На правах рукописи

Чесноков Владислав Олегович

АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ АНАЛИЗА ГРАФОВ БЛИЖАЙШЕГО ОКРУЖЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ БОТОВ И ОПРЕДЕЛЕНИЯ НЕУКАЗАННЫХ АТРИБУТОВ ПОЛЬЗОВАТЕЛЕЙ В ОНЛАЙНОВЫХ СОЦИАЛЬНЫХ СЕТЯХ

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

Автореферат

диссертации на соискание учёной степени

кандидата технических наук

Москва — 2018

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н. Э. Баумана).

Научный руководитель: **Ключарёв Петр Георгиевич**,
кандидат технических наук, МГТУ им. Н. Э. Баумана, доцент

Официальные оппоненты: **Рязанов Владимир Васильевич**,
доктор физико-математических наук, профессор, Федеральное государственное бюджетное учреждение науки Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, главный научный сотрудник

Коркин Игорь Юрьевич,
кандидат технических наук, Федеральное государственное унитарное предприятие «Центральный научно-исследовательский институт химии и механики», старший научный сотрудник

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук

Защита состоится «27» февраля 2019 г. в 13:30 на заседании объединенного диссертационного совета Д 999.216.02 при МАИ и МГТУ им. Н. Э. Баумана по адресу: 105005, г. Москва, 2-я Бауманская ул., д. 5, стр. 1, зал Ученого совета ГУК.

С диссертацией можно ознакомиться в библиотеке МГТУ им. Н. Э. Баумана и на сайте http://bmstu.ru.

Отзыв на автореферат в двух экземплярах, заверенных гербовой печатью организации, просим направлять по адресу: 105005, г. Москва, 2-я Бауманская ул., д. 5, стр. 1, ученому секретарю диссертационного совета Д 999.216.02.

Автореферат разослан «__» _____ 201_ года.

Ученый секретарь
диссертационного совета Д 999.216.02
доктор технических наук, доцент А. Н. Алфимцев

Общая характеристика работы

Актуальность темы. Онлайновые социальные сети (ОСС) стали альтернативными СМИ: многие пользователи доверяют информации, полученной из социальных сетей больше, чем традиционным источникам новостей — телевидению и газетам. ОСС являются сложными программными системами обмена, передачи и анализа глобально распределенных данных, в которых циркулирует информация, отражающая события в мире. Для анализа общественного мнения используются системы мониторинга ОСС (СМОСС), которые представляют собой программно-аппаратные комплексы для сбора, извлечения, структуризации, хранения и анализа данных, полученных из ОСС.

При этом большинство существующих СМОСС ориентированы на сбор количественных показателей, таких как количество записей по теме, количество положительных оценок к ним и их эмоциональный окрас. В связи с этим на них можно повлиять извне: при наличии достаточного количества ресурсов злоумышленник может поспособствовать искусственному росту числа сообщений с определенной точкой зрения, манипулируя таким образом общественным мнением. За счет этого в последние года ОСС стали площадкой для информационного противоборства. Для подобных действий используются учетные записи, через которые автоматически или автоматизированно публикуются сообщения, — боты, которых также называют виртуальными пользователями. Другая проблема заключается в анализе аудитории, вовлеченной в обсуждение какой-либо темы. Несмотря на то, что ОСС побуждает пользователей предоставлять достаточно много информации о себе, не все это делают, а доступ к ней может быть частично ограничен настройками приватности. Таким образом, возникают актуальные задачи разработки математического и программного обеспечения для анализа ОСС, а именно обнаружения ботов и определения скрытых или неуказанных атрибутов профилей пользователей ОСС. Решение этих задач будет способствовать увеличению эффективности анализа ОСС и повышению релевантности информации в СМОСС.

Вопросами мониторинга и анализа ОСС занимаются такие отечественные и зарубежные ученые, как Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили, М.А. Басараб, Ю. Лесковец, К. Безносов, Ч. Янг, Я. Бошмаф, Я. Раткевич, К. Ли, Дж. Кейверли, М. Коновер и др. Одной из основных моделей ОСС является граф, вершинами которого являются профили пользователей ОСС, а ребрами

— взаимодействия между ними. При этом особенностью таких графов является их неоднородность: группы вершин, расположенных близко друг к другу, и имеющих много связей между собой, образуют сообщества. В большинстве методов обнаружения ботов, основанных на анализе сообществ, рассматривается общий граф ОСС либо его достаточно большой подграф. Однако Ч. Янг, К. Ли и другие ученые показали, что боты со сложной логикой могут не образовывать сообщества, а устанавливать связи случайным образом. Отличить таких ботов от настоящих пользователей позволит анализ сообществ их локального графа — графа ближайшего окружения (ГБО). Такой анализ также является многообещающим подходом к определению скрытых или неуказанных атрибутов профилей пользователей, однако он пока плохо изучен.

Стоит отметить, что задача выделения сообществ является нетривиальной даже для человека. Существующие алгоритмы дают низкие результаты на выборках из ОСС с известной структурой сообществ. Отдельной проблемой является высокая вычислительная сложность многих алгоритмов, что ограничивает их применение в системах полномасштабного мониторинга ОСС. Кроме того, лишь немногие из них используют атрибуты вершин и позволяют выделять пересекающиеся сообщества. При этом в литературе практически не уделяется внимание возможности частичного отсутствия атрибутов вершин.

Объектом исследования являются онлайновые социальные сети.

Предметом исследования является структура графов ближайшего окружения пользователей ОСС.

Целью работы является разработка методов, алгоритмов и программного обеспечения для анализа ОСС, предназначенных для повышения релевантности информации в СМОСС путем обнаружения ботов и определения скрытых или неуказанных атрибутов профилей ОСС. Разрабатываемые методы и алгоритмы должны сочетать высокое качество работы и низкую вычислительную сложность. В рамках цели были поставлены следующие **задачи**:

1. Разработать алгоритм выделения сообществ ГБО из ОСС, использующий информацию об атрибутах вершин, допускающий пересечение сообществ и устойчивый к частичному отсутствию атрибутов.

2. Разработать метод обнаружения ботов в ОСС, основанный на анализе сообществ ГБО.

3. Разработать метод определения скрытых или неуказанных атрибутов профиля пользователя ОСС по структуре и атрибутам вершин ГБО.

4. Создать программное обеспечение для сбора и анализа данных из ОСС, реализовать программно разработанные методы.

5. Провести экспериментальную оценку качества разработанных методов и алгоритмов на выборках ГБО пользователей из ОСС.

Научная новизна:

1. Впервые для обнаружения ботов разработан метод, основанный на анализе сообществ ГБО из ОСС; его применение позволит эффективно обнаруживать ботов в ОСС.

2. Предложен новый метод определения скрытых или неуказанных атрибутов пользователя ОСС, основанный на выделении сообществ ГБО и, в отличие от существующих, использующий метки сообществ. Применение этого метода позволит существенно повысить точность и полноту автоматического определения скрытых или неуказанных атрибутов профилей пользователей.

3. Разработан алгоритм выделения сообществ ОСС. В отличие от существующих, он одновременно обладает следующими свойствами: использует информацию как о структуре графа ОСС, так и об атрибутах его вершин; может выделять пересекающиеся сообщества; основан на переносе атрибутов; имеет квазилинейную сложность; присваивает полученным сообществам метки, которые могут быть использованы для их описания.

4. Разработано оригинальное ПО, позволяющее исследовать устойчивость алгоритмов выделения сообществ и методов определения атрибутов пользователей ОСС к частичному отсутствию атрибутов.

Практическая значимость полученных результатов заключается в возможности повысить релевантность информации в СМОСС и эффективность процессов обработки данных в ней. При условии внедрения разработанных методов сложность создания правдоподобных ботов ОСС существенно возрастает. Определение атрибутов профилей позволит повысить качество анализа общественного мнения и другой информации из ОСС. Помимо этого, разработанный алгоритм выделения сообществ является универсальным и может быть использован для решения других прикладных задач в сетях с атрибутами вершин. Практическую ценность также представляет разработанное ПО для распределенного сбора и анализа данных из ОСС.

Методы исследования. Для решения поставленных задач использованы методы теории графов, теории анализа социальных сетей, математической статистики, инженерии ПО, а также методы теории сложности. В качестве языков программирования в работе были использованы Python 2.7 и C++11.

Основные положения и результаты, выносимые на защиту:

1. Квазилинейный алгоритм выделения пересекающихся сообществ в социальных сетях с атрибутами, основанный на переносе атрибутов от соседних вершин, устойчивый к их частичному отсутствию и ставящий в соответствие выделенным сообществам набор атрибутов (метку). Предложенный алгоритм превосходит аналоги по значениям F_1-меры и индекса Жаккара на выборках данных из ОСС Facebook и Twitter, что подтверждено эмпирически с помощью разработанного ПО.

2. Метод обнаружения ботов в ОСС, основанный на качественном анализе сообществ ГБО из ОСС. Характеристики сообществ в сочетании с набором из нескольких пороговых правил позволяют обнаружить ботов с высокими показателями точности и полноты.

3. Метод определения скрытых или неуказанных атрибутов профиля пользователя ОСС, основанный на анализе меток сообществ его ГБО. Объединение множеств атрибутов меток сообществ позволяет определять некоторые атрибуты профиля пользователя с высокой точностью и полнотой, что подтверждается вычислительным экспериментом с разработанным ПО.

Соответствие паспорту научной специальности. В диссертации разработаны методы и алгоритмы анализа ОСС и ее модели в виде социального графа. Предложенные методы и алгоритмы работают с его подграфами, ГБО, и реализованы программно; а для организации их взаимодействия с ОСС разработано ПО. Таким образом, результаты диссертации соответствуют пунктам 1 и 3 паспорта научной специальности 05.13.11.

Достоверность полученных результатов обеспечивается корректным применением математического аппарата и подтверждается результатами вычислительных экспериментов с использованием разработанного ПО, проведенных на больших выборках данных из ОСС с размеченными истинными данными.

Апробация работы. Основные результаты работы представлены автором на Шестой, Седьмой и Восьмой Всероссийской научно-технической кон-

ференции «Безопасные информационные технологии» (г. Москва, 2015, 2016 и 2017 гг.); XXIII Международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов-2016» (г. Москва, 2016 г.); The 6th International Conference on Network Analysis «NET 2016» (г. Нижний Новгород, 2016 г.) и обсуждены на научных семинарах в МГТУ им. Н.Э. Баумана, Институте проблем управления им. В.А. Трапезникова РАН и НИЯУ МИФИ.

Внедрение результатов работы. Разработанные автором методы и алгоритмы были использованы в научно-производственной деятельности ООО «ИНФОРИОН» в рамках проекта мониторинга социальных сетей «Система-47» и были внедрены в НИОКР ФГУП «18 ЦНИИ» МО РФ. Теоретические результаты внедрены в учебный процесс МГТУ им. Н.Э. Баумана.

Публикации и личный вклад автора. Основные результаты изложены в 12 статьях, 5 из которых изданы в журналах, рекомендованных ВАК, а две — в изданиях, индексируемых международной системой научного цитирования Scopus. Все основные результаты получены автором лично.

Диссертационная работа выполнена при поддержке гранта РФФИ № 16-29-09517 офи_м «Методы и алгоритмы выявления сообществ и организации информационного противоборства в социальных сетях на основе байесовских и теоретико-игровых подходов с использованием графовых и фрактальных моделей».

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и 10 приложений. Полный объем диссертации составляет 200 страниц текста с 22 рисунками и 20 таблицами. Объем приложений составляет 55 страниц. Список литературы содержит 191 наименование.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, аргументированы научная новизна и практическая значимость полученных результатов, представлены выносимые на защиту научные положения и результаты.

Первая глава содержит систематический обзор предметной области и основные определения, используемые в работе.

Информация в СМОСС может быть искажена путем массовой публикации сообщений (Рис. 1). Одним из основных способов массовой публикации заведомо ложных сообщений является использование ботов. В ОСС бот (виртуальный пользователь) — это учетная запись (аккаунт), связанная с программой, выполняющей некоторые действия через интерфейс ОСС автоматически и/или по заданию. Боты используются в основном для действий, приносящих прибыль своим создателям и нарушающих правила пользования ОСС. Разработка универсального метода обнаружения ботов, по крайней мере в ближайшее время, не представляется возможной, поскольку боты, управляемые человеком, могут быть практически неотличимы от реальных пользователей. Подавляющее большинство исследователей используют комбинированный метод для обнаружения ботов, который зачастую основан на машинном обучении. Исследователи отмечают, что многие боты устанавливают связи случайным образом. При этом социальные графы обычных пользователей имеют неоднородную структуру. Классификация признаков, по которым обнаруживают ботов, представлена на Рис. 2.

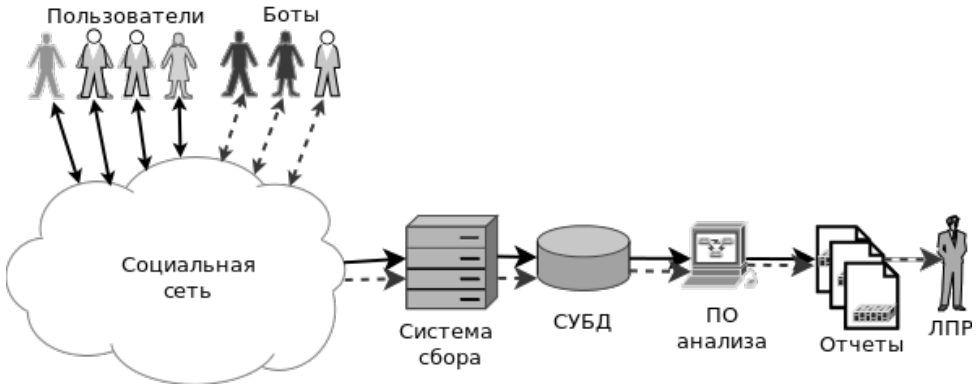


Рис. 1. Схема СМОСС. Штриховой линией обозначен путь поступления дезинформации к лицу, принимающему решения (ЛПР).

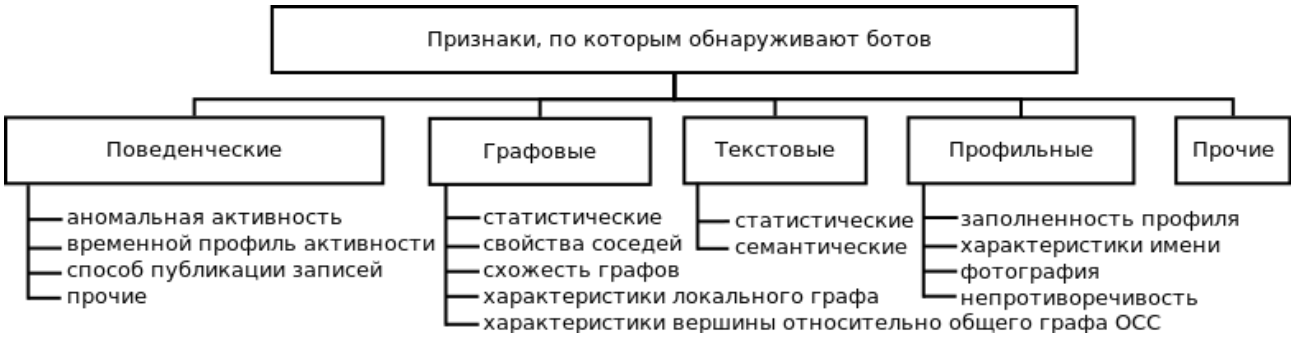


Рис. 2. Классификация признаков, по которым обнаруживают ботов.

Рассмотрена проблема частичного отсутствия атрибутов профиля пользователя ОСС. Обозначим множество всех возможных атрибутов как F. Определим функцию $f' : V \to 2^F$ получения доступных атрибутов вершины такую, что $f'(v) \subseteq f(v)$ для любой вершины $v \in V$ графа $G(V, E)$, где $f : V \to 2^F$ — функция получения всех атрибутов (недостижима на практике). Задача определения скрытых или неуказанных атрибутов профиля заключается в нахождении такого $p \subseteq F$, что $\delta(p, f(v)) \to \max$, где δ — некоторая мера схожести двух множеств. Большинство методов решения этой задачи по общедоступной информации можно отнести к одному из четырех подходов: перенос атрибутов соседних вершин путем простого голосования или по другому правилу; методы, основанные на машинном обучении; определение атрибутов по предпочтениям пользователя; методы, основанные на выделении сообществ. При этом в многих исследованиях не учитывается возможность частичного отсутствия информации об атрибутах других пользователей. Обзор литературы показал, что существующие методы либо имеют высокую вычислительную сложность, либо требуют большого объема труднодоступных данных.

По результатам анализа научных работ выявлено, что как для решения задачи определения ботов, так и для определения скрытых или неуказанных атрибутов профиля ОСС перспективны методы, основанные на анализе структуры социального графа, в частности, сообществ ГБО. Выделены основные свойства, которыми должен обладать алгоритм выделения сообществ в ОСС:

– использует как информацию об атрибутах вершин, так и информацию о ребрах;

– полученные сообщества могут пересекаться;

– имеет низкую вычислительную сложность;

– устойчив к частичному отсутствию атрибутов.

Выявлено, что исследования касательно последнего свойства не проводились. Кроме того, в литературе практически не уделено внимание проблеме автоматического получения описания сообществ (меток).

Рассмотрена проблема оценки качества полученных сообществ. На данный момент класс мер, основанных на степени схожести полученного покрытия и эталонного, считается более достоверным, поскольку такой подход позволяет проверить способность различных алгоритмов обнаружить структуру сообществ, заданную особым способом, который зависит от источника

данных или исследовательской задачи. Поскольку наборы данных в открытом доступе не всегда покрывают потребности исследования, возникает задача сбора данных из ОСС; в конце главы рассмотрены основные проблемы и подходы к решению этой задачи.

Вторая глава посвящена разработке методов анализа ГБО.

Предложен **алгоритм выделения сообществ**, который базируется на трех основных предположениях: триадной структуре социальных сетей, модели присоединения («*affiliation*») и наличии общих атрибутов у соседних вершин (гомофилия). Алгоритм состоит из пяти этапов.

Первый этап — получение ключевых атрибутов. Изначально для всех вершин $v \in V$ графа $G(V, E)$ множество ключевых атрибутов K_v пусто. На каждой итерации данного этапа для всех вершин v определяются множества их соседей $\mathcal{N}_v$ и для каждого атрибута $a \in F$ определяются множества

$$\mathcal{N}'_{v,a} = \{w | w \in \mathcal{N}_v \wedge a \in f'(w)\}, \tag{1}$$

$$Q_{v,a} = \{w | w \in \mathcal{N}_v \wedge a \in K_w\}. \tag{2}$$

Если сумма мощностей данных множеств превышает порог $\max(2, \alpha_a|\mathcal{N}_v|))$, где α_a — доля, определяющая большинство для атрибута a, то тогда в множество ключевых атрибутов K_v добавляется атрибут a. Этап завершается, когда на последней итерации не было изменений. *Второй этап* состоит в определении ключевых атрибутов для вершин на пересечении сообществ. В нем ослаблен порог для условия добавления атрибута в ключевые:

$$\max(2, \alpha_a \frac{|\mathcal{N}_v|}{|\bigcup_{w \in \mathcal{N}_v} K_w \setminus K_v|}). \tag{3}$$

На третьем этапе вершины объединяются в сообщества по ключевым атрибутам и разбиваются на компоненты связности. В *четвертом этапе* сообщества, множества вершин которых совпадают, объединяются в одно, которому сопоставляется набор атрибутов всех сообществ — метка. На *пятом этапе* происходит обнаружение сообществ, которых объединяет атрибут, не указанный ни у одной вершины, путем применения к графу модифицированных этапов 1 и 3.

Доли, определяющие большинство для атрибутов $a \in F$, могут быть заданы эвристикой, основанной на свойствах исследуемого графа, или получены в результате машинного обучения на тестовой выборке. Простейший вариант заключается в приравнивании всех коэффициентов к одному значению.

Разработанный алгоритм имеет квазилинейную сложность относительно размера графа и может быть распараллелен. Пример результата работы алгоритма представлен на Рис. 3. Тот же граф с сообществами, размеченными вручную, представлен на Рис. 4.

Также предложен вариант алгоритма для кластеризации графа.

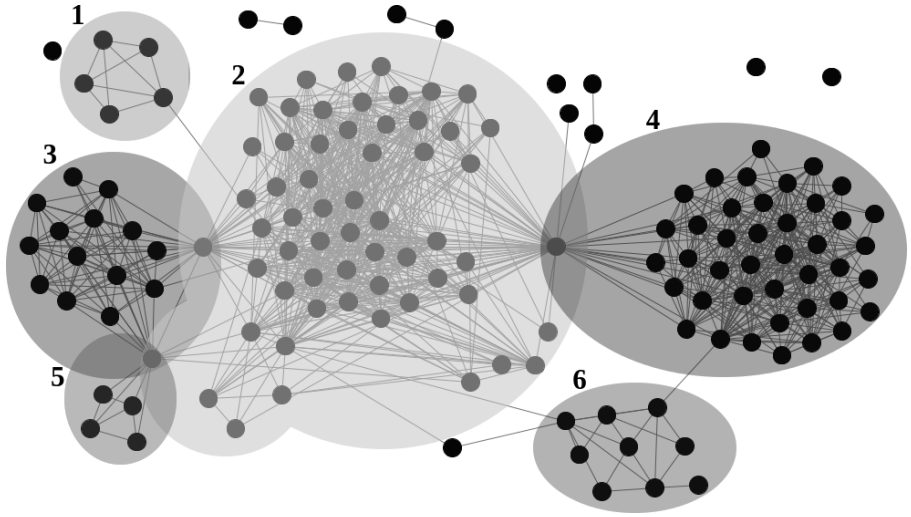


Рис. 3. Пример работы предложенного алгоритма выделения сообществ.

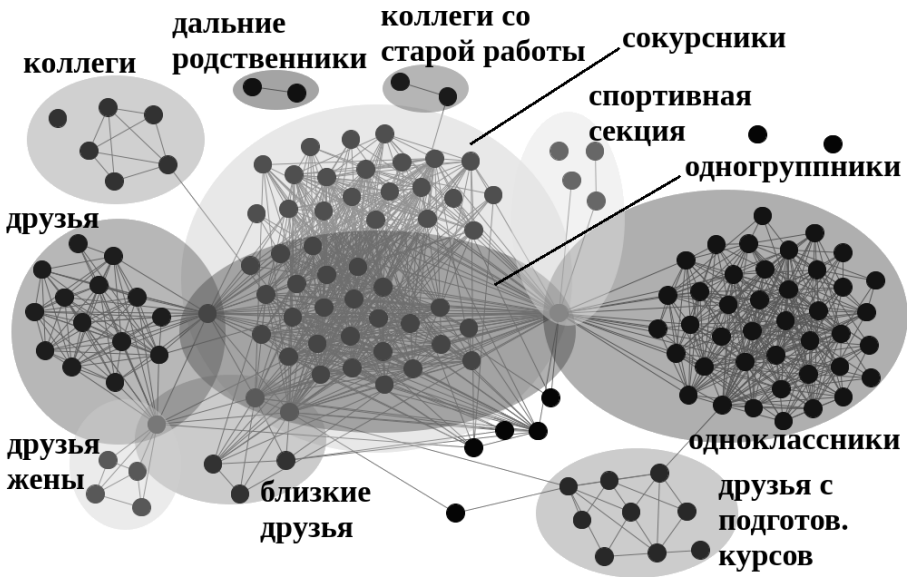


Рис. 4. Пример графа с вручную размеченными сообществами.

На основе разработанного алгоритма предложен **метод обнаружения ботов** в ОСС. Выдвинуто предположение, что структуры ГБО реальных людей и ботов будут значительно отличаться. Рассмотрены характеристики сообществ графа. Предлагаемый метод обнаружения ботов основан на следующих пороговых правилах:

1. Сообщества размером меньше минимального $S_{\min}$ или больше максимального $S_{\max}$ не учитываются. Если количество таких сообществ Q' превышает пороговое значение $Q'_{\max}$, то пользователь считается ботом.

2. Если количество друзей в сообществах B больше максимального $B_{\max}$ или меньше минимального $B_{\min}$, то пользователь — бот.

3. Если отношение количества друзей пользователя вне сообществ B' и в сообществах B больше порога $R_{\max}$, то пользователь считается ботом.

4. Если количество сообществ Q больше максимального $Q_{\max}$ или меньше минимального $Q_{\min}$, то пользователь считается ботом.

5. Иначе пользователь считается легитимным.

Предлагаемый **метод определения профиля** p центральной вершины u заключается в объединении всех меток сообществ, являющихся результатом работы предложенного алгоритма выделения сообществ: $p = \bigcup_i A_i$, где $A_i \subseteq F$ — метка сообщества C_i. Представлен упрощенный алгоритм выделения сообществ, результатом которого будет только набор ключевых атрибутов.

Как показали эксперименты, не все атрибуты профиля могут быть определены на основе информации о структуре графа и атрибутах вершин, поэтому такие атрибуты должны быть удалены из результатов анализа.

Третья глава посвящена созданию программных реализаций разработанных методов анализа, а также системе сбора данных и вспомогательному ПО. Рассмотрены основные методы организации взаимодействия ОСС и СМОСС.

Общая архитектура разработанной программной системы представлена на Рис. 5. Реализована гибкая, масштабируемая и расширяемая система сбора и анализа данных из ОСС. Для добавления обработки новой ОСС достаточно создать один или несколько классов для программы сбора. Схема потоков данных в системе представлена на Рис. 6. Система распределения задач функционирует на базе СУБД типа «ключ-значение» Redis и основана на списках задач с параметром глубины, что позволяет эффективно собирать ГБО пользователей ОСС. Система хранения данных реализована на базе реляционной СУБД PostgreSQL. Для экспорта ГБО из базы данных в текстовый формат было разработано вспомогательное программное обеспечение.

Все методы, представленные во второй главе, реализованы программно. Кроме того, создано ПО, реализующее методы определения скрытых или

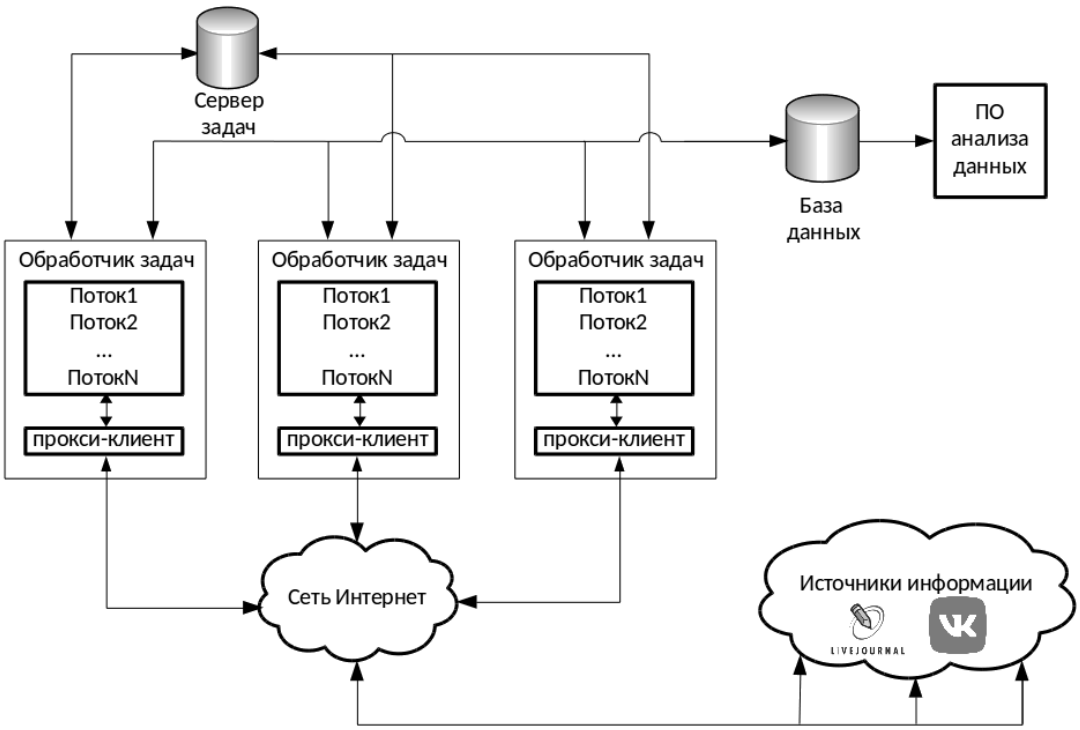


Рис. 5. Общая архитектура разработанной программной системы.

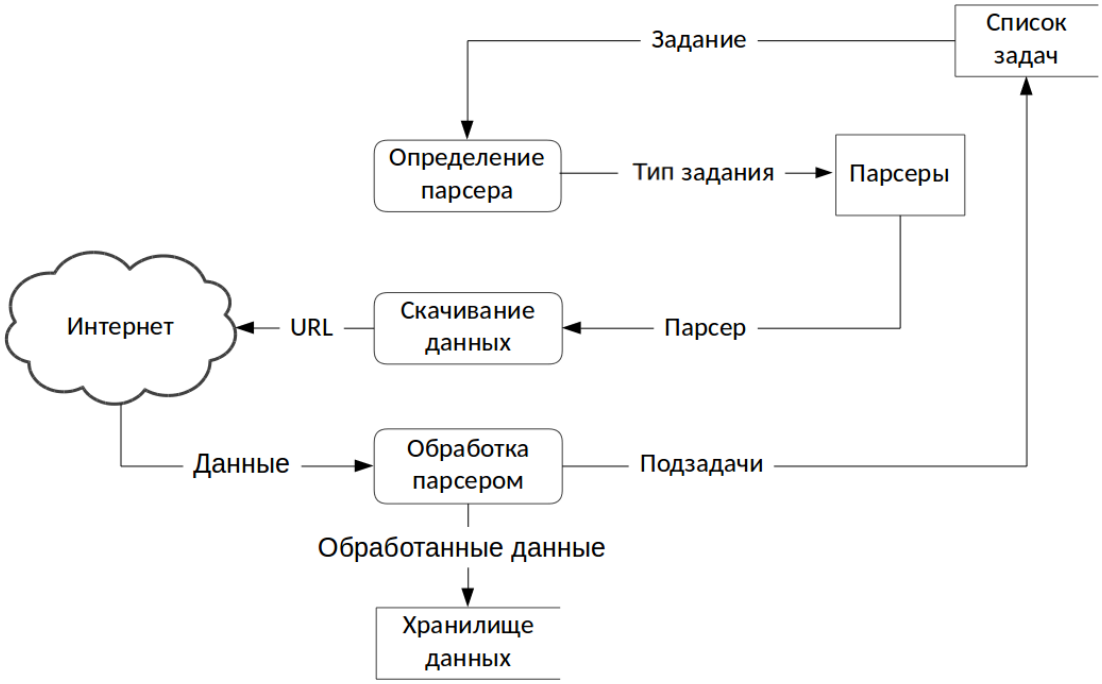


Рис. 6. Схема потоков данных в программе сбора ГБО ОСС.

неуказанных атрибутов профиля: метод простого большинства, метод на основе выделения сообществ, метод на основе весов атрибут-сообщество.

С помощью созданного ПО получены выборки более 4300 ГБО пользователей из ОСС LiveJournal и ВКонтакте, которые были использованы для экспериментальной оценки качества разработанных методов и алгоритмов в **четвертой главе**.

Для оценки качества разработанного алгоритма выделения сообществ использовалась мера схожести покрытий, предложенная Янгом и Лесковицем:

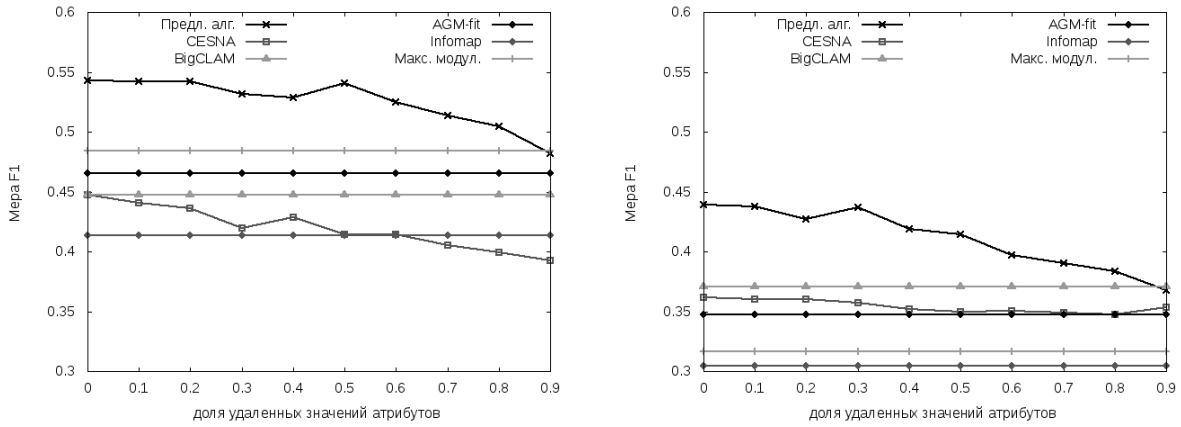
$$\mathcal{Q} = \frac{1}{2|\mathscr{C}|}\sum_{C_i\in\mathscr{C}}\max_{C_j^*\in\mathscr{C}^*}\delta(C_j^*, C_i) + \frac{1}{2|\mathscr{C}^*|}\sum_{C_j^*\in\mathscr{C}^*}\max_{C_i\in\mathscr{C}}\delta(C_j^*, C_i), \tag{4}$$

где $\mathscr{C}$ — полученное покрытие, $\mathscr{C}^*$ — эталонное, а в качестве меры схожести двух сообществ, $\delta(C_i^*, C_j)$, использовались мера F_1 и коэффициент Жаккара:

$$F_1 = \frac{2\cdot|C_j^*\cap C_i|}{|C_j^*|+|C_i|}, \quad J = \frac{|C_j^*\cap C_i|}{|C_j^*\cup C_i|}. \tag{5}$$

Сравнение производилось с алгоритмами Infomap, максимизации модулярности, AGM-fit, BigCLAM и CESNA на выборках ГБО из ОСС Facebook и Twitter от Stanford Network Analysis Project. Кроме того, были использованы несколько хорошо изученных графов с известными сообществами и граф автора из ОСС ВКонтакте. Результаты оценки качества работы алгоритмов представлены в Таблице 1. По мере F_1 и индексу Жаккара разработанный алгоритм превосходит остальные в среднем на 12–41%.

Анализ полученных сообществ на примере графа автора (см. Рис. 3 и 4) показал, что получаемые метки сообществ информативны и могут быть использованы для их автоматического описания. Эксперименты (Рис. 7) показали, что алгоритм устойчив к отсутствию до 50% значений атрибутов и позволяет выделять сообщества, которые образованы неуказанными атрибутами. Сравнение скорости работы алгоритмов показало, что предложенный алгоритм уступает только алгоритму максимизации модулярности, и имеет квазилинейную сложность.



а) Facebook б) Twitter

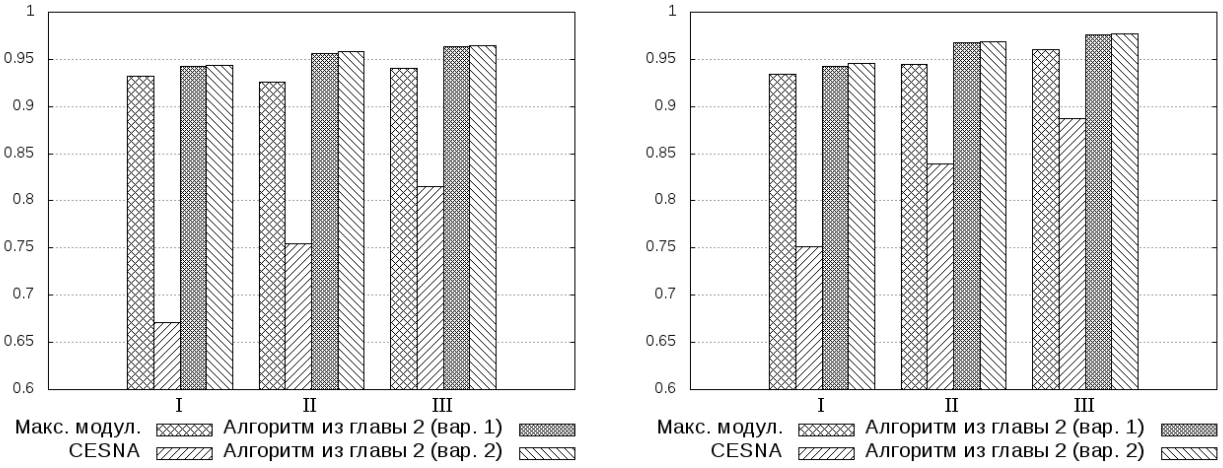
Рис. 7. Среднее значение меры F_1 в зависимости от количества отсутствующих значений атрибутов.

Таблица 1.

Оценка качества работы алгоритмов выделения сообществ на выборках ГБО из ОСС Facebook и Twitter. Средние значения мер.

	F_1-мера		Индекс Жаккара	
	Facebook	Twitter	Facebook	Twitter
Infomap	0.414	0.304	0.312	0.222
Максимизация модулярности	0.484	0.316	0.383	0.228
AGM-fit	0.466	0.347	0.366	0.251
BigCLAM	0.447	0.371	0.325	0.270
CESNA	0.447	0.362	0.339	0.264
Предложенный алгоритм	0.543	0.439	0.442	0.342

Подтверждена гипотеза, что ГБО ботов и легитимных пользователей существенно отличаются. Для проверки использованы выборки из ОСС LiveJournal. В качестве алгоритмов выделения сообществ выбраны максимизация модулярности, CESNA и разработанный во второй главе алгоритм. При обработке сообществ, полученных последним, использованы два варианта: обычный и тот, в котором исключены сообщества с пустой меткой. Пороговые значения были получены автоматически. Результаты проверки (Рис. 8) показали, что алгоритм CESNA не подходит в качестве алгоритма выделения сообществ в этой задаче. Для двух других алгоритмов метод показал высокие значения доли правильных ответов, точности, полноты и меры F_1. Вариант алгоритма, в котором отбрасываются сообщества с пустой меткой, дает небольшой прирост значений мер качества.



а) Доля правильных ответов б) Мера F_1

Рис. 8. Качество определения ботов. I — управляемые боты, II — автоматические боты, III — все боты.

Для оценки качества метода определения неизвестных атрибутов профиля были использованы ГБО из Facebook, Twitter, ВКонтакте и LiveJournal. Он был сравнен с известными методами, основанными на простом большинстве, анализе сообществ и весах атрибут-сообщество. Множества атрибутов были отфильтрованы: атрибуты, которые невозможно определить, были отброшены. На выборке из ОСС Facebook разработанный метод дает одни из наилучших значений меры F_1, а на ГБО из Twitter, ВКонтакте и LiveJournal — превосходит другие методы на 14,5%, 12% и 4% (Таблица 2). При этом он позволяет определять некоторые атрибуты профиля с точностью и полнотой, близкими к единице (Таблица 3). Метод показал высокую устойчивость к частичному отсутствию атрибутов (Рис. 9): при удалении 90% значений атрибутов качество его работы на выборке из Facebook падает лишь на 20%, а на других возрастает за счет снижения избыточности получаемых профилей.

Таблица 2.

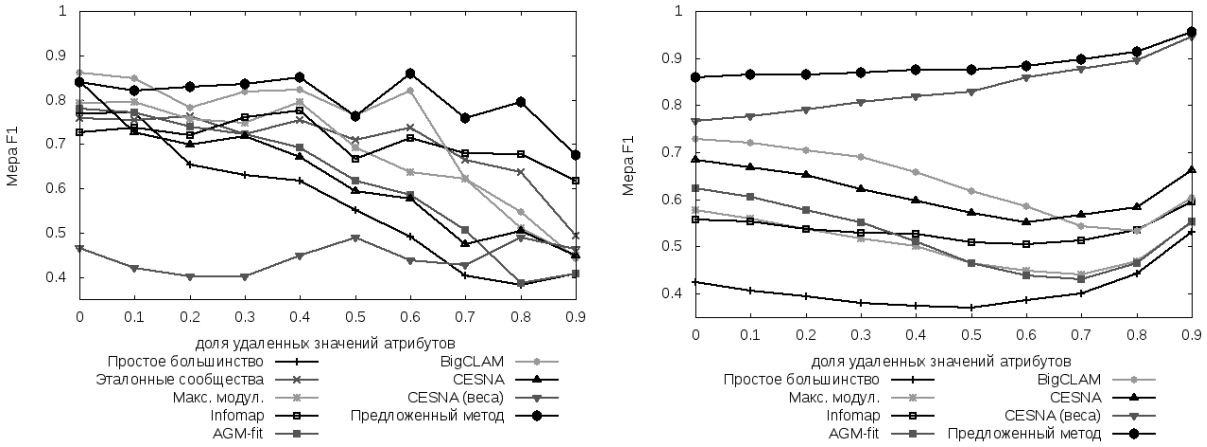
Определение атрибутов профилей. Средние значения меры F_1.

	Facebook	Twitter	ВКонтакте	LiveJournal
Простое большинство	0.769	0.642	0.424	0.678
Эталонные сообщества	0.759	0.756	н/д	н/д
Infomap	0.726	0.641	0.557	0.504
Макс. модул.	0.793	0.675	0.578	0.625
AGM-fit	0.780	0.669	0.624	0.668
BigCLAM	0.860	0.720	0.728	0.750
CESNA	0.842	0.719	0.683	0.727
CESNA (веса)	0.466	0.752	0.767	0.874
Предложенный метод	0.841	0.866	0.860	0.908

Таблица 3.

Результаты определения некоторых атрибутов профилей. Средние значения меры F_1, точности и полноты.

Набор данных	Атрибут	F_1	Точность	Полнота
Facebook	образовательное учреждение	0.902	0.967	0.875
Facebook	родной город	1.000	1.000	1.000
ВКонтакте	средняя школа	0.919	0.910	0.973
ВКонтакте	факультет (кафедра)	0.994	0.995	0.996
ВКонтакте	место работы	0.994	0.997	0.994
LiveJournal	образовательное учреждение	0.975	0.991	0.970
LiveJournal	родной город	0.984	0.977	1.000



а) Facebook б) ВКонтакте

Рис. 9. Среднее значение меры F_1 для выборок Facebook и ВКонтакте в зависимости от количества отсутствующих значений атрибутов.

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

1. Проведен анализ существующих методов выявления ботов и определения скрытых или неуказанных атрибутов профилей пользователей в ОСС, основанных на теории графов; выявлены их недостатки для решения поставленной в работе задачи. Обоснована необходимость разработки новых методов анализа ГБО пользователей ОСС.

2. Разработан быстрый и масштабируемый алгоритм выделения пересекающихся сообществ в социальных сетях с атрибутами вершин, не требующий сведений о природе атрибутов и устойчивый к их частичному отсутствию. Для каждого сообщества алгоритм предоставляет список атрибутов, которые его формируют — метку. Вычислительный эксперимент показал, что алгоритм превосходит аналоги по F_1-мере и индексу Жаккара на трех тестовых выборках в среднем на 12–41%.

3. Предложен метод выявления ботов, основанный на качественном анализе сообществ ГБО пользователя. Метод был опробован на двух выборках ботов из ОСС LiveJournal: управляемых и автоматических ботов. Эксперименты показали высокие значения точности, полноты и меры F_1 обнаружения ботов. Применение разработанного метода может существенно повысить сложность создания правдоподобных аккаунтов ботов для злоумышленника.

4. Разработан метод определения неуказанных или скрытых атрибутов пользователя путем анализа структуры ГБО и атрибутов вершин этого гра-

фа. Метод показал высокие значения F-меры, точности и полноты по определению отдельных атрибутов профиля, таких как родной город или место обучения пользователя.

5. Создано программное обеспечение для сбора данных из ОСС. Собраны тестовые выборки ГБО пользователей из ОСС ВКонтакте и LiveJournal, которые были использованы для экспериментальной оценки качества разработанных методов и алгоритмов анализа ГБО из ОСС.

6. Разработано оригинальное ПО, реализующее предложенные в работе методы и алгоритмы анализа ОСС, которое было использовано для их экспериментальной оценки.

Таким образом, все поставленные в работе задачи решены, а цель — достигнута. Внедрение результатов настоящей работы будет способствовать повышению релевантности информации и увеличению эффективности обработки данных в СМОСС и других системах обработки информации из ОСС.

Публикации автора по теме диссертации

В рецензируемых изданиях из перечня ВАК Минобрнауки РФ:

1. *Ключарёв П. Г.*, *Чесноков В. О.* Исследование спектральных свойств социального графа сети LiveJournal // Наука и образование. Электронное научно-техническое издание. 2013. № 9. С. 391—400. (0,69 п.л./0,41 п.л.)
2. *Чесноков В. О.*, *Ключарёв П. Г.* Выделение сообществ в социальных графах по множеству признаков с частичной информацией // Наука и образование. Электронное научно-техническое издание. 2015. № 9. С. 188—199. (0,85 п.л./0,72 п.л.)
3. *Чесноков В. О.* Предсказание атрибутов профиля пользователя социальной сети путем анализа сообществ графа его ближайшего окружения // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2017. № 2. С. 66—76. (0,91 п.л.)
4. *Чесноков В. О.* Применение алгоритма выделения сообществ в информационном противоборстве в социальных сетях // Вопросы кибербезопасности. 2017. № 1(19). С. 37—44. (0,61 п.л.)
5. *Чесноков В. О.*, *Ключарёв П. Г.* Современные методы выделения сообществ в социальных сетях // Наука и образование. Электронное научно-техническое издание. 2017. № 4. С. 137—152. (1,16 п.л./1,01 п.л.)

В журналах, входящих в базу цитирования Scopus:

6. *Chesnokov V.* Overlapping Community Detection in Social Networks with Node Attributes by Neighborhood Influence // Models, Algorithms, and Technologies for Network Analysis: NET 2016, Nizhny Novgorod, Russia, May 2016. Cham: Springer International Publishing, 2017. P. 187–203. (1,53 п.л.)

7. *Chesnokov V.*, *Klyucharev P.* Deanonymizing Users in Social Networking Services: an Ego-Network Analysis Approach // CEUR Workshop Proceedings. Selected Papers of the VIII All-Russian Scientific and Technical Conference on Secure Information Technologies (BIT 2017). Vol. 2081. 2017. P. 40–44. (0,51 п.л./0,42 п.л.)

В других изданиях:

8. *Чесноков В. О.* Выделение сообществ в графах по множеству признаков с частичной информацией // «Безопасные информационные технологии»: Сборник трудов Шестой Всероссийской научно-технической конференции. М.: Изд-во Научно-учебный комплекс «Информатика и системы управления», 2015. С. 18—21. (0,23 п.л.)
9. *Чесноков В. О.* Выделение пересекающихся сообществ в социальных графах по мажоритарному признаку соседей // «Ломоносов 2016»: XXIII Международная научная конференция студентов, аспирантов и молодых ученых; секция «Вычислительная математика и кибернетика». М.: Издательский отдел факультета ВМиК МГУ; МАКС Пресс, 2016. С. 49—51. (0,24 п.л.)
10. *Чесноков В. О.* Обнаружение виртуальных пользователей в онлайновых социальных сетях путем анализа графов ближайшего окружения // «Безопасные информационные технологии»: Сборник трудов Седьмой всероссийской научно-технической конференции. М.: НУК «Информатика и системы управления», 2016. С. 299—302. (0,25 п.л.)
11. *Чесноков В. О.* Деанонимизация пользователей онлайновых социальных сетей путем анализа графов ближайшего окружения // Безопасные информационные технологии. Сборник трудов Восьмой всероссийской научно-технической конференции. М.: МГТУ им. Н.Э. Баумана, 2017. С. 513—516. (0,25 п.л.)
12. *Чесноков В. О.* Программное обеспечение сбора и анализа графов ближайшего окружения из онлайновых социальных сетей // Машиностроение и компьютерные технологии. 2018. № 8. С. 34—44. (0,74 п.л.)