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 социльных медиа по их интересам и поведению на основе множества истоников 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 социльных медиа по их интересам и поведению на основе множества истоников данных
Научный руководитель Фильченков А.А., кандидат. техн. наук, доцент
Консультант

КРАТКОЕ СОДЕРЖАНИЕ МАГИСТЕРСКОЙ ДИССЕРТАЦИИ И ОСНОВНЫЕ ВЫВОДЫ

объем 21 стр., **графический материал** – стр., **библиография** 0 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, а так же группа феминисток и подгруппа радикальных феминисток. Были использованы данные пользователей из социальной сети Vk.com. Достигнуты следущие показатели для группы футбольных болельщиков: ; для группы феминисток: . Данный подход нов и так же может применятся для других групп и подгрупп пользователей.

Экономическая часть (какие использованы методики, экономическая эффективность результатов)

Данная работа не прополагает извлечения экономической выгоды из полученных результатов

Новизна полученных результатов

В рамках описываемого исследования представлен подход позволяющий опредлять 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. //СПИСОК-2016//... потом дописать

Практическая ценность работы. Рекомендации по внедрению

Полученный алгоритм дает возможность определить является ли член выбранной группы так же членом её подможества. 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 брэнда из группы его покупаетелей.

Выпускник _____

Научный руководитель _____

«____» _____ 2016 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Последнее время социальные медиа набрали огромную популярность. Такие сайты как Facebook.com, Vk.com, Twitter.com, 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. Однако содержание этой информации в основном остается не исползованным. 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, таки 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 полезна как бизнесу, так и государству. Используя восстановленные характеристики, можно уточнять таргетированную рекламу. Имея дополнительные данные об увлечениях людей, можно определять возможных преступников, что позволит предотвращать возможные нарушения или же прогнозировать конфликты.

Существует множество исследований о восстановлении данных, явно неуказанных в профилях пользователей. В них показано, что на основе информации о пользователе, его поведении в социльном медиа, можно с высокой точностью восстановить некоторые характеристики. Отдельной задачей стоит определения принадлежности ползователя к определенной группе, такой как, например, группа консерваторов или же группа любителей продукции Apple. Для решения этой задачи часто используют данные о социальных связях пользователей. Показано, что они влияют на поведение человека, на его взгляды.

В описываемом исследовании представлен подход для выделения подгруппы пользователей из определенной группы пользователей. Работа предло-

женного алогритма продемонстрирована на нескольких примерах: выделения подгруппы радикальных футбольных фанатов из группы футбольных болельщиков и выделения подгруппы радикальных феминисток из группы феминисток. 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, возникакающие 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, посвещенных 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 подразумеваюстя как социальные медиа, так и любые другие сайты, обладающие системой регистрации, так же данные могут собираться одновременно с нескольких ресурсов.

Проблема восстановления 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. Вычисление этой информации дает возможность улучшить качество таргетированных сервисов.

В социальных сетях зачастую указывается вовсе неверная информация, например, относительно возраста. Так появляется подтип задачи восстановления — определение ошибочных свойств пользователя и восстановление как неизсвестных.

Помимо широко используемых характеристик пользователя, некоторые исследования посвещены таким задачам, как определение хронотипов пользователей. 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. Подобные исследования дают новые пути исследований для психологов.

Описанные задачи зачастую сводятся к задачи класстеризации, регрессии или классификации.

Однако так же многие из них можно свести к задаче о выделении групп пользователей.

1.2. Задача о выделении групп пользователей

Задача о выделении групп пользователей является подтипом задачи о восстановлении характеристик пользователя. Она заключается в определеннии принадлежности выбранного пользователя к определенной группе, на основе имеющихся данных.

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

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

Часто задачи о выделении групп пользователей решаются путем класстеризации пользователей.

В основе всех исследований по выделению групп пользователей лежат данные, на основе которых происходит восстановление информации. В основном 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 т. д.

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

В прошлом разделе было отмечено, что основными проблемами выделения групп пользователей является приведение задачи к некой математической модели и генерация дискретных данных. Существующие решения можно условно разделить на несколько видов, основанных на виде решения проблемы. 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 в статье [**golbeck2011predicting**] применялся следующий подход. Для каждого пользователя собились все публичные характеристики его профиля.

Часто не вся информации оказывается необходимой для исследования. Поэтому часть признаков необходимо обозначить как менее информативные и удалить. Это ставит перед нами задачу определение информативности признаков. Например, в описываемой статье те данные, которые не отличались в зависимости от пользователя, те данные, которые были трудно представимы в численном виде или являлись слишком редко указываемыми характеристиками были удалены.

Так же нередко заполненных признаков пользователем бывает недостаточно, поэтому генерируются новые, например с помощью линейной регрессии [**golbeck2011predicting**].

Далее как правило решают задачу классификации или класстеризации. Где классы и кластеры соответуют 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 соответвий в исследуемом тексте. Такой подход обладает существенным недостатком, словари приходится составлять в ручную. Наглядным примером является задача определнию пола по имени[].

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

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

Другим методом приведения текста к численными данными является латентный семантический анализ[]. 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 связяй

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 определенный публичные страницы. Многие работы о выделении групп пользователей, как например в статье 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 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 оказатсья вовсе невозможным, из-за невозможности четкого определения известных членов группы.

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

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 определяющее гендерную принадлежность пользователя используя якость 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 аналиге фотографий пользователя часто анализируют мета-информацию файлов, как например в статье. Минус такого подхода заключается в возможности изменения этой мета-информации на не соответсвующую действительности.

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

Зачастую для определения принадлежности пользовталя к определнной группе используют данные о его геолокации. Для определенных типов групп, такой признак может хорошо работать. Так например в недавно рассекреченном проекте skynet по определению потенциальных террористов использовались в числе прочих данные о перемещении людей. Результатом работы алогритма являлась ложноположительное определение пользователя как террориста с вероятность менее 0.2 процента. Террористов алгоритм давал определять с вероятность в 50 процентов.

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

1.4. Постановка задачи настоящего исследования

Поставим задачу настоящего исследования следующим образом. Существует набор пользователей, публичных страниц, группа пользователей и подгруппа, которую необходимо выделить из существующей группы. Для каждой группы известны все её подписчики и все публичные сообщения созданные на этой страницей. Для каждого пользователя известны все его подписки, все его

друзья и все подписанные на него аккаунты. Так же доступна вся публичная информация о поведении пользователя, выраженная в одобрении определенного сообщения.

Имеющиеся данные можно представить в виде смешанного графа. Где узел это либо пользователь, либо публичная страница, с сопутсвующей информацией, а ребра между узлами есть отношение подписка-подписчик. Сопутствующая информация группы есть её публичные сообщения и список пользователей ододбривших 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, были описаны методы решений основанные на различных видах данных и различых моделях. Была описана задача, которая решается в данном исследовании.

ГЛАВА 2. ОПИСАНИЕ ИССЛЕДУЕМОГО ПОДХОДА

2.1. Общая схема решения

2.2. Граф социальных связей

2.3. Определение тематики публичных сообщений

2.3.1. Тривиальное решение задачи

2.3.2. Латентный сементический анализ

2.4. Случайные марковские поля

2.5. Модифицированные случайные марковские поля

2.6. Выводы по главе 2

ГЛАВА 3. РЕАЛИЗАЦИЯ ОПИСАВАЕМОГО ПОДХОДА

3.1. Сбор данных

3.2. Использование алгоритмов

3.3. Способы измерения качества результата

3.4. Результаты

3.5. Выводы по главе 3

ГЛАВА 4. ЗАКЛЮЧЕНИЕ

ГЛАВА 5. СПИСОК ИСПОЛЗОВАННЫХ ИСТОЧНИКОВ

В качестве примера таблицы приведена таблица 1.

Таблица 1 – Таблица умножения (фрагмент)

–	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
3	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51
4	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	64	68

Есть еще такое окружение `tabu`, его можно аккуратно растянуть на всю страницу. Приведем пример (таблица 2).

Таблица 2 – Таблица умножения с помощью `tabu` (фрагмент)

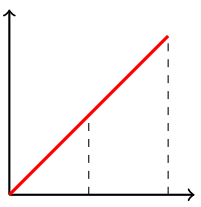
–	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
3	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51
4	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	64	68

пользователей различных социальных медиа. пользователей различных социальных медиа. пользователей различных социальных медиа.

5.1. Рисунки

Пример рисунка (с помощью `TikZ`) приведен на рисунке 1. Под `pdflatex` можно также использовать `*.jpg`, `*.png` и даже `*.pdf`, под `latex` можно использовать Metapost. Последний можно использовать и под `pdflatex`, для чего в стилевике продекларированы номера картинок от 1 до 20.

Рисунок 1 – Пример рисунка



5.2. Листинги

В работах студентов кафедры «Компьютерные технологии» часто встречаются различные листинги. Листинги бывают двух основных видов — исходный код и псевдокод. Первый оформляется с помощью окружения `lstlisting` из пакета `listings`, который уже включается в стилевике и немного настроен. Пример Hello World на Java приведен на листинге 1.

Listing 1 – Пример исходного кода на Java

```
public class HelloWorld {
    public static void main(String[] args) {
        System.out.println("Hello, world!");
    }
}
```

Псевдокод можно оформлять с помощью разных пакетов. В данном стилевике включается пакет `algorithmicx`. Сам по себе он не генерирует флоатов, поэтому для них используется пакет `algorithm`. Пример их совместного использования приведен на листинге 2. Обратите внимание, что флоаты разные, а нумерация — общая!

Листинг 2 – Пример псевдокода

function ISPRIME(N)
    for $t \leftarrow [2; \lfloor\sqrt{N}\rfloor]$ **do**
        if $N \bmod t = 0$ **then**
            return FALSE
        end if
    end for
    return TRUE
end function

Наконец, листинги из `listings` тоже можно подвешивать с помощью `algorithm`, пример на листинге 3.

Листинг 3 – Исходный код и флоат `algorithm`

```
public class HelloWorld {
    public static void main(String[] args) {
        System.out.println("Hello, world!");
    }
}
```

ГЛАВА 6. ПРОВЕРКА СКВОЗНОЙ НУМЕРАЦИИ

Листинг 4 должен иметь номер 4.

Листинг 4 – Исходный код и флоат `algorithm`

```
public class HelloWorld {
    public static void main(String[] args) {
        System.out.println("Hello, world!");
    }
}
```

Рисунок 2 должен иметь номер 2.

Рисунок 2 – Пример рисунка

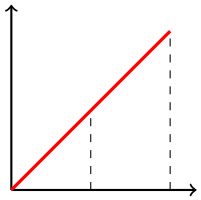


Таблица 3 должна иметь номер 3.

Таблица 3 – Таблица умножения с помощью `tabu` (фрагмент)

–	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
3	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51
4	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	64	68

Выводы по главе 6

В конце каждой главы желательно делать выводы. Вывод по данной главе — нумерация работает корректно, ура!

ЗАКЛЮЧЕНИЕ

В данном разделе размещается заключение.

ПРИЛОЖЕНИЕ А. ПРИМЕР ПРИЛОЖЕНИЯ

Пример ссылок на литературные источники: [**example-english**, **example-russian**].