федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Факультет информационных технологий и программирования
Направление (специальность) Прикладная математика и информатика
Квалификация (степень) Магистр прикладной математики и информатики
Кафедра компьютерных технологий Группа М4238

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

на тему

Выделение групп пользователей в социальных медиа по их интересам и поведению на основе множества источников данных

Автор магистерской диссертации Дмитриев С.С. _____
Научный руководитель Фильченков А.А. _____
Руководитель магистерской программы Васильев В.Н. _____

К защите допустить

Заведующий кафедрой Васильев В.Н. _____
«____» _____ 2016 г.

Санкт-Петербург, 2016 г.

Магистерская диссертация выполнена с оценкой _____

Дата защиты «____» _____ 2016 г.

Секретарь ГАК _____

Листов хранения _____

Чертежей хранения _____

федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

АННОТАЦИЯ
ПО МАГИСТЕРСКОЙ ДИССЕРТАЦИИ

Студент Дмитриев С.С.
Факультет информационных технологий и программирования
Кафедра компьютерных технологий Группа М4238
Направление подготовки Прикладная математика и информатика
Квалификация (степень) Магистр прикладной математики и информатики
Специальное звание
Наименование темы Выделение групп пользователей в социальных медиа по их интересам и поведению на основе множества источников данных
Научный руководитель Фильченков А.А., кандидат технических наук, доцент
Консультант

КРАТКОЕ СОДЕРЖАНИЕ МАГИСТЕРСКОЙ ДИССЕРТАЦИИ И ОСНОВНЫЕ ВЫВОДЫ

объем 41 стр., **графический материал** – стр., **библиография** 39 наим.

Направление и задача исследований

Целью данного исследования является создание алгоритма выделения групп пользователей социальных сетей на основе их социальных связей и поведения в социальных сетях

Проектная или исследовательская часть (с указанием основных методов исследований, расчетов и результатов)

В рамках данной работы предложен подход, позволяющий выделять подгруппы у выбранной группы пользователей в социальных сетях, основывающийся на социальных связях и видимом поведении на публичных страницах. В основе предложенного подхода лежат несколько методов и концепций: представление социальных связей в виде графа, случайные марковские поля, а так же семантический анализ. В качестве примера использования подхода взята группа футбольных болельщиков, и подгруппа радикальных футбольных болельщиков. Были использованы данные пользователей из социальной сети Vk.com. Достигнуты следующие показатели для группы футбольных болельщиков: точность 84%, f-мера 0.46. Данный подход нов и так же может применяться для других групп и подгрупп пользователей.

Экономическая часть (какие использованы методики, экономическая эффективность результатов)

Данная работа не предполагает извлечения экономической выгоды из полученных результатов

Новизна полученных результатов

В рамках описываемого исследования представлен подход, позволяющий определять принадлежность пользователя к определенной группе на основе его социальных связей и публичного поведения в социальной сети. Полученный подход является способом построения модели, не применявшимся для решения подобной задачи ранее.

Является ли работа продолжением курсовых проектов (работ), есть ли публикации

Работа не является продолжением курсовых проектов. На тему диссертации имеются публикации. Дмитриев С. С. Выделение группы радикальных футбольных болельщиков в социальных медиа по их интересам и поведению на примере сети Vk.com //Материалы всероссийской научной конференции по проблемам информатики СПИСОК-2016 - 2016. - принято в печать.

Практическая ценность работы. Рекомендации по внедрению

Полученный алгоритм дает возможность определить является ли член выбранной группы так же членом ее подмножества. Это может быть использовано правоохранительными органами, т.к. алгоритм позволяет выделить, например подгруппы, склонные к бандитизму, пользователей потенциально более способных на совершение незаконных действий, нежели среднестатистический пользователь. Так же алгоритм может быть использован для усовершенствования таргетированной рекламы, например для выделения подгруппы фанатов определенного бренда из группы его покупателей.

Выпускник _____

Научный руководитель _____

«____» _____ 2016 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Последнее время социальные медиа набрали огромную популярность. Такие сайты как Facebook.com[1], Vk.com[2], Twitter.com,[3] обладают огромной аудиторией. Совокупный размер аудитории существующих социальных сетей составляет более двух миллиардов пользователей, и их число постоянно растет. Они создают огромные массы контента, состоящего из их мнений и точек зрения. Так в течении суток на Facebook.com более 4.5 миллиардов раз пользователи ставят лайки, оставляют более 700 тысяч публичных комментариев, публикуют более 100 миллионов фотографий[4]. Однако содержание этой информации в основном остается не использованным. Тогда как оно может быть крайне важным.

Такие данные могут быть использованы для определения интересов, предпочтений и иных личных свойств пользователя. Часть подобной информации, пол, возраст, местоположение, увлечения, может быть указана в профиле пользователя. Однако, зачастую, такие данные могут быть неполными, а иногда и неверными. А некоторые признаки, например, вероисповедание, политические взгляды или же принадлежность к неким общественным движениям обычно опускаются. Из-за этой неполноты возникает задача восстановления информации о пользователе.

Получение таких данных может быть полезно как бизнесу, так и государству [1]. Используя восстановленные характеристики, можно уточнять таргетированную рекламу [2]. Имея дополнительные данные об увлечениях людей, можно определять возможных преступников, что позволит предотвращать возможные нарушения или же прогнозировать конфликты [3].

Существует множество исследований о восстановлении данных, явно неуказанных в профилях пользователей [4—7]. В них показано, что на основе информации о пользователе, его поведении в социальном медиа, можно с высокой точностью восстановить некоторые характеристики. Отдельной задачей стоит определения принадлежности пользователя к определенной группе, такой как, например, группа консерваторов или же группа любителей продукции Apple. Для решения этой задачи часто используют данные о социальных

[1] https://facebook.com

[2] https://vk.com

[3] https://twitter.com

[4] https://zephoria.com/top-15-valuable-facebook-statistics/

связях пользователей. Показано, что они влияют на поведение человека, на его взгляды [8, 9].

В описываемом исследовании представлен подход для выделения подгруппы пользователей из определенной группы пользователей. Работа предложенного алгоритма продемонстрирована на примере выделения подгруппы радикальных футбольных фанатов из группы футбольных болельщиков. Предполагается, что описываемый подход может быть использован на других группах и подгруппах пользователей различных социальных медиа.

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

В данной главе описаны основные понятия, использующиеся в предметной области.

В разделе 1.1 описана задача о восстановлении характеристик пользователя, основные трудности, возникающие при решении этой задачи.

В разделе 1.2 рассмотрена задача о разделении пользователей на группы, определения их принадлежности к подгруппам.

В разделе 1.3 разобраны существующие методы и решения, полученные результаты в разнообразных исследованиях, посвященных выделению групп пользователей.

В разделе 1.4 представлено формальное описание задачи, исследуемой в данной работе.

1.1. Задача восстановления характеристик пользователя

Задача о восстановлении характеристик пользователя, она же задача о профилировании пользователей, заключается в определении неизвестных характеристик пользователя, на основе имеющихся данных с определенных ресурсов. Под ресурсами подразумеваются как социальные медиа, так и любые другие сайты, обладающие системой регистрации, так же данные могут собираться одновременно с нескольких ресурсов.

Проблема восстановления характеристик пользователей часто встречается при необязательности заполнения некоторых полей. Часто необязательно заполнять пол, возраст, физические данные, тогда как эти данные могут быть очень важны для определенного рода ресурсов [5—7]. Вычисление этой информации дает возможность улучшить качество таргетированных сервисов.

В социальных сетях зачастую указывается вовсе неверная информация, например, данные о возрасте. Так появляется подтип задачи восстановления — определение ошибочных свойств пользователя и восстановление как неизвестных.

Помимо широко используемых характеристик пользователя, некоторые исследования посвящены таким задачам, как определение хронотипов пользователей [4]. Данные о биоритмах людей полезны врачам и рекрутерам, оценивающим подойдет ли выбранный человек на определенную должность.

Существуют исследования определяющие психотип пользователя, используя лишь данные о них из их же аккаунтов с социальных медиа [5]. Подобные работы дают новые пути исследований для психологов.

Описанные проблемы зачастую сводятся к задачам кластеризации, регрессии или классификации.

1.2. Задача о выделении групп пользователей

Задача о выделении групп пользователей является подтипом задачи о восстановлении характеристик пользователя. Она заключается в определении принадлежности выбранного пользователя к определенному сообществу на основе имеющихся данных. Под сообществом подразумевается не какая-либо страница в социальном медиа, а множество пользователей, объединенных общей ролью, функцией или атрибутами [10].

Определение психотипа, хронотипа, задачи сводящиеся к выделению группы пользователей. Так, можно поставить задачу, как принадлежность пользователя к группе холериков, сангвиников, флегматиков, меланхоликов.

Ярким примером задачи о выделении групп пользователей может служить проблема определения принадлежности пользователя к политическому движению [11—15]. В статье [11] описывается подход по определнию политических предпочтений пользователей. Исследователи предложили статистическую модель, в которой строится пространство идеологий, с построенными на них известными публичными сраницами в Twitter.com, для которых была определена их идеология. Дальше в пространство помещались пользователи, их подписчики, координаты которых определялись исходя из их подписок.

Для членов сообщества справедливы следующие свойства [16]

1. Они более тесно связаны другом с другом
2. Сообщества могут пересекаться, что согласуется с наличием нескольких социальных ролей у человека в обществе
3. Сообщества могут иметь иерархическую структуру

Для задач определения сообществ трудно использовать классические алгоритмы кластеризации, используя, например, только данные профилей. Как правило для выявления сообществ используют следующие методы [16]:

1. Модель случайного графа(null model). Заключается в сравнение исследуемого графа социальных связей с графом с равномерным распределением ребер для каждой вершины. Классическое решение, использу-

ющее целеваую функцию $modularuty$ работает для непересекающихся множеств [17], что часто не характерно для социальный сетей. Поэтому используется её обобщение [18].

2. Модель блуждания (random walk). В этой модели граф разбивается так, чтобы минимизировать длину описания случайного блуждания в графе. Для оценки этой длины кода можно брать энтропию [19].
3. Локальное изучение. Заключается в рассматривании отношения количества внутренних ребер и треугольников к внешнему и максимально возможному числу [20].
4. Агентская модель. Заключается в моделировании общения между узлами графа. Каждому узлу присваивается некое сообщество. Затем повторяется следующее действие: выбирается слушающий узел, смежные узлы отдают случайным образом одно из своих сообществ, причем, чем больше повторений сообщества у узла, тем вероятнее, что оно отправится. Слушающий узел добавляет к себе то сообщество, которое получил больше всего. В завершении алгоритма для каждого узла определяется сообщество, как самое часто сохранненное в него сообщество.

Иногда задачи о выделении групп пользователей решаются путем кластеризации пользователей, как например в статье [11].

В основе части исследований по выделению групп пользователей лежат данные, на основе которых происходит восстановление информации. Не редко это уже существующая информация из профилей пользователя. Так же часто используется информация из публичных сообщений, медиа-контент, такой как, видео, фотографии, музыка, социальные связи, поведение пользователя и так далее.

Главной проблемой в решении подобных задач является сведение задачи к математической модели. Приведение сырых данных к числовому виду так же зачастую бывает крайне непростой задачей.

1.3. Обзор существующих решений восстановления характеристик пользователя

В прошлом разделе было отмечено, что основными проблемами выделения групп пользователей является приведение задачи к некой математической модели и генерация дискретных данных. Для более общей задачи проблемы аналогичны. Существующие решения можно условно разделить на несколько

видов, основанных на виде решения проблемы: методы, использующие данные из профилей, использующие публичные текстовые сообщения, использующие социальные связи, использующие медиа-контент, использующие ссылки на другие социальные медиа в профиле. В настоящем разделе они будут описаны.

1.3.1. Методы, использующие данные из профилей

Одним из популярнейших решений является подход, использующий известные признаки пользователей, взятые из профилей. Так, например, опишем подход из статьи [21].

И так, для каждого пользователя собирались все публичные характеристики его профиля, такие как, пол, возраст и так далее.

Часто не вся информации оказывается необходимой для исследования. Поэтому часть признаков необходимо обозначить как менее информативные. Это ставит задачу определение информативности признаков. Например в описываемой статье данные охарактеризовали как мало информативные по следующим признакам: значение признака не меняется в зависимости от пользователя, характеристики, которые трудно представимы в численном виде, данные которые слишком редко указывались и информация, которая не стала бы предсказывающей, например ссылка на персональный сайт. Такие данные были удалены.

Так же нередко заполненных признаков пользователем бывает недостаточно, так как в действительности каждый признак с разной силой определяет пользователя. Поэтому обычно для признаков или же для векторов признаков считаются веса.

Далее, как правило, решают задачу классификации или кластеризации. Где классы и кластеры соответствуют принадлежности пользователя к группе или наоборот.

1.3.2. Методы, использующие публичные текстовые сообщения

Большинство социальных сетей позволяет пользователю оставлять публичные сообщения, без конкретного адресата, которые потом могут быть прочитаны другим людьми. Так же пользователь может кастомизировать такие свои персональные данные как например, имя, фамилия или ник. Использование текста такого рода возвращает нас к проблеме приведения данных к числовому виду.

Сущесвует набор примитивных решений, которые позволяют привести публичный текст к виду численной характеристики.

Одним из таких решений является использование словарей и последующего его использования для поиска соответствий в исследуемом тексте. Такой подход обладает существенным недостатком, словари приходится составлять вручную. Наглядным примером является задача определению пола по имени [22].

Так же популярна задача определения пола по текстовым сообщениям. В статье описывается, что женщины чаще используют личные местоимения, считая вхождения таких местоимений [23].

Как следствие ручного составления словаря, подобные подходы становятся более трудозатратными при исследовании мультиязычных данных.

Другим методом приведения текста к численными данными является латентный семантический анализ [24]. Этот метод позволяет уйти от ручного составление словарей, решая тем самым основную проблему приведения текста к дискретному виду.

1.3.3. Методы основанные на использовании социальных связей

Пользователь социальной сети определяется не только набором своих характеристик в профиле. Каждый пользователь является обладателем набора социальный связей. Это могут быть список друзей, подписчиков, подписок на определенные публичные страницы. Многие работы о выделении групп пользователей [8], как, например, в статье [11] используют граф социальных связей.

В работе выделяются группы либеральных пользователей твиттера. В исследовании строится идеологическая плоскость, на которой размечаются аккаунты твиттера с изначально известной позицией. Делается предположение, что вероятность того, что два пользователя соединены на графе зависит от дистанции между ними на идеологической плоскости. Получается, что чем больше у пользователя подписок на либеральные твиттер аккаунты, тем он сам более либерален.

Минус такого подхода заключается в ручном составлении списка аккаунтов с известной политической позицией. Для групп другого вида такой подход может оказаться вовсе невозможным, из-за невозможности четкого определения известных членов группы.

Существует работа в которой пользователь рассматривается как набор из всех его подписок [25]. Без дополнительных признаков подобная модель показывает плохие результаты с точность менее 50 %.

1.3.4. Другие методы

Важной группой данных при восстановлении характеристик пользователя являются медиа данные, такие, как фотографии, видеозаписи, музыка.

В качестве примера использования фотографий рассмотрим исследование [26], определяющее гендерную принадлежность пользователя используя яркость фотографий. Как признак характеризующий пользователя были взяты разности численных значений яркости каждой пары пикселей. Минус подобного подхода заключается в его крайней ресурсоемкости. Так же при анализе фотографий пользователя часто анализируют мета-информацию файлов, как например, в статье. Минус такого подхода заключается в возможности изменения этой мета-информации на не соответствующую действительности.

Существует множество исследований, использующих в качестве основы своей модели информацию о музыке пользователя [27, 28]. На таких ресурсах как last.fm[1] может использоваться, такая информация, как наиболее прослушиваемые композиции [27]. Так же применяется анализ самих аудиофайлов и отображения их в такие характеристики, как ритмичность, скорость бита, и тому подобные [28]. Минус алгоритмов, основанных на музыкальных предпочтения заключается в слабой точности результатов без дополнительных параметров.

Зачастую для определения принадлежности пользователя к определенной группе, используют данные о его геолокации. Для определенных типов групп, такой признак может работать достаточно точно. Так, например, в недавно рассекреченном проекте skynet [2] по определению потенциальных террористов использовались в числе прочих данные о перемещении людей. Результатом работы алгоритма являлось ложно положительное определение пользователя как террориста с вероятность менее 0.2 %. Террористов алгоритм давал определять с вероятность в 50 %.

Так иногда пользователи оставляют информацию о своих аккаунтах на других сайтах. Это могут другие социальные медиа. Использование подобной

[1] https://Last.fm

[2] http://arstechnica.co.uk/security/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/

информации требует решения дополнительной проблемы, определения правдивости принадлежности аккаунтов одному человеку [29].

Самым эффективным способом решения задачи выделения группы пользователей является использование нескольких видов данных.

1.4. Постановка задачи настоящего исследования

Зачастую имея дело с задачами восстановления характеристик, мы имеем ситуацию, когда нам доступны некоторые пользователи принадлежащие, какой-то группе, скажем, группе любителей определенной марки одежды, и так же известны те, кого можно назвать фанатами этого бренда, людьми не пропускающими ни одной новинки. Получая на вход эти данные хочется определять для любого пользователя, является ли он фанатом, или просто любителем.

Имея задачи, подобные выше описанной, определим проблему настоящего исследования следующим образом. Существует набор пользователей, публичных страниц(сообществ)[3], группа пользователей и сообществ и подгруппа, которую необходимо выделить из существующей группы. Для каждой публичной страницы известны все ее подписчики и все публичные сообщения созданные на этой страницей. Для каждого пользователя известны все его социальные связи, будь то, подписки, все его друзья и все подписанные на него аккаунты. Так же доступна вся публичная информация о поведении пользователя, выраженная в одобрении определенного сообщения. Для группы и подгруппы пользователей и сообществ имеется набор людей и публичных страницы принадлежащих к ним. Необходимо определить для каждой публичной страницы и каждого пользователя группы определять являются ли они членами подгруппы.

[3] здесь и далее под сообществом и публичной страницей понимается страница социального медиа, обладающая набором подписчиков, администраторами, публичными и не публичными сообщениями

Дадим формальную постановку задачи:

Пусть

$$G, G_{sub} : G_{sub} \subset G, f(g_i, g_j) = [0, 1],$$

где 1 обозначает связь между элементами

$$g_i, g_j \in G.$$

Тогда зная, что

$$g_1, g_2, g_3 ... g_k \in G_{sub},$$

нужно уметь определять входит ли $g \in G$ в G_{sub}

Имеющиеся данные можно представить в виде смешанного графа. Где узел это либо пользователь, либо публичная страница, с сопутствующей информацией, а ребра между узлами есть отношение подписка-подписчик. Сопутствующая информация группы есть ее публичные сообщения и список пользователей одобривших сообщения.

В данной задаче мы имеем три типа данных: информацию о социальных связях, тестовую информацию, а так же отношение пользователя к тестовым сообщениям.

1.5. Выводы по главе 1

В данной главе была разобрана задача восстановления данных пользователей и подзадача выделения групп пользователей. Были описаны методы решений основанные на различных видах данных и различных моделях. Была описана задача, которая решается в данном исследовании.

Сформулируем плюсы и минусы описанных подходов.

Методы, использующие данные профилей в чистом виде, могут давать плохие результаты в близких группах и подгруппах. Так же для них необходимы большие обучающие выборки.

Проблема методов использующих текст, в том, что в социальных медиа пользователи часто закрывают свои публикации. Однако использование текста для классификации тематик групп может дать хороший результат.

Задачи о выделении групп на основе социального графа наиболее близки к поставленной задаче. Однако существующие решения базируются на больших размеченных данных или решают задачу разделения на неизвестные груп-

пы, что не позволит добиться при их использовании хорошего качества результата.

ГЛАВА 2. ОПИСАНИЕ ПРЕДЛАГАЕМОГО ПОДХОДА

В данной главе описаны структуры данных, алгоритмы и методики, применяющиеся при решении поставленной в данном исследовании задаче.

В разделе 2.1 приведены основы текстового информационного поиска.

В разделе 2.2 описаны подходы для анализа публичных текстовых сообщений.

В разделе 2.3 описан подход, называемый случайными марковскими полями.

В разделе 2.4 описан собственный подход, представляющий из себя модификацию, описанного метода в прошлом разделе.

2.1. Основы текстового информационного поиска

Важной частью данного исследования является анализ сообщений, оставленных в публичных сообществах. Задачей этого анализа является определение тематики сообщения, в рамках настоящего исследования встает задача определения принадлежности текста к теме выделяемой группы.

Опишем термины, использующиеся в дальнейшем повествовании [30].

Термин, он же слово, атомарная лингвистическая единица.

Документ — конечный набор терминов. В контексте поставленной задачи, документом будет являться публичное текстовое сообщение.

Коллекция — набор документов.

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}.$$

Где D — коллекция, а D_i — документ.

Словарь — набор всех терминов встречающихся во всех коллекциях.

$$T = \{t \colon t \in \bigcup_{j=1}^{n} \mathcal{D}_j\} = \{t_i\}_{i=1}^{m}.$$

Эти понятия вводятся, для пояснения подхода, который в общем заключается в том, что у каждого термина есть характеристика, обозначающая его принадлежность к документу. Для описания такой подхода удобно использовать двумерную матрицу, в которой каждый столбец будет представлять вектор соответствующий документу.

Часто в задачах анализа текста не используется порядок слов. Так получается «bag of words», неупорядоченный набор терминов.

Для вычисления значений элементов матрицы как правило используют формулу *TF-IDF*, она имеет следующий вид:

$$d_{ij} = \mathrm{tf}_{ij} \cdot \log \frac{n}{\mathrm{df}_i}. \quad (1)$$

где tf_{ij} —число встреч термина i в документе j, df_i — число документов в которых встречается термин i, а gf_i — число встреч термина i во всей коллекции.

Формула может меняться в зависимости от исследования. Так же могут применяться и совсем иные подходы. Однако в рамках настоящего исследования они не применяются.

2.2. Определение тематики публичных сообщений

В данном разделе описаны применявшиеся методы для определения тематики сообщений.

2.2.1. Тривиальное решение задачи

Очевидным решением задачи определения принадлежности документа к определенной тематике является подсчет вхождения терминов из предварительно составленного словаря, вмещающего в себя термины искомой тематики.

Для этого сперва необходимо составить непосредственно этот словарь. Для чего необходимо определить критерии принадлежности термина к тематике. Такая задача требует лингвистического анализа.

Для получения более точных результатов, можно отказаться от представления документа, как неупорядоченного списка слов, и искать помимо отдельных терминов так же и фразы, короткие упорядоченные наборы из слов.

Таким образом, при использовании данного метода коллекция хранится как таблица, состоящая из строк — документов, столбцов — терминах составленного словаря и ячейки содержащей, информацию, обозначающую принадлежность соответствующего термина в соответствующий документ. Данный подход по сути является некой модификацией описанного ранее алгоритма подхода термин-документа.

Как уже говорилось, минус подобного подхода, в ручном составлении словаря.

2.2.2. Латентно-сементический анализ

Латентно-семантический анализ — это метод обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами в них встречающимися, сопоставляющий тематики всем документам и терминам. Таким образом автоматически решаю задачу определения тематики терминов.

Формально задачу латентно-семантического анализа можно определить так.

Пусть $D \in \mathbb{R}^{m \times n}$ — матрица «термин-документ», вычисленная каким-либо образом. Требуется выполнить следующее разложение данной матрицы:

$$D = U \cdot V^T, \quad U \in \mathbb{R}^{m \times k}, \quad V \in \mathbb{R}^{n \times k},$$

где U — матрица «термин-тема», V — матрица «документ-тема», а k — число тем. Строка матрицы U под номером i характеризует «степень принадлежности» термина i каждой из тем. Строка матрицы V под номером j обозначает «степень принадлежности» документа j каждой из тем.

Фактически данный метод можно рассматривать как нечеткую кластеризацию. Латентно-семантический анализ позволяет уменьшить набор терминов, что существенно облегчает задачу.

Существуют два подвида данной задачи, один использует вероятностную модель данных, в ячейках матрицы хранятся вероятности, другие использую особые метрики.

Формально вероятностную модель данных можно описать так.

$$p(d, w) = \sum_{t \in T} p(t) p(w|t) p(d|t),$$

где T — множество тем, $p(d, w)$ — вероятность возникновения термина w в документе d, $p(t)$ — вероятность выбрать тему t, $p(w|t)$ — вероятность выбрать термин w из темы t, а $p(d|t)$ — вероятность выбрать документ d, при условии, что выбрана тема t.

К вероятностным алгоритмам латентно-семантического анализа относят probabalistic latent semantic analysis(PLSA) [31] и так же latent Dirichlet allocation (LDA) [32].

Из невероятностных моделей следует рассказать о LSI, latent sematic indexing [33]. В методе используется сингулярное разложение, что дает возможность уменьшать объем данных, за счет увеличения плотности значений, коллекции как правило очень разрежены. К минусам этой модели можно отнести сложность при интерпретации данных.

2.3. Случайные марковские поля

Случайные марковские поля (random Markov Fields) [34] метод широко применяемый в различных областях ИИ. Его успешно используют при распознавании речи и образов, а так же в обработке текста [35, 36].

Марковским случайным полем или Марковской сетью называют графовую модель, которая используется для представления совместных распределений набора нескольких случайных переменных. Формально марковское поле состоит из нескольких компонентов.

1) неориентированный граф, где каждая вершина является случайной переменной x и каждое ребро представляет зависимость между случайными величинами u и v.

2) набор потенциальных функций для каждой клики графа. Функция представляет из себя отображение клики в неотрицательное вещественное число.

Считаем, что если вершины не смежны, то они являются условно независимыми случайными величинами.

Совместное распределение набора случайных величин в Марковском случайном поле вычисляется по формуле:

$$P(x) = 1/z \prod_{t \in T} p_(X_D),$$

где $p(X_D)$ — потенциальная функция, описывающая состояние случайных величин в k клике; z — коэффициент нормализации:

$$z = \sum_{x \in X} \prod_{k} p_k(X_k)$$

Одной из разновидностей метода случайных марковских полей является метод скрытых марковсих полей(CRF) [37, 38].

У метода есть недостатки, такие как вычислительная сложность анализа обучающей выборки, это затрудняет обновление модели с обновление обучающих данных

2.4. Модифицированные случайные марковские поля

Как сказано было в прошлом разделе случайные марковские поля обладают существенными недостатком, они медленно обновляются.

Поэтому было решено опробовать собственный упрощенный метод.

Вернемся к представлению данных в виде графа.

Введем для каждого узла характеристику p — вероятность отнесения его к определенной подгруппе. Пускай множество узлов M — это множество узлов с размеченной в ручную p. Далее для каждого смежного узла x вычисляется его p:

$$p = F_{h \in H} k * h_x$$

где H — множество признаков, таких как, например, текстовое сходство, поведенческое сходство и так далее, F — функция, считающая суммарный вклад признаков, а k — коэффициент определяющий важность параметра. В результате получается множество размеченных узлов M_1.

Путем подбора функции F можно улучшить результаты. Так, можно обучиться на выборке данных, чтобы понять какой из параметров наиболее влиятельный и представлять F в виде линейной комбинации.

Характеристика посчитана, однако, если теперь эту же характеристику пересчитать для изначально размеченных узлов, она может измениться для них. По этому процесс повторяется в рамках множества M_1. Пересчет предлагается останавлить, когда норма Фробениуса станет меньше либо равна E [39]. Важной оценкой качества такого подхода будет являться проверка изменений p размеченных узлов.

Далее алгоритм повторяет последовательность действий, до состояния полного покрытия сети. Однако этот процесс крайне ресурсоемкий и в рамках данного исследования были использованы меньшие объемы данных. Рассматривался рост в 3 шага.

2.5. Выводы по главе 2

В текущей главе были описаны некоторые алгоритмы и методики, которые используются при анализе текста, а так же в задачах структурного машинного обучения. Описанные методики и алгоритмы использовались при решении поставленной задачи.

ГЛАВА 3. РЕАЛИЗАЦИЯ ОПИСЫВАЕМОГО ПОДХОДА

В данной главе будет описан подход к решению задачи, которая была поставлена в разделе 1.4.

В разделе 3.1 описана общая схема решения исследуемой задачи.

В разделе 3.2 описана структура данных, используемая для хранения анализируемой информации.

В разделе 3.3 описан собранный набор данных, который использовался в эксперименте.

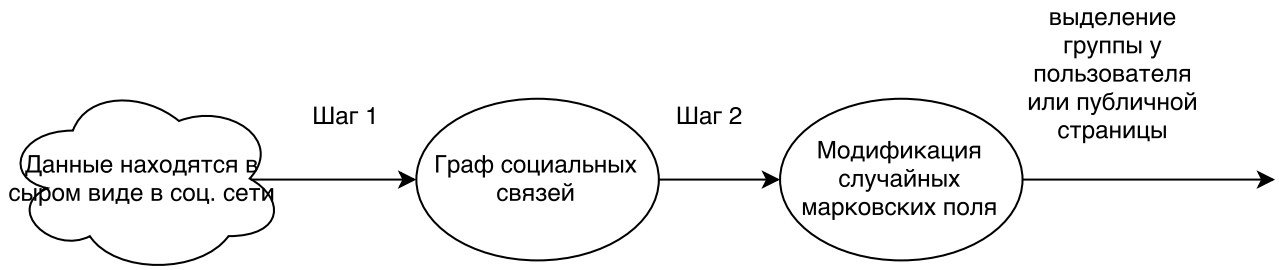
В разделе 3.4 описано применение алгоритмов из главы 2.

В разделе 3.5 кратко описаны программные средства, использованные при реализации подхода и кратко описана сама реализация.

3.1. Общая схема решения

В прошлой главе было уделено много внимания случайным марковским полям. Предлагается решать поставленную задачу использую эту концепцию, так же предлагается опробовать ее модификацию. Вместе с тем, алогритмы анализа текста помогут улучшить качество анализа тематик публичных сообщений. На рисунке 1 проиллюстрирована общая схема предлагаемого подхода.

Рисунок 1 – Общая схема решения задачи выделения подгруппы пользователей



Опишем последовательно схему представляемого подхода. Из поставленной задачи мы имеем проблему определение группы пользователей. И так, на первом этапе исходя из тематики группы необходимо собрать набор публичных страниц придерживающихся данной тематики. Для этого предлагается сделать аналог лингвистической экспертизы.

Имея набор групп определенной тематики, мы собираем всю информацию связанную с этим группами: тексты подписчики и так далее, подробнее в разделе 4.2.

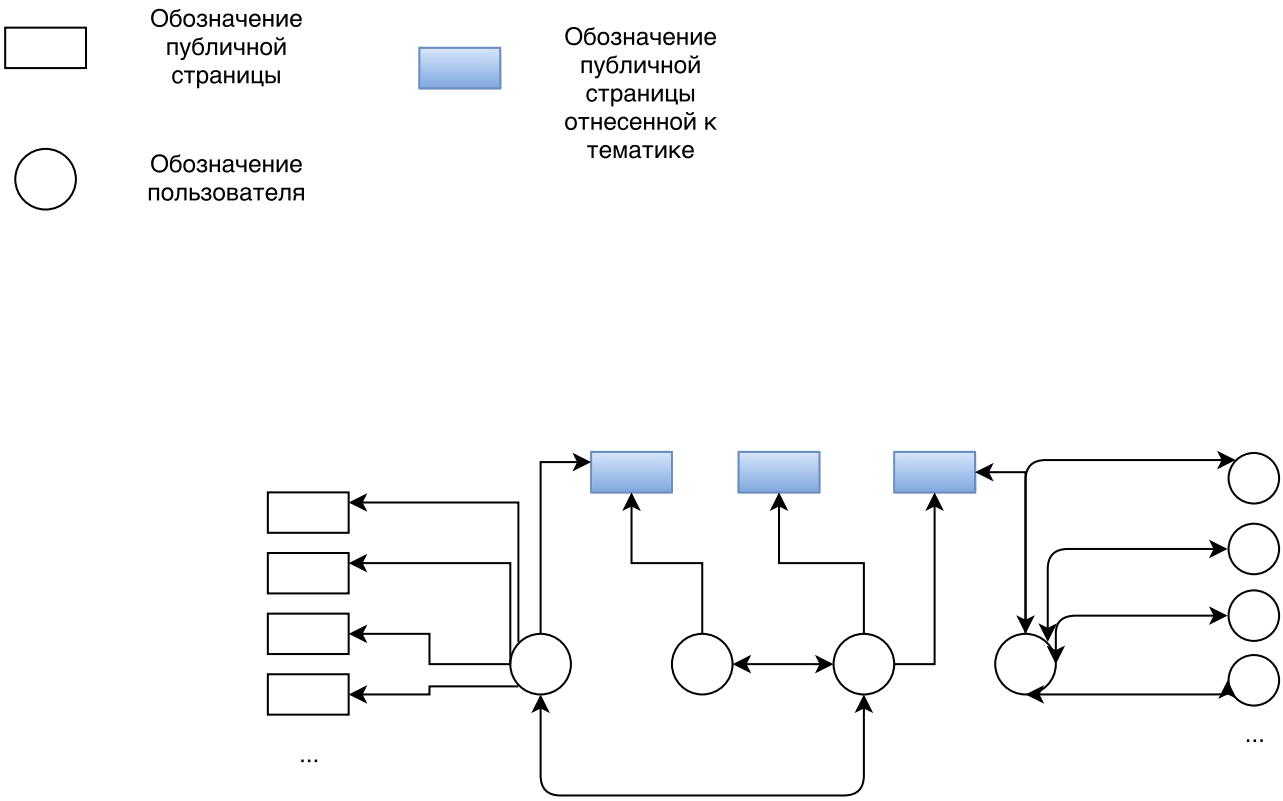
Собрав все необходимые данные строим по ним граф социальных связей с дополнительной информацией, подробнее в разделе 4.1 (шаг 1).

Преобразуем наш граф к модификации случайного марковского поля, используя дополнительную информацию, подробнее в разделе 4.3 (шаг 2).

3.2. Граф социальных связей

Для хранения собранных данных и представления их в удобном виде используется граф социальных связей. Узлом графа может являться публичная страница или же сам пользователь. Ребро обозначает подписку на публичную страницу или двусторонние отношения определяемые как взаимная подписка или отношение — дружба. На рисунке 2 проиллюстрирована схема графа.

Рисунок 2 – Граф социальных связей



Узел группы содержат о себе следующую информацию: уникальный идентификатор, список публичных администраторов, список постов группы и для каждого поста, список одобривших его. На рисунке 3 проиллюстрирована схема хранения публичной страницы.

Узел пользователя содержит о себе следующую информацию: уникальный идентификатор, фамилия, имя, текущая геолокация. На рисунке 4 проиллюстрирована схема хранения пользователя.

Рисунок 3 – Узел публичной страницы в графе

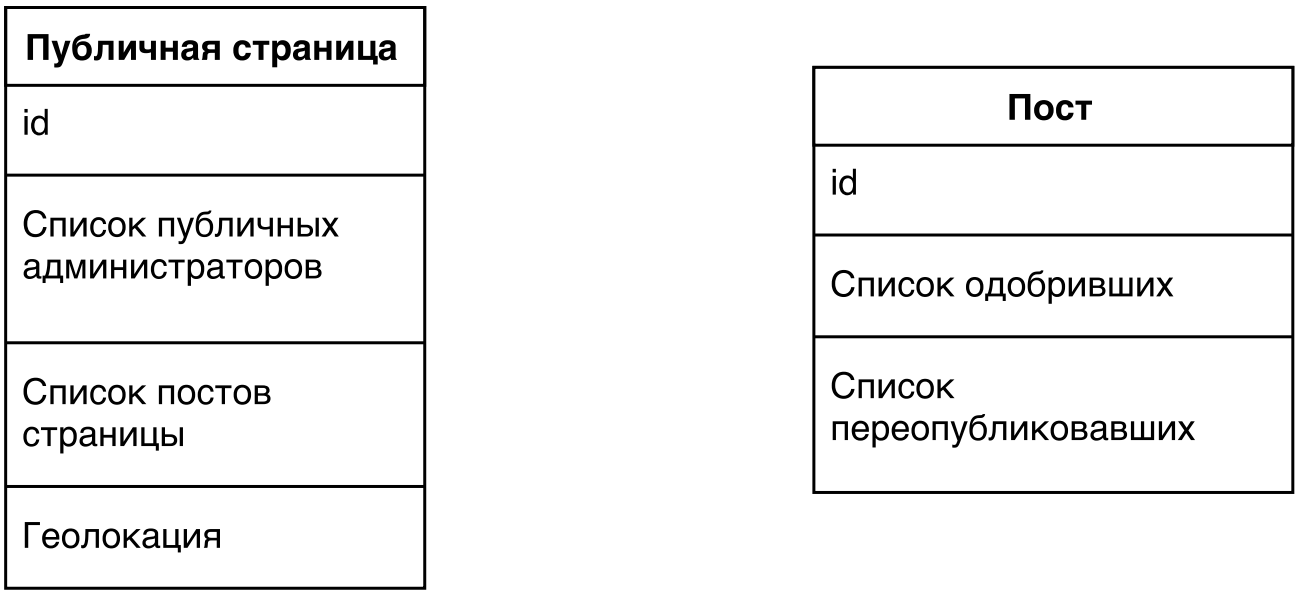


Рисунок 4 – Пользовательский узел в графе

На рисунке изображена схема хранения узлов графа.

3.3. Сбор данных

В настоящем эксперименте решается задача для группы футбольных болельщиков и подгруппы радикальных фанатов.

В рамках данного исследования используются данные из социальной сети Вконтакте[1]. Особенностью данной социальной сети является наличие двух видов публикующих страниц: обычного пользователя и группы, иначе публичной страницы. Пользователи имеют связи одновременно с группами и другими пользователями, публичные странице же связаны только с пользователями.

[1] https://vk.com

Как было отмечено выше, первым дело необходимо создать список публичных страниц тематики искомой группы. Для этого должны быть определены четкие характеристики определяемой группы.

Для группы футбольных болельщиков это: группа должна быть посвящена определенному футбольному клубу, это определялась по текстовым сообщениям, если в них присутствовали новости о футбольной команде, страница входит в группу. Для исследования были взяты группа страниц посвященных футбольному клубу "Зенит". Для определения публичных страниц входящих в подгруппу радикальных фанатов из собранных групп выбирались те, которые содержали негативные отзывы о командах соперников, ненормативные высказывания в адрес болельщиков других команд. Всего было собрано 211 групп посвященных этой тематике, 10 из которых были о футбольных хулиганах.

Выбирались примерно одинаковые по числу подписчиков сообщества, размер которых не превышал 5000 пользователей и был выше 500. Всего было собрано 225230 пользователей. 74 % собранных пользователей были мужчинами. Максимальное число подписок из перечня сообществ было —7. Сообщества сильно разнились по числу публичных сообщений. Суммарно было собрано 178022 публичного сообщения. Число лайков и репостов — 7199.

Так как при сборе подписчиков и подписок с только что добавленных узлов, граф очень быстро растет. В рамках данного исследования не проводились эксперименты с большим числом обновлений выборки. Пользователи добавлялись не более 2 раз, группы не более 3.

3.4. Использование алгоритмов

В данном разделе описано применение модифицированных алгоритмов случайных марковских полей, латентного семантического анализа. Приведены особенности реализации и использования данных методов.

3.4.1. Применение случайных марковских полей

Условия задачи ставят определенные ограничения на используемые методы. Так имеется крайне ограниченная выборка, состоящая из небольшого набора публичных страниц. Поэтому предлагается воспользоваться подходом схожим с предложенным в статье. Так мы сможем значительно увеличить объем наших данных, вместе с тем используя алгоритм предложенный в разделе 2.4 мы всегда сможем оценить качество наших результатов.

Для эксперимента было выбрано две модификации описанного в разделе 2.4 подхода. Определим эти модификации.

Множества признаков H будет одинаковым для двух модификаций, однако, оно будет отличаться от типа узла. Как уже было сказано социальная сеть Вконтакте обладает двумя типами узлов. Для узла пользователя определим два признака: наличие одобрения содержимого из смежных публичных страниц, принадлежащих к группе, и влияние характеристик смежных узлов. Для групп же: текстовая схожесть с текстами из подгруппы и влияние характеристик смежных узлов.

Использование методов, основанных на данных смежных узлов, обусловлено предположением, что пользователь, имеющий больше социальных связей с членами подгруппы, с большей вероятность сам будет принадлежать к этой подгруппе.

Использование признака одобрения контента основывается на предположении, что пользователь одобривший что-то действительно склонен одобрять данного рода информацию.

Модификации будут отличаться вычислением влияния смежных узлов.

В первом случае F будет считаться так:

$$F_{user}(x) = isApproved(x) or \sum_{y \in M_{adjacents}} p_y / n_{notnullable} > 0.5$$

$$F_{publicpage}(x) = textSimilarity(x) or \sum_{y \in M_{adjacents}} p_y / n_{notnullable} > 0.5$$

Во втором, как линейная комбинация:

$$F_{group}(x) = (k_1 * isApproved(x) + k_2 * \sum_{y \in M_{adjacents}} p_y / n_{notnullable}) / (k_1 + k_2)$$

$$F_{group}(x) = (k_3 * textSimilarity(x) + k_4 * \sum_{y \in M_{adjacents}} p_y / n_{notnullable}) / (k_3 + k_4)$$

Так изначально имея набор групп с размеченной характеристикой принадлежности к подгруппе получаем гораздо большую выборку.

Процесс пересчета стоит продолжать до некоего E. В нашем случае было взято 0.2 %. При использовании упрощенной модели, вычисляется норма Фробениуса первого рода.

Норма Фробениуса первого рода получает максимальную разницу в матрицах.

В результате такого подхода вероятности для групп могут измениться относительно изначальных. Путем сравнения с изначальными данными можно проверить качество модели. На каждом шагу увеличения выборки проверяем изначальные данные.

Далее идет обход графа социальных связей описанного в прошлой главе. На каждом этапе добавляются новые подписчики, только что добавленных узлов графа и подписки, если они у узла есть.

3.4.2. Текстовая схожесть

Пользователи и публичные страницы создают и публикуют текстовые сообщения. У пользователей они как правило закрыты правилами приватности от большинства пользователей. У публичных страниц публикации как правило открыты.

В описываемом алгоритме текстовые сообщения являются одним из важнейших типов данных.

Используя их предлагается получать информацию о том, схожи ли сообщества своим содержанием или нет.

Выше были описаны несколько подходов к определению тематики текста. Предлагается воспользоваться двумя из них: тривиальным подходом и латено-семантическим анализом.

Известно, что группа футбольных фанатов имеет набор сленговых фраз, например: «бомжи» — фанаты футбольного клуба Зенит или же жители Санкт-Петербурга, «свиньи» — фанаты футбольного клуба Спартак[2].

Как описано выше, тривиальный алгоритм, предполагает составление словаря. Предлагается для составления словаря фраз относящихся к подгруппе радикальных фанатов использовать словарь с сайта wikipedia.com из раздела «Сленг футбольных фанатов».

Так же для определения текстовой схожести предлагается воспользоваться латенто-семантическим анализом, а именно алгоритм LSI. В ходе разметки обучающей выборки, появился набор текстов относящихся к радикальным сообществам и относящихся к обычным фанатским сообществам. При использовании LSI, предлагается использоваться сингулярное разложение на две темы.

[2]https://ru.wiktionary.org/wiki/Приложение:Сленг_футбольных_хулиганов

Предварительно предлагается отбросить термины, встречающиеся реже двух раз. Это позволит отбросить опечатки.

В результате чего получаем набор терминов относящихся к радикальным сообществам.

В качестве классификатора используем метод опорных векторов.

3.5. Описание программной реализации

Кратко опишем реализацию применяемого подхода, а так же использованные библиотеки и средства.

В первую очередь стоит сказать, что подход был реализован на языке Python 3.

Для сбора данных с сайта Vk.com использовалась библиотека Vk[3]. Социальная сеть жестко ограничивает частые запросы в свое API. Поэтому для корректного сбора больших данных с Vk.com, пришлось сделать обертку над указанной библиотекой, которая позволила максимально эффективно собирать данные. Одной из особенностей используемого API является разделение запросов на два типа: требующие токена, и не требующие.

Для методов, не требующих токены, был реализовано два подхода сбора данных. К таким методам например относится метод $groups.getMembers$[4]. Использовать такие методы можно, делая таймауты только при получении ошибки. Первая реализация — ленивая, позволяющая отправлять запросы на сервера, только непосредственно при вычислениях и обычная — более удобная в ходе эксперимента, в котором, не было необходимости делать вычисления на лету.

Все необходимые методы, можно было использовать без токена. Добавление токена в редких случаях давало дополнительные данные в некоторых публичных страницах. Из-за слабого улучшения и возможных проблем с кармой, было принято решение не использовать токены.

Для хранения данных, собранных из социальной сети, использовалась ORM Peewee[5], данные хранились в реляционной базе данных PostgreSql[6].

Граф социальных связей был отдельно реализован так же на языке Python 3.

[3] https://pypi.python.org/pypi/vk

[4] https://new.vk.com/dev/groups.getMembers

[5] http://docs.peewee-orm.com/en/latest/

[6] https://www.postgresql.org/

Для стемминга слов использовалось программное средство MyStem[7]. Проверки текста на вхождение символов, не принадлежащих к алфавиту, и слов, встречающихся реже, чем один раз, были одельно реализованы.

Для работы с матрицами использовалась библиотека numpy[8].

Для обучения использовалась библиотека sklearn[9].

Обход графа социальных связей, его модификации и прочее были отдельно реализованы.

3.6. Выводы по главе 3

В настоящей главе описано применение алгоритмов описанных в прошлой главе. Показаны несколько подходов к решению каждой из подзадач. Показаны мотивации их применения и так же предполагаемые результаты.

[7] https://tech.yandex.ru/mystem/?ncrnd=325

[8] https://pypi.python.org/pypi/numpy

[9] http://scikit-learn.org/stable/

ГЛАВА 4. РЕЗУЛЬТАТЫ

В данной главе описаны результаты эксперимента проведенного в рамках исследования для апробирования подхода.

4.1. Способы измерения качества результата

Для оценки качества результатов используется несколько методов оценки.

Важным критерием является корректное определение групп из обучающей выборки после нескольких этапов добавления новых узлов в граф социальных связей.

Так же интересен результат для пользователей определенных как администраторы сообществ. До начала эксперимента считается, что они буду принадлежать к той же группе, что и сообщество.

Основным методом оценки был выбран метод кросс-валидации. Выбранные группы делились на 10 частей, 9 групп входили в обучающую выборку. Оставшаяся часть использовалась для тестирования. Процедура повторялась 10 раз, для каждой части.

Алгоритм работает в несколько шагов, так что целесообразно проверять качество работы на разных этапах.

4.2. Оценка результатов

В данном разделе приведены результаты применения различных вариаций алгоритмов на группе футбольных болельщиков и радикальных футбольных фанатов.

Так как в ходе исследования не было обнаружено аналогично поставленных задача, результаты подхода сравним с результатами тривиальных алгоритмов.

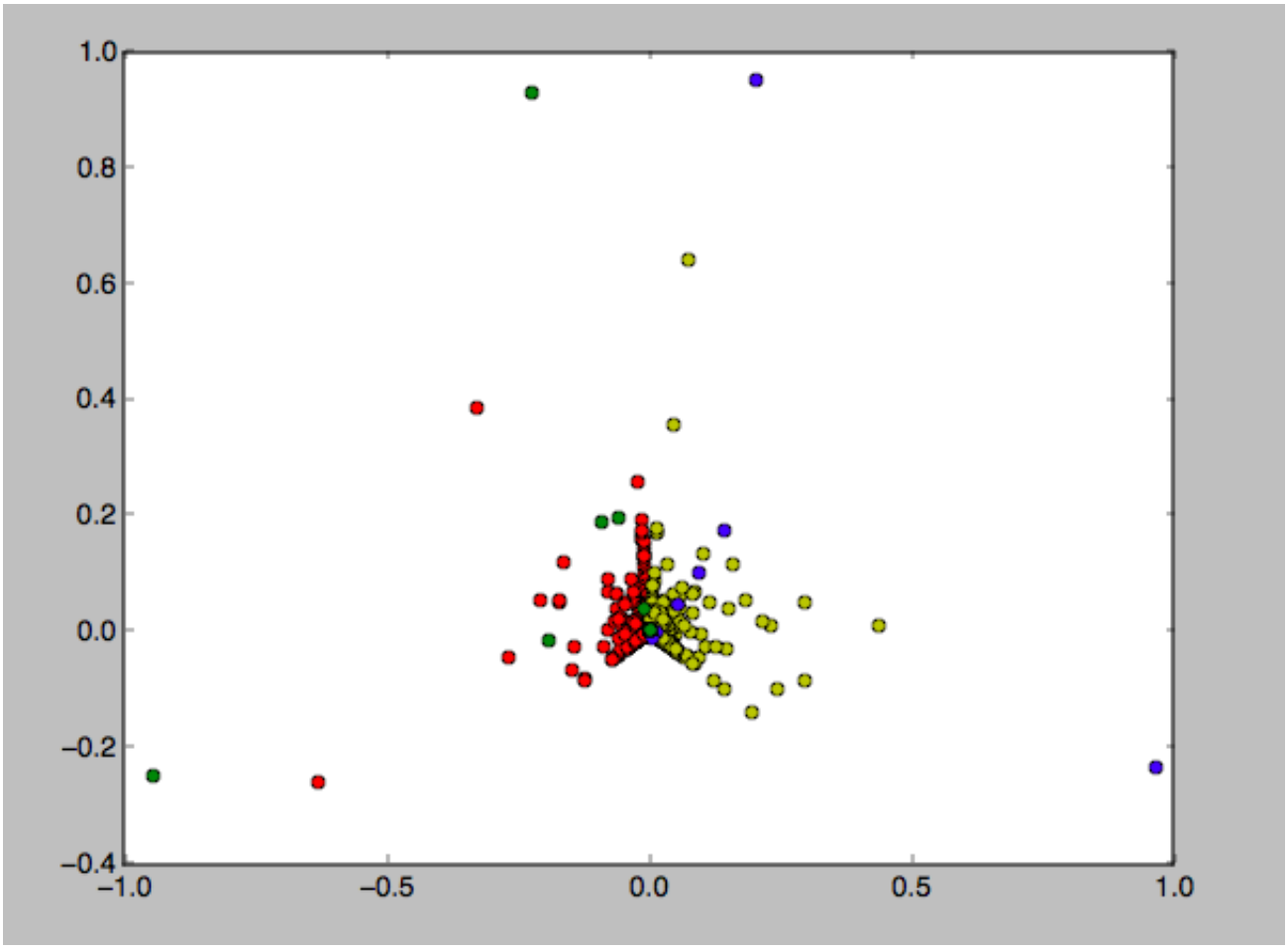
4.2.1. Результаты тривиальной оценки схожести текстов

Данный метод казался перспективным, за счет того, что подгруппы пользователей взятые в эксперимент обладали собственным сленгом. Однако оказалось, что применение сленга довольно популярно и термины из словаря сленговых выражений встречались как и в подгруппе, так и в группе. Что не позволило использовать данный метод. Средняя оценка точности не превышала 60 %. Поэтому подробное рассмотрение результатов этого метода не целесообразно.

Однако используя латенто-семантический анализ удалось добиться хорошего разделения текста по тематикам. Так на изображении 5 желтыми точками обозначены термины из группы обычных фанатов, красные из радикальных, синие — это точки фанатские группы, зеленые — радикальные фанатские группы.

На изображении данные для шести радикальных групп и шести нерадикальных.

Рисунок 5 – Разделение на тематики



4.2.2. Тривиальные подходы

Подход всегда определяющий узел как нерадикальный. При использовании подобного подхода, точность вычислений становится 95 %.

$$Precision = 0$$

$$Recall = 0$$

$$F_1 = 0$$

Подход определяющий узел как радикальный с вероятность 10 к 211. Обладает точностью 91 %.

$$Precision = 1/21$$

$$Recall = 10/211$$

$$F_1 = 0.05$$

Подход определяющий узел как радикальный с вероятность 50 %. Обладает точностью 50 %.

$$Precision = 1/21$$

$$Recall = 0.5$$

$$F_1 = 0.09$$

4.2.3. Результат метода оценки схожести основанного на операторе или

Кросс-валидация дала следующие результаты 1:

Таблица 1 – Результат кросс-валидации

Номер выборки	1	2	3	4	5	6	7	8	9	10
Точность	66%	66%	71%	62%	57%	57%	62%	57%	66%	62%

Средняя точность 63 %. Данный метод показал плохой результат, возможно это связано с частым ложно положительным результатом, обоснованным слабой функцией схожести. Однако он не является достаточно показательным. Важным критерием является возможность определять членов подгруппы. Так как за счет не сбалансированности размеров классов, алгоритм, возвращающий наиболее часто встречающийся результат будет давать точность лучшую с увеличением класса. Точность же определения подгруппы будет 0.

Посчитаем f_1 меру 2.

$$f_1 = 2 * precision * recall/(precision + recall)$$

Таблица 2 – Результат кросс-валидации

Номер выборки	1	2	3	4	5	6	7	8	9	10
Recall	1	1	1	1	1	1	1	1	1	1
Precision	1/7	1/7	1/6	1/8	1/9	1/9	1/8	1/9	1/7	1/8
F1	1/4	1/4	2/7	2/9	1/5	1/5	2/9	1/5	1/4	2/9

Алгоритм показывает очень сильный recall, однако из-за обилия ложно положительных результатов, precision очень низкий. F-мера довольно сильно превышает такую же меру для у тривиальных алгоритмов.

Что касается сохранения точности при расширении выборки, при использовании данного метода размеченные в ручную публичные страницы никогда не помечались с ошибкой.

4.2.4. Результат метода оценки схожести основанного на линейной комбинации

Эксперименты проводились с несколькими значениями параметров. Однако были выбраны $k = 1$, $k = 1$, $k = 5$, кросс-валидация дала следующие результаты 3:

Таблица 3 – Результат кросс валидации

Номер выборки	1	2	3	4	5	6	7	8	9	10
Точность	90%	81%	85%	81%	76%	90%	85%	90%	85%	76%

Средняя точность 84 %. По этой характеристике результат по прежнему слабый. Всегда определять узел, как не относящийся к подгруппе, выгоднее.

Таблица 4 – Результат кросс валидации

Номер выборки	1	2	3	4	5	6	7	8	9	10
Recall	1	1	1	1	1	1	1	1	1	0
Precision	1/2	1/4	1/3	1/4	1/5	1/2	1/3	1/2	1/3	0
F1	2/3	2/5	1/2	2/5	1/3	2/3	1/2	2/3	1/2	0

f мера показывает хороший результат 4, превышающий результаты тривиальных методов и подхода, основанного на операторе или, но в одном из номеров выборки она дала ложно отрицательный выбор. Это связано с тем, что появилась возможность поступательного уменьшения характеристики.

Выводы по главе 4

В данной главе был описан эксперимент, проведенный в рамках исследования для апробации описываемого подхода.

На примере задачи определения подгрупп радикальных футбольных фанатов из группы футбольных болельщиков была показана состоятельность данного подхода.

Выяснилось, что тривиальная оценка схожести может работать на определенных данных, однако имеет большую сложность в связи с необходимость составления правильного словаря для подгруппы. При неточном его составлении, качестве результатов сильно падает. Исследуемая модель показывает рост точности с использованием большего числа шагов, а значит большего числа узлов графа.

Использование латенто-семантического анализа дало возможность улучшить результаты.

Предложенный подход с использованием оценки схожести, основанной на операторе или дал много ложно положительных определений членов подгруппы, однако показал абсолютную полноту.

При использовании оценки схожести основанной на линейной комбинации получается максимальная точность.

Предположение о том, что администраторы публичных страниц будут принадлежать подгруппе своих страниц, не подтвердилось.

Предложенный подход позволяет достигнуть результатов на порядок лучших чем при случайном выборке. Данный подход позволяет при наличии малой обучающей выборки получать приемлемый результат.

ЗАКЛЮЧЕНИЕ

В данной работе был продемонстрирован подход, позволяющий определять принадлежат ли публичные сообщества и пользователи к подгруппе групп пользователей интернет-ресурсов, имеющих социальную составляющую.

В ходе эксперименты выделялась подгруппа радикальных фанатов из группы болельщиков футбольного клуба Зенит. Результаты показали, что наилучшего результата удалось добиться с использованием подхода, использующего латентно-семантический анализ и линейную комбинацию признаков.

Стоит отметить, что предложенный подход выделения подгруппы пользователей, может использовать не представленные в исследовании признаки схожести. Кроме текстовых данных могут использоваться данные о геолокации, данные из других социальных сетей, полученный по идентификатору указанному в профиле, данные из медиа контента, опубликованного пользователями и сообществами.

Метод может быть применим к любым видам групп и подгрупп. Подобная универсальность не гарантирует точности для любых видов групп, т.к. некоторые группы могут быть асоциальны и вовсе иметь нестандартную модель социальных связей. Однако описанный подход может дать хорошие результаты для многих из них. С другой стороны используя метод основанный на линейной комбинации можно улучшить результаты, пересчитывая на обучающих данных коэффициенты признаков.

Предложенные признаки дают высокую полноту, а при применении других алгоритмов семантического анализа или других более точных алгоритмов получения тематики текста, возможно смогут дать более точное выделение подгруппы.

К сожалению не удалось сравнить результаты с другими подобными исследованиями, так как схожих постановок задачи не было обнаружено.

В числе недостатков предлагаемого метода: для описанных в примере признаков, качество результатов может сильно ухудшаться для некоторых видов подгрупп. Эта проблема решается введением новых признаков схожести.

В дальнейшем качество данного подхода можно улучшить добавлением дополнительных признаков, таких как например геолокация. Использование других модификаций случайных марковских полей так же может улучшить

результат. Так же усовершенствования существующих признаков тоже может улучшить результаты.

Интересен результат для большего числа шагов увеличения выборки.

Развитием данного подхода может послужить апробирование представленного метода на других группах и подгруппах.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators / N. Ramakrishnan [и др.] // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2014. — С. 1799—1808.

2 *Swearingen K.*, *Sinha R.* Beyond algorithms: An HCI perspective on recommender systems // ACM SIGIR 2001 Workshop on Recommender Systems. Т. 13. — Citeseer. 2001. — С. 1—11.

3 *Grothoff C.*, *Porup J.* The NSA's SKYNET program may be killing thousands of innocent people // HAL Inria. — 2016.

4 *Blachnio A.*, *Przepiorka A.*, *D´az-Morales* J. F. Facebook use and chronotype: Results of a cross-sectional study // Chronobiology international. — 2015. — Т. 32, № 9. — С. 1315—1319.

5 Personality, gender, and age in the language of social media: The open-vocabulary approach / H. A. Schwartz [и др.] // PloS one. — 2013. — Т. 8, № 9. — e73791.

6 Определение демографических атрибутов пользователей микроблогов / Д. Турдаков [и др.] // Труды Института системного программирования РАН. — 2013. — Т. 25. — С. 179—192.

7 *Peersman C.*, *Daelemans W.*, *Van Vaerenbergh L.* Predicting age and gender in online social networks // Proceedings of the 3rd international workshop on Search and mining user-generated contents. — ACM. 2011. — С. 37—44.

8 *Trusov M.*, *Bodapati A. V.*, *Bucklin R. E.* Determining influential users in internet social networks // Journal of Marketing Research. — 2010. — Т. 47, № 4. — С. 643—658.

9 A 61-million-person experiment in social influence and political mobilization / R. M. Bond [и др.] // Nature. — 2012. — Т. 489, № 7415. — С. 295—298.

10 *Коршунов А.* Задачи и методы определения атрибутов пользователей социальных сетей // Труды. — 2013.

11 Tweeting From Left to Right Is Online Political Communication More Than an Echo Chamber? / P. Barberá [и др.] // Psychological science. — 2015. — С. 0956797615594620.

12 *Yardi S.*, *Boyd D.* Dynamic debates: An analysis of group polarization over time on twitter // Bulletin of Science, Technology & Society. — 2010. — Т. 30, № 5. — С. 316—327.

13 *Lo J.*, *Proksch S.-O.*, *Gschwend T.* A common left-right scale for voters and parties in Europe // Political Analysis. — 2014. — Т. 22, № 2. — С. 205—223.

14 *Bonica A.* Ideology and interests in the political marketplace // American Journal of Political Science. — 2013. — Т. 57, № 2. — С. 294—311.

15 *Gruzd A.*, *Roy J.* Investigating political polarization on Twitter: A Canadian perspective // Policy & Internet. — 2014. — Т. 6, № 1. — С. 28—45.

16 *Назар Б.*, *Коршунов А.* Выявление пересекающихся сообществ в социальных сетях // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов»-АИСТ. — 2012.

17 *Newman M. E.* Finding community structure in networks using the eigenvectors of matrices // Physical review E. — 2006. — Т. 74, № 3. — С. 036104.

18 Extending modularity definition for directed graphs with overlapping communities: тех. отч. / V. Nicosia [и др.]. — 2008.

19 *Rosvall M.*, *Bergstrom C. T.* Maps of random walks on complex networks reveal community structure // Proceedings of the National Academy of Sciences. — 2008. — Т. 105, № 4. — С. 1118—1123.

20 *Friggeri A.*, *Chelius G.*, *Fleury E.* Egomunities, exploring socially cohesive person-based communities // arXiv preprint arXiv:1102.2623. — 2011.

21 *Golbeck J.*, *Robles C.*, *Turner K.* Predicting personality with social media // CHI'11 Extended Abstracts on Human Factors in Computing Systems. — ACM. 2011. — С. 253—262.

22 Knowing the tweeters: Deriving sociologically relevant demographics from Twitter / L. Sloan [и др.] // Sociological research online. — 2013. — Т. 18, № 3. — С. 7.

23 *Pennebaker J. W.* Your use of pronouns reveals your personality. // Harvard business review. — 2011. — Т. 89, № 12. — С. 32.

24 Harvesting multiple sources for user profile learning: a big data study / A. Farseev [и др.] // Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. — ACM. 2015. — С. 235—242.

25 *Zheleva E.*, *Getoor L.* To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles // Proceedings of the 18th international conference on World wide web. — ACM. 2009. — С. 531—540.

26 *Baluja S.*, *Rowley H. A.* Boosting sex identification performance // International Journal of computer vision. — 2007. — Т. 71, № 1. — С. 111—119.

27 *Wu M.-J.*, *Jang J.-S. R.*, *Lu C.-H.* Gender Identification and Age Estimation of Users Based on Music Metadata. // ISMIR. — 2014. — С. 555—560.

28 *Liu J.-Y.*, *Yang Y.-H.* Inferring personal traits from music listening history // Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. — ACM. 2012. — С. 31—36.

29 Matching Entities Across Online Social Networks / O. Peled [и др.] // arXiv preprint arXiv:1410.6717. — 2014.

30 Introduction to information retrieval. Т. 1 / С. D. Manning, P. Raghavan, H. Schütze [и др.]. — Cambridge university press Cambridge, 2008.

31 *Chemudugunta C.*, *Steyvers P. S. M.* Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model // Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. Т. 19. — MIT Press. 2007. — С. 241.

32 *Blei D. M.*, *Ng A. Y.*, *Jordan M. I.* Latent dirichlet allocation // the Journal of machine Learning research. — 2003. — Т. 3. — С. 993—1022.

33 Indexing by latent semantic analysis / S. Deerwester [и др.] // Journal of the American society for information science. — 1990. — Т. 41, № 6. — С. 391.

34 *Kindermann R.*, *Snell J. L.* [и др.] Markov random fields and their applications. Т. 1. — American Mathematical Society Providence, RI, 1980.

35 *Li S. Z.* Markov random field modeling in image analysis. — Springer Science & Business Media, 2009.

36 *Романенко А. А.* Применение условных случайных полей в задачах обработки текстов на естественном языке // Москва. — 2014.

37 *Lafferty J.*, *McCallum A.*, *Pereira F. C.* Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

38 *Антонова А.*, *Соловьев А.* Метод условных случайных полей в задачах обработки русскоязычных текстов.

39 *Ланкастер П.*, *Демушкин С. П.* Теория матриц. — Издательство"Наука Главная редакция физико-математической литературы, 1978.