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 технических наук, доцент
Консультант

КРАТКОЕ СОДЕРЖАНИЕ МАГИСТЕРСКОЙ ДИССЕРТАЦИИ И ОСНОВНЫЕ ВЫВОДЫ

объем 33 стр., **графический материал** – стр., **библиография** 28 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 следущие показатели для группы футбольных болельщиков: точность 84%, f-мера 0.46. 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//...

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 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 полезна 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 социльном 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, 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 [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 посвещены 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 задачи зачастую сводятся к задачи класстеризации, 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 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 [10—14]. В статье [10] 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 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. 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

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 [15] применялся следующий подход. Для каждого пользователя собирались 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 [15].

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 [**loan2013knowing**].

Так же популярна задача определения пола по текстовым сообщениям. В статье описывается, что женщины чаще используют личные местоимения, считая вхождения таких местоимений [16].

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 [17]. 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 определенный публичные страницы. Многие работы о выделении групп пользователей [8], как, например, в статье [10] 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 оказатсья вовсе невозможным, из-за невозможности четкого определения известных членов группы.

Существует работа в которой пользователь рассматривается как набор из всех его подписок [18]. 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 [19], 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 [20, 21]. На таких ресурсах как last.fm[1] используется такая информация как наиболее прослушиваемые композиции [20]. Так же применяется анализ самих аудиофайлов и отображения их в такие характеристики как ритмичность, скорость бита, и тому подобные [21]. 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 [2] 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] https://Last.fm

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

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 социальны 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.

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

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

В разделе 2.1 приведены основы текстового информационного поиска.

В разделе 2.2 описаны подоходы 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 является ли тема сообщения тематикой выделяемой подруппы пользователей.

Опишем термины используещиеся в дальнейшем повествовании [22].

Термин, он же слово, атомарная лингвистическая единиица.

Документ — конечный набор терминов. В контексте поставленной задачи, документом будет являеться 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 подохода удобно использовать двумерную матрицу, в которой каждый стоблец будет представлять вектор соответвующий документу.

Часто в задачах анализа тексты не используется порядок слов. Так получается "bag of words неупорядоченный набор терминов.

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

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

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 болле 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 соотвествующего термина в сооствующий документ. Данный подход по сути является некой модификацией описанного ранее алгоритма подхода термин-документ.

Как уже говорилось минус подобного подхода в ручном составлении словаря.

2.2.2. Латентно-сементический анализ

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

Формально задачу латентно-сментического 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 класстеризацию. Латентно-семантический анализ позовляет уменшить 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 алогритмам латентно-семантического анализа относят probabalistic latent semantic analysis(PLSA) [23] и так же latent Dirichlet allocation (LDA) [24].

Из неверотностных моделей следует рассказать о LSI, latent sematic indexing [25]. 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 интепртитации данных.

2.3. Случайные марковские поля

Случайные марковские поля (random Markov Fields) [26] метод широко применяемый в различных областях ИИ. Его успешно используют при распозновании речи и образов, а так же в обработке текста [**\IeC {\cyrr }\IeC {\cyro }\IeC {\cyrm }\IeC {\cyra }\IeC {\cyrn }\IeC {\cyre }\IeC** 27].

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) неорентированный граф, где каждая вершина является случайной переменной x и каждое ребро представляет зависимость между случайными велечинами 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 велечинами.

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 по-

лей(CRF) [**\IeC {\cyra }\IeC {\cyrn }\IeC {\cyrt }\IeC {\cyro }\IeC {\cyrn }\IeC {\cyro** 28].

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 [**\IeC {\cyrl }\IeC {\cyra }\IeC {\cyrn }\IeC {\cyrk }\IeC {\cyra }\IeC {\cyrs }\IeC {**]. 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 ОПИСАВАЕМОГО 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 анализиремой информации.

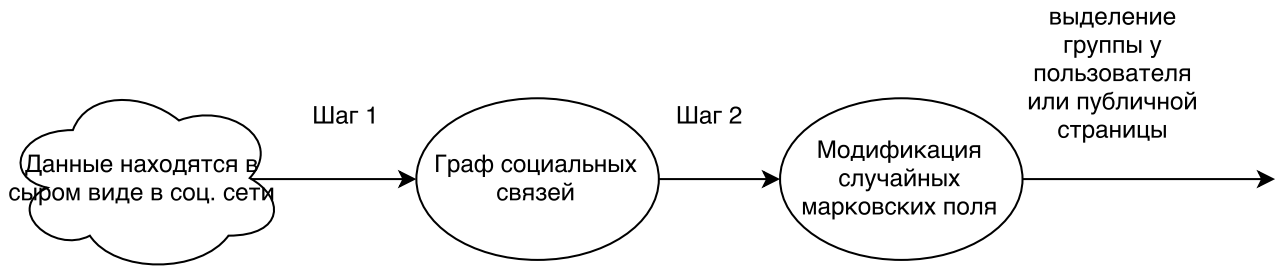
В разделе 3.3 описан собранный набор данных, который использовался в эксперименте.

В разделе 3.4 описано применение алгоритмов из главы 2.

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 её 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.

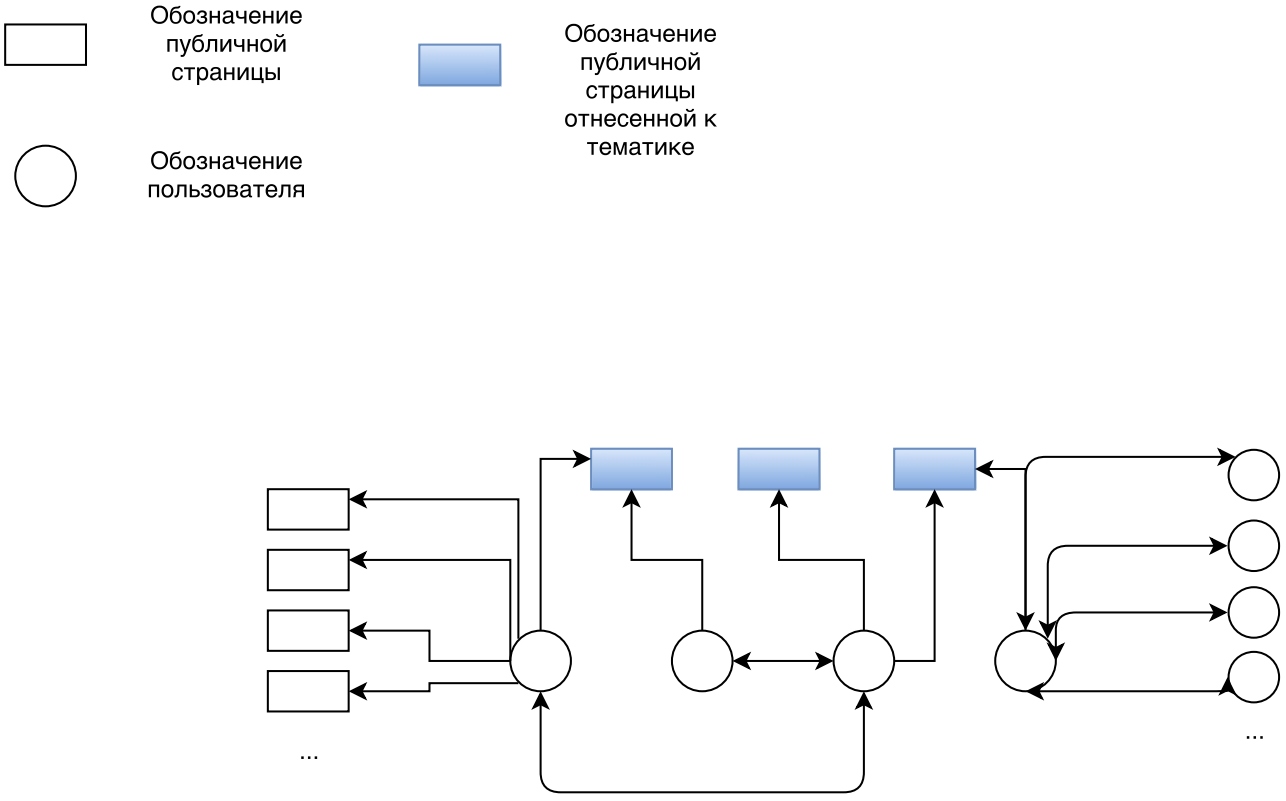
Собрав все необходимые данные строим по ним граф социальных связей с дополнительной информацией, подробнее в разделе 4.1 (шаг 1).

Преобразуем наш граф к модификации случайного марковского поля, используя допольнительную информцию, 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 2проиллюстрирована 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 сдедующую информцию: уникальный идентификатор, фамилия, имя, текущая геолокация. На рисунке 4 проиллюстрирована схема хранения пользователя.

Рисунок 3 – Узел публичной страницы в графе

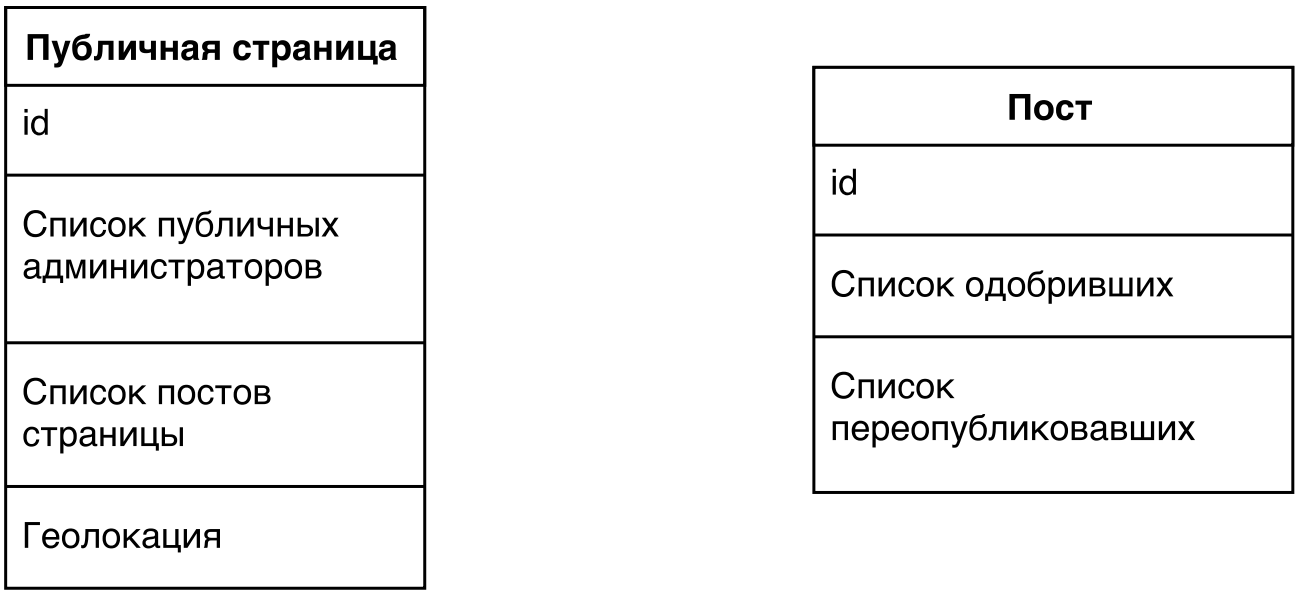


Рисунок 4 – Пользовательский узел в графе

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 посвещена 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 посвещенных футбольному клубу "Зенит". Для определения побличных 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 посвещенных этой тематике, 10 из которых были о футбольных хулиганах.

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 модифицрованных 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 воспользоватсья подходом схожим с предложеннным 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 модифкации 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 результатах 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 процента. При испльзовании 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 сравненения 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

Одной из основных характеристик публичной страницы является её текстовые сообщения. Для определения текстовой схожести предлагается воспользоваться латенто-семантическим анализом, а именно алгоритм LSI.

Данные о постах групп преобразуются к виду термин-документ.

Предварительно предлагает отбросить термины встречающиеся реже двух раз. Это позволит отбросить опечатки.

Использование алгоритма основано на предположении, о том, что публичные страницы принадлежащие к определенной подргруппе могут содержат тексты отличной тематики, так же зачастую они могут обладать сленговыми выражениями присущими исключительно данной подгруппе, например сленг радикальных футбольных фанатов. Исходя из этого предлагается опробовать и тривиальный метод.

3.5. 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.

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 обучающий выборки после нескольких кругов переобучения.

Так же интересен результат для пользорвателей определенных как администртаоры сообществ.

Основным методом оценки был выбран метод кросс-валидации. Выбранные группы делились на 10 частей, 9 групп входили в обучающую выборку. Оставшаяйся часть используется 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.

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 процентов. Поэтому подробное рассмотрение результатов этого метода не целесообразно.

4.2.2. Результат метода оценки схожести основанного на операторе или

Кросс валидация дала следующие результаты 1: Средняя точность 63 %. Данный метод показал плохой результат, возможно это связано с частым

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%

ложноположительным 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.

f_1 мера для случайного распределения, где вероятность посчитать узел принадлежащим к подгруппе 50 % : 1/11.

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 ложноположительных результатов, precision очень низкий. F-мера довольно сильно превышает такую же меру для случайного результата.

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

4.2.3. 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:

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

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%

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,но в одном из номеров выборки она дала ложноотрицательный выбор. Это связано с тем, что появилось возможность поступательного уменьшения характеристики.

4.3. Выводы по главе 4

На примере задачи определения подгрупп радикальных футбольных фатантов 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 праивльного 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.

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 обчающей выборки получать приемлемый результат.

Выводы по главе 4

ЗАКЛЮЧЕНИЕ

В данной работе был продемонстрирован подход, позволяющий выделять подгруппы групп пользователей интернет-ресурсов имеющих социальную составляющую.

Метод применим к разнообразным видам данных.

Так же он легко маштабируем и способен использовать соврешенно иные характеристики для уточнения результатов.

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

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

Полученный результат не имеет точного конкурирующего аналогам.

В дальнейшем качество данного подхода можно улучшить добавлением дополнительных признаков, таких как например геолокация. Использование других модифиакций случайных марковских полей так же может улучшить результат. Так же усоврешенствования существующих признаков тоже может улчшение результата.

Так же интересен результат для большего числа шагов увеличения выборки.

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

11 *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.

12 *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.

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

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

15 *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.

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

17 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.

18 *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.

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

20 *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.

21 *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.

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

23 *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. T. 19. — MIT Press. 2007. — C. 241.

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

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

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

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

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

ПРИЛОЖЕНИЕ А. ПРИМЕР ПРИЛОЖЕНИЯ

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