ОТЗЫВ

официального оппонента на диссертационную работу Гомзина Андрея Геннадьевича «Методы и программные средства определения значений стационарных демографических атрибутов пользователей социальных сетей», представленную к защите на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальность

Диссертационная работа посвящена методам и программным средствам определения значений признаков пользователей: пола, года рождения и др., которые используются в различных прикладных задачах анализа данных. В частности, подобная информация используется в системах рекомендаций для анализа целевой аудитории и поиска потенциальных клиентов в социальных сетях, эффективного поиска потенциальных сотрудников. Указанные явно и предсказанные значения признаков также полезны для анализа общественных тенденций. Существующие методы предсказания значений признаков обладают недостатками. В частности, они не учитывают значимость связей со специфичными для предсказываемого признака вершинами, например, повышенную значимость подписок на профессиональные сообщества для предсказания рода деятельности.

Диссертационная работа состоит из введения, четырёх глав, заключения, двух приложений. Список литературы содержит 105 наименований. Общий объём диссертации составляет 143 страницы.

Во введении обосновывается актуальность темы исследования, формулируются цель работы и поставленные для её достижения задачи, определяется научная новизна, теоретическая и практическая значимость, приводятся основные положения, выносимые на защиту. В первой главе приводится обзор существующих методов определения значений демографических атрибутов пользователей. Рассматриваются различные постановки задач предсказания значений признаков. Особое внимание уделяется методам предсказания по текстам сообщений пользователей и по социальному графу. Также описываются подходы, применяемые при сборе данных из социальных сетей, в частности, социального графа и референсных значений. В конце главы описываются выявленные в результате обзора

недостатки существующих методов предсказания значений демографических признаков пользователей. Во второй главе описывается новый подход для предсказания значений демографических атрибутов пользователей по социальному графу. В основе подхода лежит введённое в работе свойство вершин социального графа, специфичность контекста. В работе представлено формальное определение этого свойства, основанное на дивергенции Кульбака-Лейблера. Идея подхода заключается в том, что связь с более специфичными вершинами является более значимой для предсказания значений признаков. Подход заключается в вычислении значений специфичности контекста для вершин и использовании их в качестве количественной оценки важности, значимости связей с вершинами. Экспериментально обосновывается применимость предлагаемого подхода. Для этого используются несколько существующих и новых, собранных в рамках работы, наборов данных из реальных социальных сетей. Было экспериментально показано, что вероятность совпадения значений признаков увеличивается при увеличении суммарной специфичности контекста. Также было показано, что специфичность общего контекста сильней коррелирует со значениями признаков, чем размер общего контекста и наличие ребра между вершинами.

В третьей главе предлагаются новые методы предсказания значений атрибутов на основе предложенного подхода. Предложены модификации алгоритма распространения меток. Разработаны новые признаки, описывающие вершины графа, для применения методов машинного обучения с учителем. В отличие от используемых в литературе представлений вершин, использующих только структуру графа, предлагаемые представления помимо структуры используют значения атрибутов других вершин: при вычислении значений этих признаков учитывается специфичность контекста инцидентных вершин. Предложен метод, использующий введённые признаки как отдельно, так и совместно с векторными представлениями вершин DeepWalk. Предлагается метод на основе свёрточной графовой нейронной сети. Специфичность контекста вершин используется для регуляризации обучаемых в рамках нейронной сети векторных представлений вершин. Было экспериментально показано, что каждый из разработанных методов превосходит по качеству соответствующий базовый метод, не использующий специфичность контекста. В работе приводится теоретическая оценка вычислительной сложности для всех разработанных методов, а также фактическое время работы методов на различных наборах данных. На основе полученных

оценок качества и времени работы разработанных методов были сформулировать рекомендации к их выбору и использованию.

В четвёртой главе описывается программная система, в которой реализованы описанные методы предсказания значений демографических атрибутов пользователей социальных сетей по социальному графу. Описываются возможности программной системы, перечисляются используемые языки программирования, библиотеки и программы, представляется архитектура программной системы, приводится диаграмма классов, реализованных в рамках программной системы.

Можно отметить следующие научные результаты, представленные автором в диссертации:

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, 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.

Предлагаемый подход ранее не использовался в методах предсказания значений признаков пользователей. Методы, реализованные в рамках подхода, показывают более высокое качество по сравнению с аналогами. Диссертационная работа, безусловно, имеет практическую значимость.

Результаты исследования опубликованы в 5 печатных работах (в том числе 3 в журналах из перечня ВАК, докладывались на 7 российских и международных конференциях, семинарах.

Работа не лишена некоторых недостатков. В частности, в работе рассмотрены только свёрточные графовые нейронные сети. Не рассмотрены другие виды графовых нейронных сетей, в частности GAT (Graph Attention Network). На странице 107 противопоставляются два подхода к сравнению

качества методов, оба из которых (а не только первый) относятся к вариантам скользящего контроля. В формулировке Теоремы 1 на стр. 67 допускается небрежность: необходимо указать существование пары вершин x и y, для которых значение $h(x,y) > 0$. Слово «атрибут» используется в работе наряду с классическим для русскоязычной литературы по анализу данных термином «признак». Встречаются термины-кальки из профессионального жаргона, использование которых можно было бы избежать в тексте. Имеются некоторые опечатки, хотя их количество весьма мало.

Обозначенные недостатки не влияют на общую положительную оценку диссертационной работы. Диссертация Андрея Геннадьевича Гомзина «Методы и программные средства определения значений стационарных демографических атрибутов пользователей социальных сетей» является завершенной работой, в которой автору удалось решить поставленные перед диссертационным исследованием задачи. Основные результаты диссертации полностью опубликованы, докладывались на российских и международных конференциях. Автореферат диссертации отражает ее содержание.

Диссертация отвечает требованиям Положения ВАК РФ о порядке присуждения ученых степеней, а ее автор, Андрей Геннадьевич Гомзин, заслуживает присуждения ему учёной степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент

Доктор физико-математических наук,
руководитель департамента анализа
данных и искусственного интеллекта
Национального исследовательского университета
«Высшая школа экономики»
С.О. Кузнецов