Отвлечение - вопрос, который может быть задан в любом месте диалога.

В задаче восстанавления будет использоваться несколько другой граф. В нём вершинами будут являться кластеры и фраз оператора, и фраз человека. А ребрами последовательные пары фраз из кластеров в диалоге.

1.2. Анализ

На вход подаётся набор диалогов по которым нужно получить граф для диалоговой системы, который бы мог проводить аналогичные диалоги.

Граф, если получать его путём обычной кластеризации операторских фраз, получается очень громоздким. В нём плохо видно структуру, его сложно анализировать.

Поскольку конечной целью является построить автомат аналогичный тому, что работает в продуктовой части команды, то появляется требование привести его в состояние, когда его можно изучать вручную.

То есть так же как и в графе создаваемом вручную, появляется необходимость выделить отвлечения. В этом случае его структура ста-

новится более удобной для изучения. Его можно использовать как вспомогательный инструмент.

Такие выделения отвлечений позволяют грамотно обрабатывать сценарии, которых не было в изначальном наборе диалогов. Если такие отвлечения не выделить, то в случае вопроса, не предусмотренного в этом месте диалоговая система либо ответит невпопад, либо зависнет.

Необходима хорошая кластеризация текстов, поскольку любой алгоритм выделения отвлечений так или иначе будет опираться на эту кластеризацию. При таких кластеризациях полезно учитывать имеющуюся информацию, то есть фразы до текущей и после.

1.3. Постановка задачи

В связи с описаной выше проблемой громоздкости графа, появляется задача выделить отвлечения, которые могли бы встретиться в любом месте.

В случае, когда диалог проводит оператор такие вопросы нет необходимости расписывать, но диалоговой системе нужен чёткий скрипт.

Необходимо выделить отвлечения и перестроить граф таким образом, чтобы он покрывал большее количество диалогов и его было проще анализировать.

1.4. Выводы по первой главе

Рассмотрена предметная область графовой структуры диалогов. Разобрана структура продуктового графа и структура графа для восстанавливаемой модели. Проведён анализ задачи восстановления графа. Поставлена задача выделения отвлечений.

ГЛАВА 2. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

2.1. Модель предлагаемого решения

Поскольку граф на данном этапе всё ещё перестраивается, то информация о произнесенных человеком фразах является полезной, так как содержит в себе контекст. Поэтому мы оставляем фразы человека в качестве вершин.

Нужно понимать, что изначально кластеризация проводилась только по фразам оператора. Фразы вершин же были разделены на группы, где для каждой группы совпадала предыдущая и последующая вершины оператора. И уже внутри этих групп бились на некоторые подгруппы.

Рассмотрим два подхода в решении задачи выделения отвлечений:

— Первый подход заключается в том, чтобы выделить вершины, в которые идёт много рёбер. Порог считается функцией от количества кластеров на которые бъются фразы оператора.

  Мы предполагаем, что поскольку отвлечение встречается в разных местах, то и рёбра будут идти в него из множества различных вершин. Такая гипотеза является хорошей, поскольку в обычном графе в вершину обычно приходит одна или две ветки, в случае же отвлечения их должно быть много, или же оно встречается крайне редко.

— В качестве другого подхода можно выделить циклы и сказать, что вершина следующая в диалоге за вершиной повторения с некоторой вероятностью будет являться началом отвечения.

  Здесь мы пользуемся наблюдением, что после отвлечения на стронний вопрос, оператор зачастую повторяет ту же или схожую фразу для возвращения в сценарий. Более того, эта идея используется в графе, который реализован для реального окружения. Там диалоговая система так же повторяет фразу, её сокращенную версию или её иную формулировку, которую произносил, перед тем, как перейти в отвлечение.

Поскольку подходы используют разные идеи их так же имеет смысл комбинировать и использовать данные полученные в обоих подходах.

2.2. Соответствие поставленным требованиям

Такие алгоритмы позволят найти для диалогов которые были проведены с системой, отвлечения, и сопоставить их с выделенными лингвистами вручную для графа.

ГЛАВА 3. РЕАЛИЗАЦИЯ

3.1. Подход с большим количеством рёбер входящих в вершину

Выбираются вершины, в которые входит много ребер. Для значений размеров кластеров в интервале от двенадцати до пятнадцати был выбран параметр три. При увеличении количества кластеров соответственно должен увеличиваться и порог.

Во время тестирования на реальных диалогов человека с человеком была выявлена важная особенность. На работу алгоритма очень сильно влияет качество кластеризации. Поскольку фразы в телефонных диалогах не всегда верно переводятся в текст, сами фразы сравнительно короткие. А в случае людей-операторов ещё и очень вариативные, то кластеризация оказалась очень некачественной.

3.2. Подход с поиском циклов

После кластеризации ищутся пары вершин для которых кластеры совпадают. Так же было добавлено ограничение на расстояние между ними в диалоге. Оно должно быть не слишком велико, поскольку ясно что отвлечение бывает длинным очень редко, а иначе можно случайно выкинуть почти весь диалог.

В этом случае в качестве потенциальных отвлечений мы выбираем вершины человеческих фраз которые первые следуют после начала цикла. После перевода графа в продуктовый режим у нас эти вершины станут рёбрами и таким образом это будут триггеры, которые будут перенаправлять русло диалога в отвлечение.

3.3. Улучение кластеризации

В качестве решения проблемы некачественной кластеризации было предложено использовать данные из фраз человека. До этого для всех фраз человека между двумя фразами оператора, они кластеризировались и никак не использовались.

Было решено кластеризировать тексты пользователей. Но поскольку как описывалось выше, кластеризация не достаточно хорошая, то необходимо было отсеять плохие кластеры во избежание каскадных ошибок.

Для этого использовалось попарное сравнение фраз внутри каждого кластера. Для этого сравнивались наборы слов внутри фразы с весами. Если точность превышала порог 0.5, то пара считалась хорошей.

Константа 0.5 была выведена эмперически. Для большего значения в кластре начинало содержаться большое количество фраз разных по смыслу.

Проверять все пары оказалось очень долго $O(n^2)$, а в предыдущей части восставновления все операции имели ассимтотику не более чем $O(n \log n)$, то получилось так, что эта часть занимала значительно больше половины времени от всего восстановления графа.

Тогда было решено для каждого кластера брать случайную выборку, равная утроенному размеру кластера, ассимптотически это занимало уже $O(n)$. В силу достаточно больших размеров кластеров (размер их для основной массы данных состовляет несколько сотен фраз нескольких сотен фраз), статистически показывало те же результаты, что и полная выборка.

Ниже приведён псевдокод для данной функции:

ГЛАВА 4. ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

4.1. Используемые данные

В качестве данных использовалось два принципиально разных типа наборов диалогов. Первые проводились уже с существующим роботом скриптом. Второй тип, это данные диалогов человека с человеком.

Особенность первого заключается в том, что там легко кластеризовать фразы оператора, так как они произносятся всегда одинаково. Так же для этого случая есть возможность перевести полученный граф в тот же формат и сравнить полученный результат с оригиналом.

Особенности второго соответсвенно следующие: во первых там говорят разные операторы и у них разный стиль подачи одних и тех же данных. Во вторых диалоги более сложные и зачастую отходят от скрипта. В третьих люди решают более сложные вопросы и умеют давать ответы на не предусмотренные скриптом вопросы. На этот вариант данных стоит ориентироваться, но в связи с отсутствием оригинального скрипта в удобном формате напрямую сравнить полученный результат представляется возможным только вручную.

В обоих случаях есть сложности с переводом речи людей в текст, поэтому иногда даже рассматривая текст диалога вручную нельзя понять что человек имел ввиду.

ГЛАВА 5. ЗАКЛЮЧЕНИЕ

5.1. Улучение кластеризации

Алгоритм вносит небольшие изменения и способен объединять некоторые одинаковые кластеры.

5.2. Алгоритм поиска отвлечений по рёбрам

Для данного алгоритма критично качество кластеризации и наличие небольшого количества выбросов.

В случае с данными из реальных диалогов зачастую не было видно правильно выделеных отвлечений из-за смешения кластеров и операторских и человеческих вершин.

Для данных из диалогов с существующей диалоговой системой алгоритм находил практически все отвлечения. Основная причина столь большого различия в том, что операторские фразы во втором случае кластеризуются практически идеально.