МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра математической теории упругости и биомеханики

Автоматизация профилирования пользователей социальных сетей на примере социальной сети "ВКонтакте"

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 442 группы

направления 09.03.03 «Прикладная информатика (прикладной бакалавриат)»

механико-математического факультета

Головашова Антона Юрьевича

Научный руководитель
доцент, к.ф.-м.н. _____ Л.В. Бессонов
подпись, дата

Зав. кафедрой
д.ф.-м.н., профессор _____ Л.Ю. Коссович
подпись, дата

Саратов 2019

ВВЕДЕНИЕ

Социальные сети (от англ. social networking service) представляют собой платформу, обеспечивающую построение и организацию социальных взаимоотношений. Созданные для общения и обмена информацией между людьми, социальные сети ежедневно генерируют огромный поток информации. На сегодняшний день существует огромное количество различных социальных сетей, например: Facebook, Twitter, ВКонтакте, Мой Мир, Instagram, Одноклассники, MySpace, LiveJournal, и другие. С каждым годом число пользователей социальных сетей стабильно растет, меняется структура аудитории пользователей по возрастным, демографическим, гендерным и другим критериям. По данным TNS Web Index за март 2019 [1], в топ-5 самых популярных интернет ресурсов в России по показателю среднесуточной аудитории вошли социальные сети «Вконтакте» (аудитория 546 млн человек) и «Одноклассники» (350 млн человек). Все социальные сети отличаются своими целевыми аудиториями, спецификой взаимодействия, принципами формирования сообществ и связей между пользователями. Структура, представляющая собой визуализацию социальной сети, в которой ключевыми элементами являются люди и связи между ними, называется социальным графом. На сегодняшний день во всех социальных сетях, как правило, круг контактов конкретного пользователя составляют те, с кем он взаимодействует и в жизни (одноклассники, друзья, коллеги по работе, родственники и т.д.). Получается, что социальные сети зачастую не являются источником формирования взаимосвязей пользователей, так как пользователи уже установили контакт в условиях реальных обстоятельств, а уже после этого начали взаимодействовать в социальной сети. Однако в реальной жизни, если человек стремится к обретению нового опыта или новых знаний, он, скорее всего, обращается к новым людям, ищет новые связи. Поэтому социальные сети преследуют новую цель – обеспечить создание взаимосвязей пользователей по интересам в пределах социальной сети. Пользователи, которые не знакомы в жизни и не

связаны социальными обстоятельствами, начинают взаимодействовать в социальной сети, обсуждая интересные им темы. В связи с этим возникает новое понятие граф интересов [2], расширяющее понятие социального графа. Граф интересов описывает структуру, состоящую из людей, интересов и связей между ними. Если в социальном графе существует только один тип связи «человек-человек», то в графе интересов таких типов связей три: «человек-человек», «человекинтерес», «интерес-интерес». Необходимо отметить, что социальные сети представляют особый интерес в том числе и для субъектов бизнеса. Появляется возможность продвигать свой товар, определяя конкретную целевую аудиторию и воздействуя на нее. В данном контексте, целесообразно говорить о необходимости анализа интересов пользователей на основе информации, которую они сами предоставляют. Однако сложность заключается в том, что зачастую пользователь либо не предоставляет о себе информацию, либо предоставляет ложные данные. Так, например, согласно исследованию Всероссийского центра изучения общественного мнения (ВЦИОМ) [3], сообщать о себе в социальных сетях и блогах недостоверную информацию хотя бы однажды приходилось половине пользователей этих ресурсов (51%), причем наиболее часто искажается информация об имени и возрасте (по 29%), о хобби (22%), о половой принадлежности пользователя, музыкальных и художественных пристрастиях (по 18%). Это значительно усложняет задачу анализа интересов пользователей. В этом контексте возникает проблема автоматического определения интересов пользователей на основании доступной информации. Задача моделирования интересов пользователя состоит в нахождении тем, которые интересны пользователю. В общем случае из доступной информации имеются только сообщения пользователя и структура его социальных связей (друзья, подписчики). Для автоматизации профилирования пользователей социальной сети "ВКонтакте" необходим сбор данных, связанных с доступной информацией, такой как: основная информация о владельце страницы, геотеги фотографий, которые

выкладывает пользователь, сбор информации на основе сообществ пользователя. При этом необходимо обобщить всю абстрактную информацию о «профиле пользователя» или «модели пользователя», которые включают в себя основную характеристику пользователя и данные о поведении пользователя. С точки зрения данных, пользователь является ключевым источником получения мета-данных [4].

Целью данной дипломной работы будет являться исследование и разработка приложения для автоматизации методов профилирования пользователей социальных сетей на примере социальной сети "ВКонтакте".

Для достижения заявленной цели необходимо реализовать следующие задачи:

1. Исследовать существующие подходы к автоматизации профилирования пользователей социальных сетей.

2. Разработать метод автоматизации профилирования пользователей социальных сетей.

3. Реализовать выбранный метод с помощью приложения, способного собирать данные о пользователе социальной сети.

Структура и объем работы. Бакалаврская часть состоит из введения, 3 глав, заключения и списка используемых источников, включающего 26 наименований. Работа изложена на 42 листах машинописного текста без приложений, содержит 7 рисунков.

Основное содержание работы. Первая глава ВКР состоит из двух разделов. В ней были рассмотрены основные понятия, возможности и сферы использования профилирования пользователей социальных сетей.

Первым этапом процесса сбора данных о пользователе является фаза предварительной обработки информации. Фаза подготовки данных может быть разделена на две фазы:

1. получение данных из социальной сети;
2. подготовка данных;

Для того, чтобы упорядочить данные необходимо произвести их очистку. Набор данных необходимо отфильтровать от записей, генерируемых автоматически совместно с загрузкой страницы.

Удаление записей, не отражающих активность пользователя. Веб-боты в автоматическом режиме просматривают множество различных страниц в сети. Их поведение сильно отличается от человеческого, и они не представляют интереса с точки зрения анализа использования веб-ресурсов.

Сбор информации о каждом пользователе будет подразумевать набор упорядоченных данных о зарегистрированных в социальной сети пользователей. Можно применять информацию о зарегистрированных пользователях, по средствам cookie-файлов для определения предпочтений каждого пользователя.

Каждый идентифицированный пользователь при каждом визите оставляет о себе данные, содержащие перечень просмотренных страниц, посещенных сообществ, а так же профилей других участников выбранной социальной сети. Также система пытается оценить, когда пользователь покинул социальную сеть. Первая проблема, как правило, побочный эффект посреднических прокси устройств и локальных сетевых шлюзов. Кроме того,

многие пользователи могут иметь доступ к одному компьютеру. Вторая проблема возникает, когда провайдер выполняет балансировку нагрузки используя несколько прокси-серверов. Другим средством хорошей идентификации пользователя является назначение пользователям имени пользователей и пароля [5].

Еще одной проблемой, с которой можно столкнуться при сборе информации о конкретном человеке является нахождение полного пути. Множество людей используют кнопку "назад" для возвращения к ранее просмотренной странице. Если это происходит, то браузер отображает страницу, ранее сохраненную в кэше. Это приводит к "дырам" в журнале веб-сервера. Знания топологии веб-сайта могут быть использованы для восстановления таких пропусков.

Для решения этой проблемы используется идентификация транзакции. Страницы, которые пользователь посещает в течение сеанса могут быть классифицированы в качестве вспомогательных или содержательных (страниц с контентом) страниц. Вспомогательные страницы используются для навигации, то есть пользователь не заинтересован в содержании, а лишь пытается переходить от одной страницы к другой. Содержательные страницы обеспечивают пользователя полезным содержанием. Процесс генерации транзакции, как правило, пытается определить различие между вспомогательными страницами и страницами содержания, чтобы провести независимо друг от друга так называемые вспомогательные сделки (состоящие из вспомогательных страниц и в том числе первой страницы содержания) и контент-сделки (состоящий только из содержательных страниц) [6].

Большой объем информации приносит ряд проблем пользователю, а также затрудняет возможность автоматизации профилирования. Представленная информация зачастую является произвольной смесью текста, речи, изображений и видео, объединенной в один документ и распределенной по разным частям социальной сети. Дополнительной

проблемой является разная целевая аудитория и сообщества, созданные на основе одних и тех же данных. При рассмотрении автоматизации профилирования пользователей социальной сети необходимо учитывать некоторые подходы к автоматическому анализу информации на основе профиля пользователя [7].

Последние исследования предлагают новые решения, помогая пользователям принять правильное и быстрое решение в выборе информации, в которой он заинтересован. Некоторые из аспектов интеллектуального анализа данных включают в разработку моделей для распознавания метаданных фотографий, фраз, лингвистических и грамматических свойств текста, а также извлечения информации из больших объемов данных.

Одним из первых в управлении данными рассматривается вопрос о представлении данных. Часто используется векторное представление, где все геоданные, собранные из фотографий пользователя, сохраняются игнорируя порядок геометок или их структуру. Одной из характеристик таких данных, является возможность производить поиск фотографий, сделанных поблизости от определенного места путём ввода координат в поисковую систему с поддержкой геотегинга. В таких случаях поисковые системы с поддержкой геотегинга могут быть полезными для поиска привязанных к определенному месту новостей, веб-сайтов, или других ресурсов.[8]

В области информационного поиска, один из устоявшихся методов классификации геоданных, является представление каждой геометки, используя автоматический геотегинг фотографий. Для автоматической записи GPS-координат места съёмки в EXIF-данные используется фотокамера, оснащённая встроенным GPS-приёмником, записывающая координаты в файл непосредственно в момент создания снимка.

Во второй главе были проанализированы способы автоматизации профилирования пользователей, а также выбраны наиболее подходящие технологии для их реализации [9].

В социальных сетях генерируются большие потоки данных (создаются профили, связи, контент). Анализируя эти данные можно получить много полезной информации как по различным группам, сообществам и обсуждениям, так и по каждому пользователю в отдельности [1,2]. Большой интерес к социальным сетям испытывают различные коммерческие организации, использующие их как инструмент взаимодействия с аудиторией. Применяя специализированные сервисы, компании анализируют информацию о пользователях, их активностях и персонализируют предложения для отдельно взятых сегментов своей целевой аудитории, тем самым повышая конверсию и снижая затраты на рекламную кампанию. В работе рассматривается метод повышения эффективности подобного рода инструментов и сервисов, который основан на психологии и паттернах человеческого поведения. Предлагаемый метод основан на следующих фактах:

- многие пользователи Интернета имеют аккаунты сразу в нескольких популярных социальных сетях (ВКонтакте, Facebook, Instagram и Twitter);
- многие пользователи социальных сетей скрывают информацию о себе от незнакомых людей (в том числе информацию о наличии аккаунтов в других социальных сетях);
- поскольку социальные сети являются предметом социализации людей, то для каждого пользователя можно выделить хотя бы одно сообщество людей такое, что пользователи этого сообщество попарно знакомы друг с другом;
- человек имеет аккаунты в разных социальных сетях и контактирует с одними и теми же людьми [10].

В третьей главе будет разработана структура системы, а также частично реализовано приложение по автоматизации профилирования пользователей социальных сетей.

Данная часть находится в стадии разработки.

ЗАКЛЮЧЕНИЕ

В работе рассмотрены основные понятия и приемы, используемые в задачах автоматизации профилирования пользователей социальных сетей. В качестве объекта изучения была выбрана социальная сеть «Вконтакте». Был проведен анализ возможности применения методов автоматизации для упрощения основных действий, связанных с профилированием персональных страниц пользователей и возможностей, предоставляемых владельцу страницы непосредственно со стороны социальной сети. Из существующий и наиболее распространенных социальных сетей, наиболее популярной является «Вконтакте», по этой причине именно по ней был произведен анализ, целью которого являлось выявление наиболее оптимальных вариантов для автоматизации профилирования пользователей данной социальной сети.

Целью данной работы являлось исследование существующих подходов к автоматизации профилирования пользователей социальных сетей, а также проектирование и реализация приложения с использованием технологии VK API для сбора данных о пользователе социальной сети «Вконтакте».

В ходе проектирования построена структура приложения, удовлетворяющее требованиям по автоматизации возможностей владельцев страниц в социальной сети.

С помощью технологии VK API будет реализовано приложение, позволяющее собирать данные для автоматизации профилирования пользователей социальной сети «Вконтакте».

Цель работы полностью достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Главные понятия социальных сетей [Электронный ресурс]: сайт. URL: http://itua.info/internet/15459.html (дата обращения: 21.03.2019)

2. О сайте «Вконтакте» [Электронный ресурс]: сайт. URL: https://vk.com/page-47200925_44240810. (дата обращения: 22.03.2019)

3. Социальная сеть «ВКонтакте» [Электронный ресурс]: сайт. URL: http://ru.wikipedia.org/wiki/vk (дата обращения: 22.03.2019)

4. Википедия. Свободная энциклопедия [Электронный ресурс]: сайт. URL: https://ru.wikipedia.org/wiki/Метаданные (дата обращения: 20.04.2018).

5. Воройский Ф. С. Информатика. Новый систематизированный словарь -справочник (Вводный курс по информатике и вычислительной технике в терминах). — 2-е изд., перераб. и доп.. — М.: Издательство Либерия, 2006. — С. 536

6. Task Force on Metadata. Summary Report. // American Library Association. — 1999. — T. June.

7. Википедия. Свободная энциклопедия [Электронный ресурс]: сайт. URL: https://ru.wikipedia.org/wiki/Геотегинг (дата обращения: 21.04.2018).

8. Willison, Simon. (2008). "Wininear.com, OAuth and Fire Eagle" SimonWillison.net, Mar 22 2008

9. The Internet Engineering Task Force. "Geographic registration of HTML documents". Retrieved 2008-07-30.

10. Документация социальной сети «Вконтакте» [Электронный ресурс]: сайт. https://vk.com/dev/openapi (дата обращения: 25.04.2018).