УДК 004.67

ПРИМЕНЕНИЕ СРЕДСТВ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ИНФОРМАЦИИ ПРИ ВЫЯВЛЕНИИ ИНФОРМАЦИОННОГО ОБРАЗА

Тропников А.С.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.полит.н, Чугунов А.В.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В статье представлены основные результаты эксперимента по применению автоматизированных средств обработки данных для изучения поведенческих особенностей пользователей социальных сетей. Полученные данные были использованы для создания алгоритма работы приложения по определению информационного образа пользователя.

Все большее количество социальных интеракций и общественных действий фиксируется в сети. В настоящее время социальные сети медиа и социальные сети являются местом постоянного обмена большими объемами персональных данных между пользователями, не обладающими какими-либо специализированными навыками, но способными создать свой онлайн-образ и наладить взаимодействие со всем киберсообществом.

Пользователи делают «сэлфи» из отпусков, комментируют в блоге только что просмотренный фильм в кинотеатре, ставят «лайки» понравившимся записям. Совершая подобные действия, пользователи оставляют «цифровые отпечатки» – набор разнообразных данных, начиная от «лайков» и фотографий, заканчивая данными геолокации и MAC-адресов устройств, с которых осуществлялся доступ в Интернет.

Цифровые отпечатки создают имидж человека в социальных сетях и оказывают влияние уже на реальную личную жизнь человека, его профессиональную, социальную и политическую деятельность. Данные «цифровые отпечатки» используются в самых различных областях: в таргетинговой рекламе и маркетинге, рекомендательных системах, персонализированных поисковых запросах, политическом маркетинге, системах рекрутинга персонала. Однако выявление подобных взаимосвязей сопряжено с трудозатратным процессом обработки и анализа больших данных. Для упрощения данного процесса можно применять различные технологии автоматизации: по сбору данных, их обработке и непосредственному анализу.

Большая часть проектируемой системы состоит из инструментов выгрузки данных из различных источников, а также их дальнейшей обработки для приведения к необходимому формату. Однако, основная функция системы – прогнозирование «информационного образа» или отдельных личностных черт, происходит на основе заранее выработанных алгоритмов и выявленных корреляций.Для их нахождения, выявления и подтверждения могут быть использованы как существующие данные, предоставляемые третьими лицами, так и получены в ходе самостоятельных исследований.

В ходе ранее проведенных исследований, мы применяли методы кластерного анализа, статистического анализа и корреляционного анализа для поиска различных взаимосвязей. Используя базу испытуемых, вручную прошедших психологическое тестирование и предоставивших доступ к своим контактам, мы получаем доступ к клиентским социальным сетям: страницам ВКонтакте, Твиттеру, Инстаграмму – ко всему, что либо указал сам испытуемых, либо было получено через поисковые сервисы при помощи предоставленных данных (имя, фамилия, город и пр.). Из этих источников собирается вся возможная информация, которая затем приводится к формализованному виду.

Далее необходимо решить, с помощью каких именно методов и при помощи каких инструментов мы сможем получить наиболее достоверные корреляции. Применение одновременно машинного обучения, кластерного и корреляционного анализа, методов

Спирмена и т.п. позволяет получить целый набор вариативных данных, полученных при помощи различных методик, из которых мы сможем выбрать наиболее достоверные и точные с нашей точки зрения.
Используя выявленные ранее взаимосвязи между пользовательскими данными и информационным образом, мы сможем сопоставить информационный образ исследуемого человека с заранее исследованным испытуемым с ярко выраженными личностными чертами (интроверт, экстраверт, сангвиник, холерик, флегматик и т.п.) : кто из них чаще всего отмечается в социальных сетях, кто из них имеет больше друзей, кто чаще ругается матом, кто чаще путешествует, кто пишет на иностранном языке, кто общается при помощи сложных речевых оборотов, кто часто фотографируется в компании и т.д.
И чем ближе информационный образ исследуемого пользователя будет к образу того или иного «эталонного» испытуемого – тем более высокая или низкая оценка будет присвоена парметру.
Для реализации системы предложено использовать следующие модули: модуль оценки (непосредственная алгоритмизированная оценка пользователя на основе полученных данных), модуль парсинга (сбор данных из социальных сетей) и модуль API (обработка и передача необходимых запросов сторонним сервисам для анализа данных).
Под модулем оценки понимается набор скомпилированных алгоритмов, способных рассчитывать вероятностные показатели отдельных личностных черт пользователя и его информационного образа.
Под модулем парсинга понимается спрограммированный набор "краулеров" и "парсеров", способных машинным образом сканировать профили в социальных сетях и копировать оттуда всю необходимую информацию.
Предполагается что оценка будет являться бальным обозначением принадлежности пользователя к какой-либо личностной черте.
При запросе на выставление новой оценки, система просматривает, есть ли у данного клиента уже существующая оценка, в случае её отсутствия происходит инициация процесса скорринга. Процесс скорринга заключается в сравнении показателей анализируемого клиента с показателями уже изученных клиентов, попадающих в определенную категорию (Например, в категорию неплательщиков), и выдаче оценки на основании близости их атрибутов. Имеющие данные об ФИО, адресах в социальных сетях и различных интернет-сервисах, передаются в модуль парсинга, который использует их для получения как можно больше данных о клиенте.
Собранные данные сортируются на два вида: готовые статистические данные, которые уже можно применять для оценки клиента (количество лайков, постов, друзей) и «сырые» данные (фотографии, текстовые сообщения, посты) которые необходимо предварительно обработать. Модуль API формирует запрос согласно полученным данным и направляет его в необходимые облачные сервисы машинной обработки данных. После того как «сырые» данные будут обработаны и формализированы в сторонних сервисах, они загружаются в модуль оценки, где уже на основании всей полученной информации происходит скорринг.
В результате проделанных работ были разработаны модули по выгрузке и обработке данных через API в сторонние облачные сервисы. Данные модули в автоматизированном порядке способны отправлять запросы на обработку фотографий, выгружая сразу выходные данные из сторонних сервисов. Благодаря внедрению подобных модулей, мы смогли значительно сократить время, необходимо на анализ отдельного профиля пользователя.
В качестве перспективы работы планируется дальнейшая автоматизация элементов проектируемой системы, чтобы в конечном итоге полностью автоматизировать процесс прогнозирования человеческих психологических особенностей на основе пользовательских профилей.