ОТЗЫВ

ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Гомзина Андрея Геннадьевича

«Методы и программные средства определения значений стационарных демографических атрибутов пользователей социальных сетей», представленную на соискание учёной степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальность темы. Одной из актуальных задач анализа онлайновых социальных сетей является определение значений демографических атрибутов пользователей. Так, например, в рекомендательных системах для выявления целевой аудитории предлагаемых продуктов и поиска клиентов используется пол, возраст и интересы пользователей социальных медиа. В организациях информация о демографических атрибутах необходима для эффективного поиска потенциальных сотрудников. Эта же информация может применяться для анализа демографических тенденций в обществе в целом и ее отдельных группах. Однако профили пользователей в онлайновых социальных сетях часто являются неполными или некорректными, и в этом случае возникает задача восстановления значений атрибутов пользователей. Диссертационная работа посвящена методам и программным средствам определения значений атрибутов пользователей, которые являются необходимыми для решения ряда прикладных задач в социально-экономической сфере.

Оценка содержания работы. Диссертационная работа Гомзина А.Г. состоит из введения, четырёх глав, заключения, списка литературы из 105 наименований и

двух приложений. Объём диссертации составляет 143 страницы.

Во введении обосновывается актуальность темы исследования, формулируется цель работы и поставленные для её достижения задачи, определяется научная новизна, теоретическая и практическая значимость, а также выносимые на защиту положения.

В первой главе приводится обзор существующих методов определения значений демографических атрибутов пользователей. Рассматриваются задачи предсказания значений атрибутов по текстам сообщений пользователей и по структуре связей между пользователями социальной сети (так называемому социальному графу). Рассматриваются сопутствующие задачи, в частности задача сбора данных о социальном графе сети и референсных значениях атрибутов, которые необходимы для обучения моделей и сравнения качества различных методов. Отмечаются недостатки существующих методов предсказания значений демографических атрибутов пользователей.

Во второй главе предлагается новый подход для предсказания значений демографических атрибутов пользователей по социальному графу. В основе подхода лежит идея того, что связь пользователя с более специфичными вершинами графа является более значимой для предсказания значений атрибутов. Для специфичности контекста вершин размеченного графа дается формальное определение, которое позволяет количественно оценить важность связей с вершинами графа. Адекватность предлагаемого подхода экспериментально обосновывается при помощи данных из реальных онлайновых социальных сетей. Для этого оценивается зависимость между суммарной специфичностью контекста общих вершин и совпадением значений атрибута для случайных пар вершин. Показано, что вероятность совпадения значений атрибутов увеличивается при увеличении суммарной специфичности контекста. Кроме того, показано, что специфичность имеет большую связь со значениями атрибутов, чем размер общего

контекста и наличие ребра между вершинами.

Третья глава посвящена разработке методов предсказания значений атрибутов на основе похода, рассмотренного во второй главе. Предложены методы LP-CS и LP-CS-Gen, которые являются модификациями алгоритма распространения меток и принимают в расчет специфичность вершин социального графа. Предложены методы Distr2-CS-XGB и Distr2-CS+DW[n]-XGB, которые используют новые признаки для вершин графа, представляющие не только структуру сети, но и значения атрибутов вершин социальной сети. При вычислении таких признаков учитывается специфичность контекста инцидентных вершин. В методе Distr2-CS+DW[n]-XGB эти признаки используются совместно с векторными представлениями вершин, отражающими структуру социального графа. Кроме того предложен метод GConv-CS[n], который основан на современных свёрточных графовых нейронных сетях. В этом методе специфичность контекста вершин применяется для регуляризации обучаемых в рамках нейронной сети векторных представлений вершин. Экспериментально показано, что каждый из разработанных методов превосходит по качеству соответствующий базовый метод, в котором специфичность контекста не используется. В работе приводится теоретическая оценка вычислительной сложности для всех разработанных методов, а также время работы методов на различных наборах данных. Оценки качества и времени работы алгоритмов позволили сформулировать в третьей главе рекомендации по выбору и применению разработанных методов.

В четвёртой главе рассматривается программная система, в которой реализованы введенные ранее методы предсказания значений демографических атрибутов пользователей социальных сетей по социальному графу. Описываются такие возможности программной системы, как предсказание значений демографических атрибутов, сбор данных из реальных социальных сетей, сравнение различных методов предсказания значений атрибутов, а также получение

визуальных представлений результатов сравнения. Описывается архитектура программной системы, приводится реализованная в системе модель классов, обозначаются применяемые библиотеки и программы.

В заключении сформулированы основные результаты работы.

Оценка научной новизны и достоверности. В работе получены новые теоретические и практические результаты. Автором разработан подход для предсказания значений демографических атрибутов на основе специфичности контекста вершин социального графа. В рамках предложенного подхода созданы методы предсказания значений демографических атрибутов по социальному графу и даны рекомендации по их применению. Данные методы применяются в специально разработанной программной системе предсказания значений атрибутов пользователей социальных сетей по социальному графу, которая позволила экспериментально подтвердить качество разработанных методов.

Предлагаемый подход к предсказанию значений атрибутов является новым, а реализованные в рамках подхода методы показывают высокое качество по сравнению с аналогами. Достоверность результатов обосновывается экспериментальной проверкой возможности использования специфичности контекста для предсказания значений атрибутов на основе данных реальных социальных сетей, а также экспериментальным сравнением разработанных методов с аналогичными современными методами определения демографических атрибутов.

Значимость для науки и практики результатов диссертационного исследования. Теоретическая и практическая значимость полученных в диссертации результатов заключается в том, что в работе предложен подход к предсказанию значений стационарных демографических атрибутов пользователей социальных сетей на основе специфичности контекста вершин социального графа. Разработанные в рамках подхода методы могут быть включены исследователями в

набор референсных методов и использоваться для развития исследований в области методов анализа сложных социальных систем с сетевой структурой. Предложенные методы превосходят по качеству аналогичные методы и могут быть эффективно использованы для решения прикладных социально-экономических задач с учетом предложенных в работе рекомендаций. Применимость предложенного подхода и методов показана на данных из реальных социальных сетей.

Подтверждение опубликования основных результатов диссертации. Результаты работы Гомзина А. Г. докладывались на семи российских и международных конференциях, а также на научных семинарах. Основные результаты диссертации опубликованы в 5 печатных работах (в том числе три в журналах из перечня ВАК).

Соответствие содержания автореферата и диссертации. Автореферат соответствует основным результатам, выводам и положениям диссертационного исследования и верно отражает его содержание.

Недостатки и замечания по диссертационной работе. Диссертационная работа Гомзина А. Г. не содержит серьезных недостатков, однако к содержанию работы могут быть сделаны следующие замечания.

- Нет обоснования выбора определения специфичности для пары вершин сети *cs* как суммы специфичности контекста по их общим соседям. Не обсуждается вопрос, почему рассматривается именно такая величина, почему, в частности, величина является ненормированной. Не обсуждается вопрос, как влияет плотность первой и второй окрестности вершины на предсказываемое значение атрибута (например, в методе LP-CS). Представляется, что величины специфичности могут сильно коррелировать в плотной группе соседей вершины.
- В разделе 3.5 для разработанных методов предсказания значения атрибута

хотелось бы увидеть рекомендации по их применению не только исходя из вычислительных особенностей и возможностей, но и из содержательных различий методов.

- На странице 53 приведено определение двухшаговой окрестности вершины x, ее формальное определение «$\{x: (x,v) \in E, (v,z) \in E, z \neq x\}$» неточно соответствует описанию «множество вершин, достижимых из x ровно в два шага …» для графов с циклами.
- Опечатка на странице 62 «$B := B \backslash \{n\}$» (строка 16 алгоритма 5), должно быть «$B := B \cup \{n\}$».
- Опечатка в формуле дивергенции Кульбака-Лейблера на страницах 65 и 85, должно быть «$\sum_i p_i \log \frac{p_i}{q_i}$», а не «$\sum_i \log p_i \frac{p_i}{q_i}$».
- В строке 6 алгоритма 6 (страница 88), вероятно, допущена опечатка, должно быть не "$d[x][y_x] := d[x][y_x] + 1$", а "$d[x][y_v] := d[x][y_v] + 1$"; в строке 15 появляется не определенная ранее в алгоритме величина a_i.
- На странице 90 размерность матрицы W задана как n x $|A|$, а должна быть $|A|$ x n.
- В диссертации используется индексирование объектов по значению атрибута, которое в общем случае может быть не целочисленным. В ряде случаев тип объектов не обозначен, что затрудняет чтение и понимание работы.
- Описания алгоритмов являются довольно лаконичными, следовало бы оставить больше пояснений к используемым обозначениям и переменным.
- В тексте диссертации и автореферата имеется довольно значительное число ошибок и опечаток.

Указанные замечания не являются принципиальными и не снижают общей положительной оценки диссертационной работы.

Заключение. Диссертационное исследование Гомзина А. Г. является завершенной научно-квалификационной работой, обладает актуальностью, научной новизной и практической значимостью, проведено на высоком научно-техническом уровне.

Результаты диссертационного исследования соответствуют специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей», а разработанные теоретические положения и полученные результаты имеют важное научное и прикладное значение.

Диссертация соответствует требованиям, предъявляемым ВАК Российской Федерации к диссертациям на соискание ученой степени кандидата физико-математических наук, а её автор, Гомзин Андрей Геннадьевич, заслуживает присуждения учёной степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент,
кандидат технических наук,
старший научный сотрудник
Федерального государственного бюджетного учреждения науки
Института проблем управления им. В.А. Трапезникова РАН

Д. А. Губанов