На правах рукописи

Гомзин Андрей Геннадьевич

Методы и программные средства определения значений стационарных демографических атрибутов пользователей социальных сетей

Специальность 05.13.11 —
«Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2021

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В. П. Иванникова РАН и на кафедре системного программирования факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова».

Научный руководитель: кандидат физико-математических наук
Турдаков Денис Юрьевич

Официальные оппоненты: **Кузнецов Сергей Олегович,**
доктор физико-математических наук,
Национальный исследовательский университет «Высшая школа экономик»,
руководитель департамента анализа данных и искусственного интеллекта

Губанов Дмитрий Алексеевич,
кандидат технических наук,
Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова РАН,
старший научный сотрудник

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

Защита состоится 10 июня 2021 г. в 14 часов на заседании диссертационного совета Д 002.087.01 при Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В. П. Иванникова РАН по адресу: 109004, г. Москва,ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Институте системного программирования им. В. П. Иванникова РАН.

Автореферат разослан «___» _____ 2021 года.

Ученый секретарь
диссертационного совета
Д 002.087.01,
кандидат физ.-мат. наук Зеленов С.В.

Общая характеристика работы

Актуальность темы.

В современном мире широко распространены такие способы коммуникации посредством сети Интернет, как *социальные медиа*: блоги, сайты знакомств, форумы, микроблоги, социальные сети. Особый интерес среди социальных медиа представляют социальные сети. *Социальная сеть* – платформа, онлайн-сервис и веб-сайт, предназначенные для построения, отражения и организации социальных взаимоотношений в Интернете. Основными элементами социальной сети являются публичные страницы, Они могут являться как персональными страницами пользователей, так и страницами, представляющими организации, тематические сообщества, события и т.д. Отношения между страницами представлены *социальными связями*. Примерами социальных связей являются дружба между пользователями, подписка на сообщества, события и т.д. Социальная сеть или её часть моделируется с помощью *социального графа*. Социальный граф состоит из вершин, представляющих страницы пользователей, сообществ, организаций и т.д., и рёбер, представляющих социальные связи между соответствующими вершинами.

Под *демографическими атрибутами* пользователей социальных сетей понимаются пол, возраст, семейное положение, уровень образования, род деятельности, трудоустроенность, место жительства, доход, политические, религиозные взгляды, интересы, национальность и другие. Множество значений демографических атрибутов пользователя составляют его *демографический профиль*. Множество явно указанных и публично доступных значений демографических атрибутов пользователя назовём *публичным профилем*. Не все значения указываются пользователями явно, поэтому лишь часть значений атрибутов могут быть определены с использованием публичного профиля. В связи с этим возникает задача предсказания неуказанных значений демографических атрибутов пользователей социальных медиа по доступным данным, таким как тексты публичных сообщений, социальный граф. Кроме того, некоторые пользователи преднамеренно указывают ложные данные. Отличие указанных в публичном профиле значений атрибутов от предсказанных на основе анализа поведения пользователя может служить признаком для определения ложных значений.

Для решения задачи предсказания значений демографических атрибутов необходимо специальное программное обеспечение, позволяющее собирать открытые данные из социальных сетей, применять к ним методы и модели с целью получения и восстановления демографических профилей пользователей, оценивать качество различным моделей и методов с использованием различных наборов данных. Программное обеспечение,

позволяющее восстановить демографические профили пользователей, являются базовым и необходимым инструментом при решении различных прикладных задач. Так, например, значения демографических атрибутов пользователей могут использоваться коммерческими компаниями для определения целевой аудитории предлагаемых продуктов, а также для поиска потенциальных клиентов в социальных медиа. Организации могут использовать демографические профили пользователей для поиска потенциальных сотрудников с целью найма. Значения демографических атрибутов также могут быть полезными и для таких задач государственного управления, как изучение современных демографических тенденций, оценка переизбытка или нехватки специалистов в различных областях.

В диссертационная работе исследуются и разрабатываются методы и программные средства для предсказания значений *стационарных* демографические атрибутов, то есть таких, которые редко меняются и актуальны на протяжении жизни пользователей. Такими атрибутами являются пол (меняется крайне редко), год рождения (не меняется), семейное положение (в среднем меняется 1-2 раза), уровень образования (меняется по уровням, 1-2 раза), род деятельности (в среднем не меняется для взрослого человека). Методы предсказания таких атрибутов, как интересы, отношение к определённым событиям, не рассматриваются в рамках данной работы. Также в работе не рассматривается вопрос о зависимости между различными атрибутами, задача предсказания ставится независимо для каждого стационарного атрибута.

На практике одной из частых постановок задач определения значений демографических атрибутов для заданного множества целевых пользователей. Это множество может представлять собой как некоторое сообщество (студенты университета, подписчики сообщества), так и всех пользователей социальной сети. Недоступные из публичных профилей значения демографических атрибутов можно предсказывать с использованием соответствующих методов и программных средств по другим доступным данным, например, по текстам публичных сообщений пользователей. В применении к обозначенной задаче методы и программные средства для предсказания значений атрибутов по текстам имеют ряд недостатков, связанных с недоступностью, разнородностью и затруднённым сбором текстовых данных для заданного множества пользователей. Социальные связи являются более доступным источником публичных данных о пользователях. Поэтому в диссертационной работе особое внимание уделено методам, моделям и программным средствам для предсказания значений демографических атрибутов с использованием социального графа.

Существующие методы предсказания значений атрибутов по социальному графу обладают недостаточным качеством, что показывается примерами, где эти значения предсказываются неверно. В общем случае задача предсказания значений атрибутов сводится к задаче классификации

или регрессии, поэтому под качеством методов понимаются традиционные для задач классификации и регрессии метрики: F-1 мера с микро- и макроусреднением, среднеквадратичная ошибка (MAE), коэффициент детерминации (R2).

Целью диссертационной работы является разработка методов и программных средств для определения значений стационарных демографических атрибутов пользователей социальных сетей. Разработанные методы должны превосходить по качеству предсказания существующие методы при доступности информации только о социальном графе.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать существующие методы определения демографических атрибутов пользователей;
2. Разработать и реализовать методы предсказания значений демографических атрибутов по социальному графу, превосходящие по качеству существующие методы;
3. Провести экспериментальное сравнение разработанных методов с существующими методами с использованием общепринятых метрик качества;
4. Разработать программную систему для определения значений стационарных демографических атрибутов пользователей социальной сети.

Научная новизна: Разработаны методы предсказания значений демографических атрибутов пользователей, основанные на введённом в диссертационной работе свойстве вершин социального графа, *специфичности контекста* для заданного атрибута. Разработанные методы показывают более высокое качество предсказания по сравнению с методами, не использующими специфичность контекста.

Теоретическая и практическая значимость заключается в использовании разработанных методов в Talisman, комплексе взаимосвязанных программных инструментов для автоматизации типовых задач обработки данных, включая их сбор, интеграцию, анализ, хранение и визуализацию. Результаты работы были применены при выполнении работ по договору с Министерством образования и науки Российской Федерации №14.514.11.4111 «Построение социо-демографического профиля пользователей сети Интернет». Разработанные методы позволят повысить эффективность решения прикладных задач, использующих значения демографических атрибутов пользователей.

Личный вклад. Все выносимые на защиту результаты получены лично автором.

Основные положения, выносимые на защиту:

1. Разработан подход для предсказания значений демографических атрибутов на основе специфичности контекста вершин социального графа;
2. В рамках подхода созданы новые методы предсказания значений демографических атрибутов по социальному графу *LP-CS*, *LP-CS-Gen*, *Distr2-CS+DW[n]-XGB*, *GConv-CS[n]*, *Distr2-CS-XGB*, превосходящие по качеству существующие аналоги; даны рекомендации по их применению;
3. Реализована программная система предсказания значений атрибутов пользователей социальных сетей по социальному графу, позволившая экспериментально подтвердить превосходство созданных методов над существующими аналогами по качеству решения задачи.

Достоверность полученных результатов обеспечивается проведенной экспериментальной проверкой возможности использования специфичности контекста для предсказания значений атрибутов, с использованием данных из реальных социальных сетей, а также экспериментальным сравнением разработанных методов с аналогичными методами определения демографических атрибутов, описанными в литературе, с использованием данных из реальных социальных сетей.

Апробация работы. Основные результаты диссертационной работы докладывались в рамках следующих мероприятий:

- Научный семинар отдела «Информационных систем», Москва, 2016 г.
- 190-е заседание Московской секции ACM SIGMOD, Москва, 2016 г.
- Семинар по социофизике имени Д.С.Чернавского, Москва, 2016 г.
- Международная открытая конференция ИСП РАН 2016, Москва, 2016 г.
- 24-я международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», Москва, 2018 г.
- Ломоносовские чтения, факультет ВМК МГУ им. М.В. Ломоносова, Москва, 2020 г.
- Международная конференция «55th Annual Conference on Information Sciences and Systems (CISS)», дистанционно, 2021 г.

Публикации. Автор имеет 12 публикаций в печатных изданиях, 2 работы индексируются в Scopus и Web of science. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 3 из которых изданы в журналах, рекомендованных ВАК, 1 — в тезисах докладов. Получено 5 свидетельств о регистрации программ для ЭВМ. Основная часть работы [1] выполнена автором, редакторские правки и анализ результатов экспериментов выполнялись совместно с соавторами. Основная часть работ [4], [3]

выполнена автором, редакторские правки выполнялись совместно с сооавторами. В работе [2] автором был собран набор данных, описание методов и анализ результатов был выполнен совместно с соавторами. Работа [5] полностью выполнена автором. В рамках программы [6] автором реализована часть методов сбора социальных графов. В Talisman [7] автором реализованы методы предсказания значений атрибутов пользователей социальных сетей. Большая часть программ на ЭВМ [8], [9] и [10] была реализована автором и использована для сбора данных и оценки качества работы методов.

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объем диссертации составляет **143** страницы текста, включая **52** рисунка и **15** таблиц. Список литературы содержит **105** наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, формулируется цель, ставятся задачи работы, перечисляются основные положения, выносимые на защиту, излагается научная новизна и практическая значимость представляемой работы.

Первая глава посвящена исследованию существующих методов определения значений демографических атрибутов пользователей социальных медиа. Проводится обзор и выделяются основные направления и подходы к определению демографических атрибутов.

Раздел 1.1 посвящён методам предсказания значений атрибутов по текстам авторов. Еще до появления социальных сетей проводились исследования текстов и их авторов. При этом анализировались такие тексты, как эссе студентов, личные дневники, сообщения электронной почты (e-mail). В этих работах проводились статистические исследования лингвистических особенностей текстов мужчин и женщин, а также людей с различными моделями личности.

В начале 2000х годов начали набирать популярность блоги. Одной из актуальных задач являлась задача предсказание неизвестных значений пола и возраста пользователя по текстам его авторства, полученных из сообщений электронной почты и постов в блогах. В эти годы активно разрабатывались методы предсказания значений атрибутов на основе машинного обучения с учителем. Типичными признаками, извлекаемыми из текстов блогов, являлись n-граммы.

В конце 2000х годов возросло количество пользователей микроблогов. Тексты сообщений, которые пользователи публикуют в микроблогах, обладают некоторыми существенными для анализа лингвистическими особенностями, такими как короткая длина, орфографические ошибки, специальные обозначения символы. Методы предсказания были адаптированы

к соответствующим условиям: решения включали в себя нормализацию текстов, выделение эмотиконов как отдельных признаков, использование дополнительных доступных данных, например, из профилей пользователей. Из-за увеличения словаря и уменьшения размера текстов, методы на основе машинного обучения сталкивались с проблемой переобучения. Для преодоления переобучения применялась предобработка признаков. При этом использовались как методы отбора наиболее информативных признаков, так и способы проецирования исходных признаков на пространство более низкой размерности, применяя, например, метод главных компонент. Во втором случае разреженные представления в пространстве высокой размерности проецировались на плотные представления в новом пространстве более низкой размерности.

В рамках диссертационной работы было проведено экспериментальное сравнение различных методов предсказания значений возраста и уровня образования по текстам публичных комментариев пользователей. Для этого был собран набор данных из социальной сети Вконтакте. В рамках экспериментального сравнения методов предсказания возрастного интервала и уровня образования с использованием этого набора данных развивающиеся в последнее время методы, основанные на нейронных сетях (например, LSTM) не показали преимущества над классическими методами (например, методом опорных векторов над n-граммами). По результатам исследования можно выделить ряд недостатков использования текстов комментариев для предсказания значений демографических атрибутов для заданного множества целевых пользователей. Таргетированный сбор текстов комментариев для заданного множества пользователей социальной сети часто затруднён, так как комментарии могут располагаться на различных публичных страницах и относиться к различным публикациям. Кроме того, не все пользователи активно пишут публичные комментарии. Как следствие, для многих пользователей не удаётся собрать достаточное количество текстов, что оказывает существенное влияние на качество методов предсказания значений атрибутов по текстам комментариев. Предсказание значений атрибутов с использованием социальных связей лишено обозначенных недостатков.

Раздел 1.2 посвящён методам предсказания значений атрибутов по социальным связям пользователей. Наиболее популярные решения задачи в такой постановке можно разделить на три подхода. Первый подход заключается в применении методов машинного обучения без учителя. Задача предсказания значений характеристик пользователей сводится к задаче кластеризации вершин в графе. Методы кластеризации графа разбивают множество вершин на кластеры таким образом, чтобы максимизировать количество связей (рёбер) внутри кластеров и минимизировать количество связей вершин из разных кластеров. В контексте задачи предсказания значений атрибутов предполагается, что пользователи объединяются в

кластеры согласно их демографическим характеристикам, т.е. кластер представляет собой множество пользователей с равными или похожими значениями рассматриваемо атрибута. Наиболее популярными методами кластеризации, являются алгоритмы распространения меток в графе. На каждом шаге алгоритма метка каждой вершины обновляется, используя метки соседних вершин. В нашем случае метка представляет собой значение рассматриваемого атрибута.

Второй подход заключается в применении методов машинного обучения с учителем. В основе подхода лежит построение векторных представлений вершин социального графа. Векторные представления строятся только с использованием информации о структуре графа, т.е. связях между вершинами. Будем называть такие представления статическими. Полученные векторы используются в качестве признаков для алгоритмов машинного обучения с учителем. Алгоритм обучается на этих признаках и значениях предсказываемого атрибута для тех пользователей, у которых это значение известно. Обученная модель используется для предсказания значений по признакам для пользователей с неизвестным значением атрибута.

В последнее время развиваются методы, основанные на использовании графовых нейронных сетей (GNN, англ. Graph Neural Networks). Графовые нейронные сети в применении к задаче классификации позволяют учитывать не только связи между вершинами, но и информацию о вершинах, например, векторные представления, извлеченные из текстов пользователей, а также известные для некоторых пользователей значения предсказываемого атрибута. Таким образом, векторные представления вершин графа, возникающие в слоях нейронной сети, подстраиваются под конкретную прикладную задачу, с использованием дополнительной доступной информации. Частным случаем графовых нейронных сетей, используемых для классификации, являются свёрточные графовые нейронные сети.

В разделе 1.3 обсуждаются и другие способы комбинирования данных различной природы (тексты и граф) для решения задачи предсказания значений демографических атрибутов, а также методы совместного предсказания значений нескольких атрибутов.

Раздел 1.4 посвящён методам и особенностям сбора открытых данных из социальных сетей. Рассматриваются вопросы извлечения референсных значений атрибутов, методы сэмплинга, описаны результаты экспериментального сравнения двух способов сбора социального графа для заданного множества целевых пользователей.

Методы предсказания значений демографических атрибутов по социальным связям не лишены недостатков, обозначенных в разделе 1.5. В частности, методы неверно работают в следующей ситуации. Допустим, пользователь, являющийся разработчиком ПО, подписан на две страницы крупных сообществ, например, на новостной источник и страницу,

посвящённую творчеству популярного артиста. Профессия подписчиков каждого из этих сообществ равномерно распределена, с небольшим преобладанием, например, водителей. Кроме того, пользователь подписан на одно небольшое тематическое сообщество, подписчики которого являются преимущественно разработчиками ПО. Алгоритм распространения меток, например, будет предполагать, что этот пользователь является водителем. Для алгоритма подписка пользователя на популярную страницу и подписка на узкоспециализированное профессиональное сообщество имеют одинаковую значимость. Методы, описываемые в диссертационной работе, учитывают более высокую значимость узкоспециализированных сообществ и показывают более высокое качество по сравнению с другими методами.

Таким образом, в первой главе проводится обзор методов определения значений демографических атрибутов пользователей социальных медиа, в том числе, социальных сетей, по текстам сообщений и социальному графу. Обозначаются недостатки существующих методов и подходов.

Во **второй главе** описывается новый подход к предсказанию значений демографических атрибутов. Для этого вводится понятие специфичности контекста вершины для заданного демографического атрибута, формулируется и экспериментально проверяется гипотеза о зависимости между специфичностью контекста и значениями атрибутов.

В разделах 2.1 и 2.2 вводятся обозначения и приводится формальная постановка задачи. Рассмотрим ненаправленный социальный граф без петель $G(V, E)$. Пусть $U \subseteq V$ – множество вершин, представляющих пользователей социальной сети. Будем их также называть вершинами-пользователями. Для некоторого подмножества вершин-пользователей $Y \subseteq U$ известны значения $y_x \in A$, $x \in Y$ рассматриваемого атрибута $\mathcal{A}$. Задача заключается в предсказании неизвестных значений атрибута $\mathcal{A}$ для вершин-пользователей $U \setminus Y$.

В разделе 2.3 описываются четыре набора данных, используемых в диссертационной работе. Два из них, *twitter* и *pokec*, представляют собой существующие наборы данных, используемые в других исследованиях. Еще два набора, *vk1*, *vk2* были собраны автором и представляют собой социальные графы социальной сети Вконтакте.

Набор данных twitter включает в себя социальный граф, полученный из социальной сети Twitter, и известные для некоторых вершин значения рода деятельности и дохода. Значения рода деятельности извлекались автоматически из поля «описание» профиля. Доход[1] определялся из рода деятельности. Социальный граф формировался следующим образом. Для каждой из размеченной вершины-пользователя собиралось множество подписок в социальной сети, соответствующие рёбра добавлялись в социальный граф.

[1]Количество различных значений дохода (53) в наборе данных twitter существенно меньше количества размеченных вершин (4625).

Набор данных pokec является снимком социальной сети Pokec. Он содержит все социальные связи между пользователями и профили пользователей Pokec, находящимися в открытом доступе. Профили содержат значения различных демографических атрибутов, включая пол и возраст[2]. Эти значения были указаны пользователями социальной сети в их публичных профилях.

В рамках работы было собрано два набора данных из социальной сети Вконтакте. Первый набор данных, vk1, состоит из пользователей с известными значениями рода деятельности и множеством соседей для этих пользователей. Множество размеченных пользователей ограничено студентами и выпускниками МГУ, значения рода деятельности размечались ручном режиме, при помощи аннотаторов. Дополнительно для размеченных пользователей были извлечены значения пола и возраста из профиля. Социальный граф собирался по множеству размеченных вершин-пользователей аналогично набору данных twitter.

Набор данных vk2 собирался на основе явно указанных значений пола и возраста в профиле. Множество размеченных пользователей было собрано с использованием алгоритма сэмплинга «Forest fire». Затем, аналогично наборам данных twitter и vk1, был собран социальный граф.

В разделе 2.4 определяется и анализируется свойство специфичности контекста и предлагается подход к предсказанию значений атрибутов на ей основе.

Пусть $N_x^1 = N_x = \{z \in V : (x, z) \in E\}$ – множество соседних вершин вершины x в графе $G(V, E)$. $A = (a_1, a_2, \ldots, a_{|A|})$ – возможные значения атрибута $\mathcal{A}$. Определим функцию d: $2^V \to \mathbb{R}^{|A|}$, которая для заданного множества вершин X определяет вектор, представляющий дискретное распределение по значениям атрибута:

$$\begin{cases} d(X)_i = \dfrac{\left|\{x \in X \cap Y : y_x = a_i\}\right|}{\left|X \cap Y\right|}, & \text{если } X \cap Y \neq \emptyset \\ d(X) \text{ не определено} & \text{иначе.} \end{cases} \tag{1}$$

Если $d(X)$ существует, то сумма его элементов равна 1, порядок элементов соответствует порядку в A. Функция задаёт дискретное распределение значений атрибута среди пользователей из заданного множества.

Теперь определим значение специфичности контекста $s(z)$ для заданной вершины z следующим образом:

$$s(z) = \begin{cases} KL\big(d(N_z)\big|\big|d(Y)\big) & \text{, если } \exists d(N_z) \\ 0 & \text{, иначе} \end{cases} \tag{2}$$

Здесь $KL(p||q) = \sum_i \log p_i \frac{p_i}{q_i}$ – дивергенция Кульбака–Лейблера.

[2]Значения возраста представлены с точностью до года.

Специфичность контекста $s(z)$ показывает, насколько распределение $d(N_z)$ значений $\mathcal{A}$ для соседей вершины z отличается от распределения генеральной совокупности $d(Y)$, т.е. распределения для всех размеченных вершин Y. Если распределение значений атрибута среди соседей близко к распределению генеральной совокупности, то значение специфичности вершины близко к 0. Чем больше расстояние между этими распределениями, тем больше значение специфичности контекста вершины.

Рассмотрим предположения «гомофилии» и зависимости между общим контекстом и значениями атрибутов, на основе которых строятся существующие методы. «Гомофилия» предполагает, что две вершины, соединённые ребром, чаще имеют одинаковые значения атрибутов, чем не соединённые. Предположение о зависимости между общим контекстом и значениями атрибутов заключается в том, что значения атрибута у двух вершин совпадают тем чаще, чем больше размер общего контекста, т.е. множества общих смежных вершин.

Предположение о зависимости между специфичностью общего контекста и значениями атрибутов, которое вводится в диссертационной работе, помимо размера контекста учитывает специфичность каждой из общих вершин. В работе проводится сравнительный анализ этих трёх свойств. Для каждой пары вершин (x,z) введём свойства h, c и cs и их значения.

Значение свойства h определим как *наличие ребра* между вершинами:

$$h(x,z) = \begin{cases} 1 & \text{, если } (x,z) \in E \\ 0 & \text{, иначе} \end{cases} \tag{3}$$

Значение свойства c определим как *размер общего контекста* вершин:

$$c(x,z) = |N_x \cap N_z| \tag{4}$$

Значение свойства cs определим как *специфичность общего контекста*:

$$cs(x,z) = \sum_{v \in N_x \cap N_z} s(v) \tag{5}$$

Специфичность пары вершин (x,z) является численной характеристикой общего контекста этих двух вершин, то есть множества вершин, связанных как с x, так и с y. Она зависит как от количества общих соседей, так и от специфичности контекста $s(v)$ каждой вершины их общих соседей $N_x \cap N_z$. Чем больше общих соседей с бóльшим значением $s(v)$, тем больше специфичность контекста для пары вершин $cs(x,z)$.

Значения h, c и cs вычисляются на наборах данных, описанных выше и затем анализируются. При вычислении значения округляются вниз до 0.001.

Пусть $\Omega = \{(x,z) : x,z \in Y, x \neq z\}$ – множество всех возможных пар размеченных вершин. Рассмотрим вероятностное пространство с множеством элементарных исходов Ω, сигма-алгеброй $\sigma = 2^{\Omega}$ и функций вероятности $\mathbb{P}(\omega) = \frac{1}{|\Omega|}, \forall \omega \in \Omega$.

Описанные выше функции h, c и cs далее будем рассматривать как случайные величины в вероятностном пространстве $(\Omega, \sigma, \mathbb{P})$. Множество размеченных вершин в наборе данных конечно, следовательно, количество различных значений, которые принимают рассматриваемые случайные величины также конечно. Для каждой из случайных величин h, c и cs упорядочим по возрастанию соответствующие множества значений. Получим три последовательности: $(h_i)_{i=1}^{N_h}$, $(c_i)_{i=1}^{N_c}$, $(cs_i)_{i=1}^{N_{cs}}$. Здесь N_h, N_c и N_{cs} – количество различных значений величин h, c и cs, соответственно, полученных из набора данных.

Пусть ξ – одна из случайных величин h, c или cs. Для каждого из значений, которое может принимать случайная величина ξ, рассмотрим множество вершин $\{(x,z) \in \Omega : \xi(x,z) \geq \xi_i\}$ и определим для него:

$$\alpha_i^{\xi} = P(\xi(x,z) \geq \xi_i), \text{ где } i = \overline{1,N_{\xi}} \tag{6}$$

Множество пар вершин, соответствующих значению α_i^{ξ}, обозначим как $B_{\xi}(\alpha_i^{\xi}) = \{(x,z) \in \Omega : \xi(x,z) \geq \xi_i\}$. Стоит отметить, что $\alpha_i^{\xi} = 1$, с увеличением ξ_i значение α_i^{ξ} уменьшается.

Теорема 1 Пусть диаметр графа $D(G) > 2, \exists\, v \in V : |N_v| > 1, s(v) > 0$. Тогда $h_1 = c_1 = cs_1 = 0$ и $\exists h_2 > 0, c_2 > 0, cs_2 > 0$; при этом α_2^h является плотностью графа; α_2^c равно вероятности того, что для случайной пары вершин существует путь длины 2 между ними.

Для всех рассматриваемых наборов данных выполняются условия теоремы 1: $D(G) > 2$, $\exists v \in V : |N_v| > 1, s(v) > 0$. Следовательно, для каждой из величин h, c, cs существует не менее двух различных значений. Таким образом, можно проводить анализ наборов данных на предмет связи этих величин со значениями атрибута вершин. Дальнейшие рассуждения предполагают выполнение условий теоремы 1.

Значению α_i^{ξ} соответствует множество пар $B_{\xi}(\alpha_i^{\xi}) \subset \Omega$. Определим способ получения множества пар вершин $B_{\xi}(\alpha)$, соответствующей произвольному значению $\alpha \in (\alpha_{i+1}^{\xi}, \alpha_i^{\xi})$. Для этого воспользуемся следующей интерполяцией. Рассмотрим множество $B_{\xi}(\alpha_{i+1}^{\xi})$ и дополним его случайными парами из множества $\{(x,z) \in \Omega : \xi(x,z) = \xi_i\}$. Количество дополнительных пар равно $\lfloor |\Omega|(\alpha - \alpha_{i+1}^{\xi}) \rfloor$.

Утверждение 1 Для значения α_1^{ξ} множество пар $B_{\xi}(\alpha_1^{\xi}) = \Omega$.

Полученные выборки использовались для оценки связи между значением свойств h, c, cs и совпадением значений атрибута у пары вершин. Кроме того, сравнивалось, насколько частота совпадения значений атрибута среди вершин с наибольшими значениями свойства cs выше, чем частота

среди вершин с наибольшими значениями h и c, при одинаковом размере выборок. Была рассмотрена нулевая гипотеза H_0 для двух выборок, заключающаяся в том, что совпадение значений атрибута у пары вершин (x,z) не зависит от выборки. Для этой цели рассматривалась случайная величина, принимающая значение 1, если $y_x = y_z$, и 0 в противном случае. Для каждой из случайных величин h, c, cs сравнивались выборки $B_\xi(\alpha)$ и Ω. Также сравнивались выборки $B_{cs}(\alpha)$ и $B_h(\alpha)$, $B_{cs}(\alpha)$ и $B_c(\alpha)$. Заметим, что гипотеза H_0 в случае этих сравнений выполняется для $\alpha = 1$ в силу утверждения 1. Для обозначенных сравнений выборок при значениях $(1-\alpha) \in \{0.25, 0.5, 0.75, 0.8, 0.95, 0.99, 0.999\}$ были проведены статистические t-тесты Стьюдента.

Результаты тестов показали, что выборки $B_h(\alpha)$, $B_c(\alpha)$ и $B_{cs}(\alpha)$ статистически значимо (с p-value менее 1%) отличаются от всех пар вершин Ω по совпадению значений атрибута, на всех рассматриваемых наборах данных. Во всех случаях наблюдается, что значения атрибута для пар из B_ξ совпадают чаще, чем для произвольной пары из Ω, за исключением свойства h в наборе данных pokec для атрибута пол. Также в большинстве случаев тесты показывают, что совпадение значений атрибута на выборках $B_{cs}(\alpha)$ происходит значимо чаще, чем на выборках $B_h(\alpha)$ и $B_c(\alpha)$.

Для наглядной демонстрации результатов анализа данных построим графики с вероятностями совпадения значений атрибутов. Определим:

$$r^\xi(\alpha) = P\big(y_x = y_z | B_\xi(\alpha)\big) \tag{7}$$

Пусть $r_i^\xi = r(\alpha_i^\xi)$, где $i = \overline{1,N_\xi}$. Значение r_i^ξ показывает вероятность совпадения значений атрибута для выборки размера α_i^ξ, включающую пары, для которых значение свойства ξ не менее заданного порога.

Дополнительно определим референсное значение $R = P(y_x = y_z)$ – вероятность того, что две случайно выбранные размеченные вершины имеют одинаковую метку, независимо от значения ξ.

Утверждение 2 $R = r_1^h = r_1^c = r_1^{cs}$.

Построим график, показывающий изменение вероятности совпадения значений атрибута при увеличении значений величин h, c, cs. С увеличением этих значений α^ξ будет уменьшаться. Поэтому на оси абсцисс расположим значения $1-\alpha_\xi$, по оси ординат расположим значения r^ξ. Изобразим точку (α_2^h, r_2^h) на графике. Аналогично изобразим точки (α_i^c, r_i^c), $i \in \overline{2,N_c}$ и точки $(\alpha_i^{cs}, r_i^{cs})$, $i \in \overline{2,N_{cs}}$. Из теоремы 1 и утверждения 2 следует, что $(\alpha_1^h, r_1^h) = (\alpha_1^c, r_1^c) = (\alpha_1^{cs}, r_1^{cs}) = (1, R)$. Для большей наглядности вместо трёх одинаковых точек отобразим референсное значение R в виде горизонтальной линии.

На рисунке 1 представлены графики с результатами анализа набора данных twitter для атрибута род деятельности и набора данных vk2 для атрибута пол. Графики показывают вероятности r совпадения значений атрибута для доли пар (равной α) с наибольшим значением свойств h, c и cs.

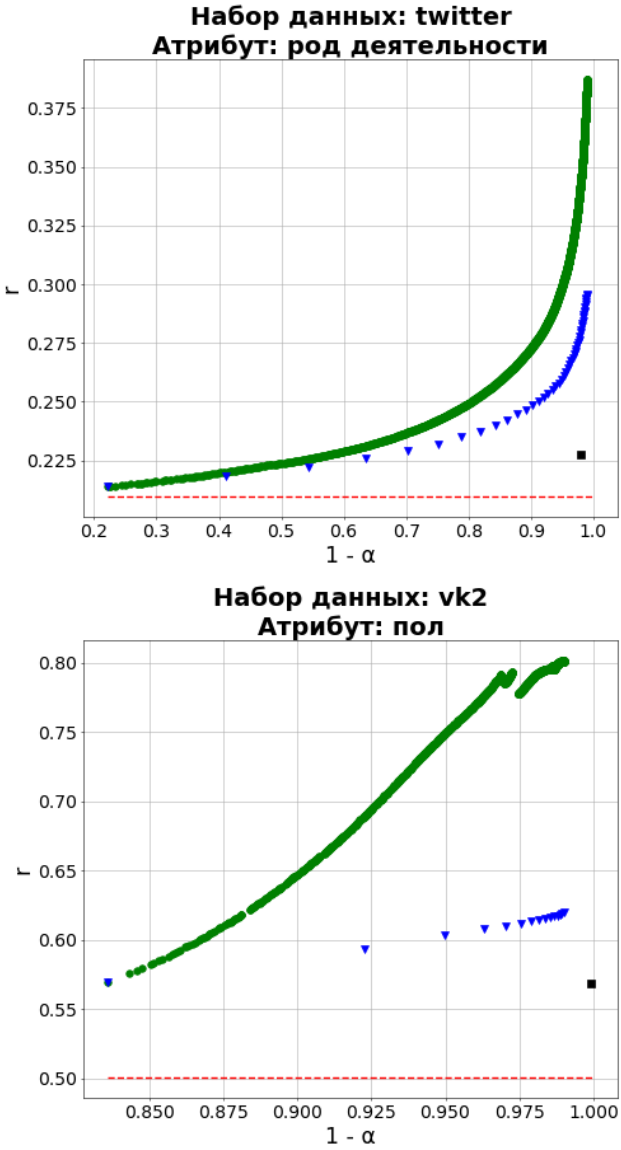

Рис. 1 — Анализ наборов данных. Представлены значения α (ось абсцисс) и r (ось ординат). Значение R – красная пунктирная линия; значение для h – черный квадрат; значение для c – синие треугольники; значение для cs – зеленые круги.

Из графиков видно, что с увеличением h, c и cs увеличивается вероятность совпадения значений атрибута для двух вершин с заданным значением свойства. Стоит отметить, что эта тенденция наиболее выражена у свойства cs. Аналогичные результаты наблюдались и для остальных пар (набор данных, атрибут). Дополнительно стоит отметить, что в наборе данных vk1 для атрибутов род деятельности и возраст наблюдается выраженная «гомофилия».

Проведённые статистические тесты и наблюдения из построенных графиков позволяют сделать вывод о применимости свойства специфичности контекста для предсказания значений атрибутов.

В разделе 2.5 предлагается и формулируется подход для предсказания значений атрибутов на основе специфичности контекста. Подход заключается в вычислении специфичности контекста для каждой из вершин социального графа и использования полученных значений в качестве веса, количественной характеристики важности связи с этой вершины, представленной ребром.

В **третьей главе** представляется несколько методов на основе предложенного подхода, использующих специфичность контекста для предсказания значений атрибутов пользователей, оценивается их вычислительная сложность, проводится экспериментальное сравнение с существующими методами, даются рекомендации к использованию.

Раздел 3.1 посвящён описанию предлагаемых методов. **Метод *LP-CS*** является модификацией 2-шагового синхронного алгоритма распространения меток. На первом шаге алгоритма для каждой вершины $z \in V$ вычисляются значение специфичности $s(z)$ и метка $v(z) \in A$, являющееся некоторым значением атрибута. Специфичность вычисляется согласно формуле (2). Метка $v(z)$ для вершины z выбирается как самая частая метка среди меток соседних вершин $x \in N_z$ в случае задачи классификации (атрибуты пол, род деятельности) или как среднее значение атрибута среди значений соседних вершин. На втором шаге предсказываемое значение $p(x)$ атрибута $\mathcal{A}$ для вершины x вычисляется как метка a_i, соответствующая максимальной сумме значений специфичности среди соседних вершин вершины x с заданным значением атрибута a_i:

$$p(x) = \arg\max_{a_i \in A} \sum_{z \in N_x} \mathbb{1}_{v(z)=a_i} \cdot s(z) \tag{8}$$

Значения $p(x)$ для вершин $x \in V$ являются выходом *LP-CS* алгоритма.

Смысл алгоритма заключается в следующем. Величина $v(z)$ характеризует контекст вершины z и представляет собой наиболее специфичное, выделяющееся, частое значение атрибута среди её соседей в графе. Значение атрибута $p(x)$ для вершины x предсказывается на основе меток $v(z)$ соседних вершин, представляющих характеристику их контекста, с учётом специфичности соседних вершин. На первом шаге известные значения

атрибута y_x распространяются к вершине z от её соседей $x \in N_z$ и сохраняются в виде характеристики контекста вершины z – наиболее специфичным значением атрибута соседей $v(z)$ и количественной мерой специфичности $s(z)$. Так как каждая вершина x является частью контекста своих соседей, предсказываемое для неё значение $p(x)$ оценивается на втором шаге алгоритма путём «обратного распространения» вычисленных на первом шаге характеристик контекста $v(z)$ и $s(z)$. Значение атрибута $p(x)$ выбирается как метка с наибольшим весом среди меток контекста соседних вершин N_x: каждое значение $v(z)$ учитывается с весом, равным $s(z)$. Таким образом, алгоритм считает подписку на публичную страницу, посвящённую, например, рыбалке и охоте, более значимой, чем подписку на страницу артиста, одинаково популярно среди мужчин и женщин, для предсказания неизвестных значений пола для пользователей.

Метод *LP-CS-Gen* является модификацией алгоритма *LP-CS*, учитывающей неравномерность распределения значений атрибута среди размеченных пользователей. Он отличается лишь способом выбора метки $v(z)$ на первом шаге. Метка $v(z)$ для вершины z выбирается таким образом, что $d(N_z)$ максимально превышает распределение генеральной совокупности $d(Y)$ в выбранной точке:

$$v(z) = \arg\max_{a_i \in A} kl_i\big(d(N_z)\big|d(Y)\big) \tag{9}$$

Здесь $kl_i = \log p_i \frac{p_i}{q_i}$ – член суммы в формуле дивергенции Кульбака–Лейблера в случае дискретных распределений, соответствующий значению атрибута a_i.

Для **метода *Distr2-CS-XGB*** вводятся новые признаки для вершин графа. Признаковый вектор *Distr2-CS* представляет собой дискретное распределение значений $\mathcal{A}$, определение которого будет представлено далее. Так как этот вектор строится, чтобы с его помощью предсказывать значения $\mathcal{A}$ для некоторой вершины x, логично потребовать от него быть независимым от значения y_x. Другими словами, вектор для вершины x будет одинаковым, независимо от того, известно ли значение атрибута y_x для этой вершины или нет.

Сначала определим для вершины v относительную специфичность контекста $\hat{s}(v|x)$, игнорирующую вершину x:

$$\hat{s}(v|x) = KL\big(d(N_v \backslash \{x\})\big|\big|d(Y)\big) \tag{10}$$

Определим относительную специфичность контекста $\hat{cs}(z|x)$ для вершин z и x относительно вершины x:

$$\hat{cs}(z|x) = \sum_{v \in N_x \cap N_z} \hat{s}(v|x) \tag{11}$$

Теперь определим признаковый вектор *Distr2-CS*. Вектор представляет собой взвешенное распределение (англ.**Distr**ibution) значений $\mathcal{A}$ среди пользователей из **2**-шаговой окрестности N_x^2), построено с применением ***CS***-свойства (англ. **C**ontext **S**pecificity). Каждая компонента вектора $Distr2\text{-}CS(x) \in \mathbb{R}^{|A|}$ определяется следующим образом:

$$Distr2\text{-}CS(x)_i = \frac{\sum\limits_{z \in N_x^2 \cap Y} \mathbb{1}_{y_z = a_i} \cdot \hat{cs}(z|x)}{Norm} \tag{12}$$

Здесь $N_x^2 = \{z : (x,v) \in E, (v,z) \in E, z \neq x\}$ – *двухшаговая окрестность*, множество вершин, достижимых из x ровно в два шага (перехода по рёбрам), кроме самой вершины x, $Norm$ – нормализационная константа, такая что $Distr2\text{-}CS(x)$ удовлетворяет условию $\sum_{i=1}^{|A|} Distr2\text{-}CS(x)_i = 1$. Если все компоненты в числителе равны 0, определим $Distr2\text{-}CS(x) \equiv \vec{0}$. Стоит отметить, что в рассматриваемых в диссертационной работе наборах данных такой случай не встречался.

Признаки *Distr2-CS* для всех вершин вычисляются в два этапа. Сначала для каждой вершины вычисляется ненормализованное распределение значений атрибута среди соседних вершин. Затем на основе этих распределений для каждой вершины вычисляются значения ***Distr2-CS***. Применив классификатор или регрессор XGBoost для описанных признаков получим новый метод *Distr2-CS-XGB*.

Метод *Distr2-CS+DW[n]-XGB* основан на комбинировании (конкатенации) двух векторных представлений вершин. *DW[n]* – векторными представлениями вершин графа, извлекаемое только с использованием структуры графа, полученными методом DeepWalk. *Distr2-CS* – описанное выше представление, использующее как структуру 2-й окрестности вершины, так и значений атрибута других пользователей со 2-й окрестности.

Метод GConv-CS[n] представляет собой классическую однослойную свёрточную графовую нейронную сеть с 2 модификациями. Так как рассматривается только структура социального графа и отсутствуют признаки для вершин графа, в качестве входных параметров нейронной сети рассматриваются обучаемые векторы размера n. Вторая модификация заключается в использовании специфичности контекста вершин для выделения наиболее важных для предсказания значений атрибута вершин. Для этой цели дополнительно к функции потерь добавляется регуляризация $MSE(l2(x), s(x))$. Метод применяется только для задач классификации, т.е. для предсказания атрибутов пол и род деятельности.

В разделе 3.2 формулируется теорема о вычислительной сложности разработанных методов.

Теорема 2 Пусть n – размер векторных представлений вершин. Тогда вычислительная сложность методов составляет:

– *LP-CS* и *LP-Gen* – $O\Big(|V| + |E|\Big)$;

– *Distr2-CS-XGB* – $O\Big(|V|\log|V| + |E|\Big)$;

– *Distr2-CS+DW[n]-XGB* – $O\Big(n|V|\log|V| + |E|\Big)$;

– *GConv-CS[n]* – $O\Big(n|E| + n|V|\Big)$.

В разделе 3.3 обсуждаются другие теоретические особенности разработанных методов. В частности, предложенные признаки Distr2-CS сравниваются со статическими векторными представлениями вершин графа. Отмечается применимость всех разработанных методов к двудольным графам, описывается, какие рёбра графа используются в методах, а какие необязательны для сбора. Кроме того, описываются какие методы и при каких условиях применимы к задачам классификации и регрессии, что зависит от рассматриваемого атрибута.

В разделе 3.4 описывается постановка эксперимента и результаты сравнения разработанных методов с аналогичными методами, не учитывающими специфичность контекста вершин социального графа.

Сначала описываются базовые методы. *LP[1]* и *LP[2]* – синхронный алгоритм распространения меток с одной и двумя итерациями, соответственно. Метка представляет собой значение атрибута, на каждом шаге алгоритма новое значение метки для вершины вычисляется как самое частое значение среди меток соседей (для атрибутов пол и род деятельности), или среднее значение метки (для атрибутов возраст и доход).

Distr2-XGB – модификация авторского метода *Dist2-CS-XGB*, игнорирующая значения специфичности контекста. Вместо специфичности общего контекста используется размер общего контекста:

$$Distr2(x)_i = \frac{\sum\limits_{z \in N_x^2 \cap Y} \mathbb{1}_{y_z = a_i} \cdot |N_x \cap N_z|}{Norm} \tag{13}$$

Метод *DW[n]-XGB* заключается в построении векторных представлений, полученных методом DeepWalk, и применении к ним классификатора или регрессора XGBoost.

GConv[n] аналогичен методу *GConv-CS[n]*. Отличие в том, что в *GConv[n]* не применятся регуляризация, норма $l2$ не связывается со специфичностью контекста вершины. Аналогично *GConv-CS[n]*, этот метод применяется только для задач классификации.

Опишем процесс экспериментального сравнения методов. Всего имеется 9 различных комбинаций (набор данных, атрибут). 5 из них – задачи классификации (род деятельности, пол), оставшиеся 4 – задачи регрессии (возраст, доход). Для каждой из 9 комбинаций выполнялась следующая последовательность действий. Множество размеченных пользователей Y случайным образом разбивалось на обучающую (80%) и проверочную (20%) части. Каждый из методов был применен к обучающей выборке для

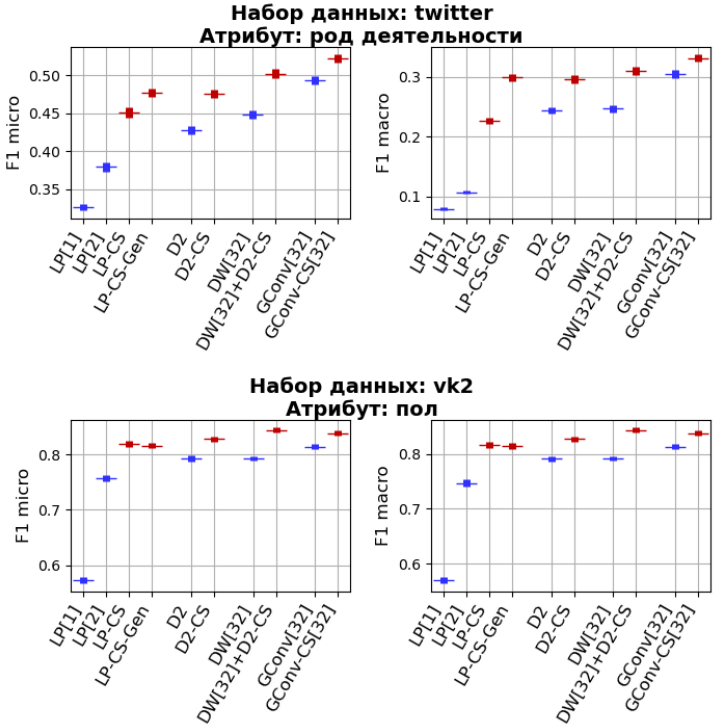

Рис. 2 — Результаты экспериментального сравнения методов. К признакам *Distr2*, *Distr2-CS*, *DW[n]* и *DW[n]+Distr2-CS* применялся XGBoost. Представлены среднее значения метрик качества и доверительные интервалы для уровня доверия $p = 0.95$.

построения модели предсказания, затем полученная модель была использована для оценки качества на проверочной части. Для задач классификации использовались значения F1-меры с микро- и макроусреднением. Для задач регрессии использовались метрики R2 и среднеквадратичная ошибка (MAE). Эта последовательность действий повторялась по 30 раз для каждой комбинации (набор данных, атрибут). После чего были подсчитаны средние значения метрик и доверительные интервалы, полученные с использованием t-распределения Стьюдента, с уровнем доверия $p = 0.95$.

Результаты экспериментального сравнения показывают, что методы, основанные на специфичности контекста, *LP-CS* и *LP-CS-Gen*, *Distr2-CS-XGB*, *DW[n]+Distr2-CS-XGB*, *GConv-CS[n]*, показывают более высокое качество по сравнению с методами *LP[2]*, *Distr2-XGB*, *DW[n]-XGB*, *GConv[n]*, соответственно. На рисунке 2 представлены результаты для атрибута род деятельности на наборе twitter и для атрибута пол на наборе vk2. На других парах (набор данных, атрибут) наблюдалось аналогичные результаты.

В разделе 3.5 на основе теоретической оценки вычислительной сложности и экспериментального сравнения разработанных методов представлены рекомендации к их использованию.

Рассмотрим задачу классификации. Если необходимо быстрое решение с приемлемым качеством, то рекомендуется использовать метод LP-Gen-CS. Если необходимо максимально качественное решение, то рекомендуется использовать GConv-CS[n]. Значение n необходимо подбирать эмпирически, при бóльшем количестве вершин в графе рекомендуется использовать бóльшее значение n. Если необходим компромисс между скоростью и качеством работы, предлагается воспользоваться методами Distr2-CS и Distr2-CS+DW. Второй из них показывает бóльшее качество, однако для его работы необходимо вычислить векторные представления DeepWalk. Стоит отметить, что эти представления не зависят от атрибута и могут быть переиспользованы при изменении атрибута, но при неизменном графе.

Не все из представленных методов применимы для задач регрессии. Рекомендуется использовать метод LP-CS, который показывает как высокую скорость, так и высокое качество. В некоторых случаях более высокое качество достигается с использованием методов Distr2-CS или Distr2-CS+DW. Однако в случае регрессии эти методы применимы только когда $|A| << |V|$. Вычислительная сложность Distr2-CS+DW выше, чем Distr2-CS, но в некоторых случаях Distr2-CS+DW показывал более высокое качество.

Итак, в третьей главе представлены методы предсказания значений демографических атрибутов пользователей по социальному графу, на основе специфичности контекста. Экспериментально показано их преимущество над существующими методами. Даны рекомендации к применению разработанных методов.

Четвёртая глава посвящена программной системе для определения значений стационарных демографических атрибутов пользователей социальных сетей.

Программная система для предсказания значений демографических атрибутов пользователей социальных сетей по социальному графу и представляет собой фреймворк. В нём реализованы методы предсказания значений атрибутов пользователей по социальному графу. Фреймворк позволяет сравнить качество различных методов предсказания значений демографических атрибутов. Имеется возможность добавить новые методы, признаковые описания вершин графа, настраивать методы оценки качества. Кроме того, система позволяет оформить результаты экспериментального сравнения методов в виде графиков с настраиваемыми цветами, подписями, типами линий и точек. Реализована возможность проанализировать набор данных с целью оценки свойств «гомофилии» (h), зависимостей между размером общим контекстом и значениями атрибутов

(c), зависимости между специфичностью общего контекста и значениями атрибутов (cs). Результаты анализа оформляются в виде графиков. Исходный код программной системы составляет около 4100 строк на языке Python 3 и 520 строк на языке C++.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработан подход для предсказания значений демографических атрибутов на основе специфичности контекста вершин социального графа;
2. В рамках подхода созданы новые методы предсказания значений демографических атрибутов по социальному графу *LP-CS*, *LP-CS-Gen*, *Distr2-CS+DW[n]-XGB*, *GConv-CS[n]*, *Distr2-CS-XGB*, превосходящие по качеству существующие аналоги; даны рекомендации по их применению;
3. Реализована программная система предсказания значений атрибутов пользователей социальных сетей по социальному графу, позволившая экспериментально подтвердить превосходство созданных методов над существующими аналогами по качеству решения задачи.

В рамках обзора были рассмотрены методы определения значений демографических атрибутов пользователей по текстам сообщений и социальному графу, были выявлены недостатки существующих методов. Было описано и определено свойство специфичности контекста вершины графа для заданного атрибута. На нескольких социальных графах, собранных из реальных социальных сетей, экспериментально показано, что специфичность контекста может быть использована для предсказания значений атрибутов. Предложен подход для предсказания значений атрибутов пользователей на основе специфичности контекста и несколько методов на основе подхода. Методы *LP-CS* и *LP-CS-Gen* являются вариацией алгоритма распространения меток. В методах *Distr2-CS-XGB* и *Distr2-CS+DW[n]-XGB* использовались предложенные признаки *Distr2-CS*, основанные на специфичности контекста Метод *GConv-CS[n]* основан на свёрточных графовых нейронных сетях. Даны рекомендации по применению разработанных методов.

Публикации автора по теме диссертации

1. *Gomzin A., Drobyshevskiy M., Turdakov D.* Context specificity matters: profile attributes prediction for social network users //Conference on Information Sciences and Systems (CISS), Johns Hopkins University.– 2021.

2. *Gomzin A., Laguta A., Stroev V., Turdakov D.* Detection of author's educational level and age based on comments analysis //Dialogue. – 2018.
3. *Гомзин А. Г., Кузнецов С. Д.* Метод автоматического определения возраста пользователей с помощью социальных связей //Труды Института системного программирования РАН. – 2016. – Т. 28. – №. 6.
4. *Гомзин А. Г., Кузнецов С. Д.* Методы построения социо-демографических профилей пользователей сети Интернет //Труды Института системного программирования РАН. – 2015. – Т. 27. – №. 4.
5. *Гомзин А. Г.* Предсказание рода деятельности пользователей социальной сети //Ломоносовские чтения-2020. Секция Вычислительной математики и кибернетики. – Секция Вычислительной математики и кибернетики. – М.: М., 2020. – С. 56–57.
6. *Гомзин А.Г., Коршунов А.В. и др.* Система сбора пользовательских данных из онлайновых социальных сетей //Свидетельство №2015616047 о государственной регистрации программы для ЭВМ – 2015
7. *Гомзин А.Г., Турдаков Д.Ю. и др.* Talisman //Свидетельство №2018615539 о государственной регистрации программы для ЭВМ – 2018
8. *Гомзин А.Г., Турдаков Д.Ю.* Веб-приложение для разметки рода деятельности пользователей социальной сети //Свидетельство №2019661808 о государственной регистрации программы для ЭВМ – 2019
9. *Гомзин А.Г. Турдаков, Д.Ю.* Программное средство методов предсказания рода деятельности пользователя социальной сети по его социальным связям //Свидетельство №2019663796 о государственной регистрации программы для ЭВМ – 2019
10. *Гомзин А.Г., Дробышевский М.Д., Турдаков Д.Ю.* Фреймворк для сравнения методов предсказания значений атрибутов пользователей социальных сетей //Свидетельство №2020666741 о государственной регистрации программы для ЭВМ – 2020

Гомзин Андрей Геннадьевич

Методы и программные средства определения значений стационарных демографических атрибутов пользователей социальных сетей

Автореф. дис. на соискание ученой степени канд. физ.-мат. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____