Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

кафедра социальной информатики

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ ПРОГРАММНЫХ СРЕДСТВ В ИССЛЕДОВАНИИ КОНТЕНТА СОЦИАЛЬНЫХ СЕТЕЙ

(автореферат бакалаврской работы)

студентки 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Воронковой Ольги Александровны

Научный руководитель
кандидат социологических наук, доцент _____ С.В.Ситникова
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г.Малинский
подпись, дата

Саратов 2017

Введение. Актуальность проблемы. На сегодняшний день внимание многих ученых, работающих в различных научных сферах, сосредоточено вокруг Интернет исследований. Высокие возможности современных компьютеров, широкое распространение Всемирной паутины, возрастающая доступность средств связи для различных групп населения сопровождается ежедневным приростом информации.

Этот процесс информатизации привел к созданию особых форм социальных отношений, к образованию социальных групп, основанных на связях через Интернет, к применению новых средств формирования общественного мнения посредством социальных медиа. Интернет стал не только индикатором общественных процессов, но и электронным банком данных о социальной жизни в ее разнообразных проявлениях. «Цифровой след», который оставляют люди, пользуясь электронными средствами связи, открывает новые возможности для наук о человеке: социологии, социальной философии, демографии, этнографии и др.

Степень научной разработанности проблемы. Несомненно, большим потенциалом обладают данные, которые размещены в открытом доступе с согласия индивидуума. Речь идет об анкетах пользователей в электронных социальных сетях. Термин «социальная сеть» возник еще до появления современной вычислительной техники. Считается, что в научный оборот его ввел социолог Д. Барнс в 1954 году: «Социальная сеть – это социальная структура, состоящая из группы узлов, которыми являются социальные объекты (люди или организации), и связей между ними (социальных взаимоотношений)». Другими словами – это некая группа знакомых людей, где сам человек является центром, а его знакомые ветками. Между всеми членами сети есть двусторонние или односторонние связи. Д. Барнс развил подход Дж. Морено, который заключался в исследовании взаимосвязей между людьми с помощью социограмм, то есть визуальных диаграмм, в которых отдельные лица представлены в виде точек, а связи между ними - в виде линий. Другой известный ученый – Радклифф-Браун первым использовал терминологию

сетевых исследований и призвал рассматривать общественную структуру как сеть социальных отношений. Впоследствии основные направления исследования сетевых сообществ были заложены в методологическом аппарате П. Лазарсфельда, Дж. Хоманса, Л.Фримена, С. Милграма, М. Грановеттера.

Электронные социальные сети, появившиеся в начале 2000-х годов, можно считать своеобразным «отражением» или «проекцией» социальных сетей, как их понимают в социологии. В дальнейшем в своей работе я буду использовать термин *социальная сеть* для обозначения именно электронной социальной сети — фактически, базы данных о пользователях и связях между ними.

В зависимости от социальной сети, доступные о пользователях данные, конечно, варьируются. Если причислять к социальным сетям популярные сервисы блогов, микроблогов или фото-блогов (например, livejournal.com, twitter.com и instagram.com), то, из «анкетной» информации о пользователе гарантированно можно рассчитывать только на какое-то имя, возможно вымышленное.

Если же ограничиться «классическими» социальными сетями (vk.com, facebook.com, ok.ru), то анкеты пользователей становятся более развернутыми — пользователям предлагается указать свое имя, дату рождения и пол, а также сведения о местах проживания, учебы и работы. В отличие от опросного листа, анкета пользователя социальной сети имеет заранее определенный список полей (он задан создателями конкретной социальной сети) и исследователь не может на него влиять. Конечно, наличие полей в анкете не гарантирует, что они будут заполнены, и, тем более, что они будут содержать достоверные сведения. Однако, учитывая, что основная цель социальной сети как сервиса — способствовать общению и установлению контактов, указание достаточно подробных сведений о себе представляется разумным с точки зрения пользователей, которые хотят, чтобы их было легче найти их одноклассникам, сокурсникам, коллегам по работе.

Еще одна проблема для социологов - обеспечение и доказательство репрезентативности выборки. При таком «электронном» опросе сделать это гораздо сложнее, чем при проведении классических «поквартирного» или «телефонного» опросов. Но проблема составления выборки сходит на нет, если есть возможность извлечь и обработать всю генеральную совокупность. Подобные массивы информации — яркий пример «больших данных» (англ. «big data»). Разумеется, проведение таких исследований крайне трудоемко без использования специальных программ, позволяющих анализировать как можно большее количество контента за меньшее количество времени.

Таким образом, актуальность данной бакалаврской работы обусловлена необходимостью развития методологического аппарата, который позволил бы использовать большие объемы анкет пользователей социальных сетей для решения комплекса задач по автоматизации мониторинга общественного мнения.

В связи с этим **целью** работы является разработка и апробация приложения для автоматического сбора данных о пользователях электронной социальной сети для последующего анализа.

Для достижения заданной цели необходимо решить следующие **задачи**:

1. Представить и описать возможности, преимущества и недостатки использования Online Big Data в социологических исследованиях;
2. Представить методологические, методические и технические особенности создания авторского программного продукта по сбору и анализу данных из «классических» социальных сетей, а именно «ВКонтакте»;
3. Описать проблемы, возникающие при создании и использовании программного средства, а также последующей формализации и интерпретации данных;

4. Представить итоговые результаты экспериментальной апробации программного продукта по сбору и анализу контента социальных сетей (на примере «ВКонтакте»).

Объектом исследования выступают технологии извлечения больших массивов социологических данных на примере авторского программного средства (VKtoExel).

Предметом исследования являются особенности применения программного продукта VKtoExel в изучении контента социальных сетей.

Практическая значимость исследования заключается в возможности использования автоматизированного средства, которое позволяет извлекать большие массивы информации из социальных сетей. Это способствует уменьшению трудоемкости работы социологов, HR – менеджеров и многих других специалистов.

Структура бакалаврской работы. Квалификационная работа состоит из введения, трех разделов, заключения, списка литературы и приложений.

Основное содержание работы. В разделе 1 «Теоретико-методологические основания исследования контента социальных сетей» описываются возможности, преимущества и недостатки использования Online Big Data в социологических исследованиях при изучении содержания социальных сетей. Получение ценных знаний через автоматизированный анализ больших объемов данных (которые в частности предоставляют социальные сети) сегодня является крайне популярным направлением и тесно сопряжено с такими терминами, как Big Data («большие данные») и Data Mining (в русском языке нет устоявшегося перевода для данного термина, но принято понимать его как «добыча знаний из данных» или «интеллектуальный анализ данных»).

Преимущество автоматизированных методов онлайн-исследований в том, что хранение больших данных предусматривает доступ ко всему разнообразию Big Data, независимо от того, какие из них актуальны на данный момент, а к

каким будут обращаться в дальнейшем. Опираясь на эти данные, социологи производят сегментацию и кластеризацию наблюдаемых явлений, строят прогностические модели социальных процессов. Недостатком является то, что ещё не разработана концептуально-методологическая схема анализа Big Data, в которой все результаты сведены к общему знаменателю, а по каждому показателю известны алгоритмы получения числовых значений, с помощью которых вся эта информация должна быть осознана и использована для проверки гипотез и формирования обоснованных выводов.

Применение автоматизированных методов онлайн-исследований поднимает вопросы стандартизации веб-измерений. Несмотря на наличие многочисленных и дополняющих друг друга программных средств, собрать надёжные данные о структуре аудитории Интернета довольно сложно. В существующих сервисах исследователю чаще всего не доступна модификация заложенных показателей, что не позволяет ему осуществлять независимую статистическую интерпретацию, необходимую для проверки собственных гипотез и объяснения интересующих фактов. Еще одной преградой для использования уже разработанного программного обеспечения также является отсутствие универсального программного продукта с типичными функциями. Регулярность усовершенствования сервисов, пополнение доступных операций приводит к дезориентации аналитиков, которые не успевают разобраться с одним сервисом, как выходит другой с более сложным и интересным пакетом функций, в результате чего много времени уходит на их установку, настройку и освоение. В связи с этим необходима разработка и апробация концептуально-методологической схемы анализа данных, собранных с помощью автоматизированных веб-измерений.

В разделе 2 «Алгоритм работы программного приложения. Основные достоинства и сложности реализации» представлены методологические, методические и технические особенности создания авторского программного продукта по сбору и анализу данных из «классических» социальных сетей, а именно «ВКонтакте» Приложение написано на языке программирования C#.

Оно состоит из двух компонентов: окно, где задаются фильтры и базы данных. При запуске программы появляется форма, где нужно установить фильтры (например, страна, город, возраст), необходимые для выполнения исследовательских задач. Данные поля можно оставлять пустыми или же заполнять частично, исходя из задач предстоящего анализа.

При нажатии кнопки «Поиск» программой производится случайный отбор анкет, которые можно увидеть в окне слева. После окончания работы, приложением формируется список пользователей, который при нажатии кнопки «Exel» переносится в соответствующий документ. Выбор такого редактора таблиц обусловлен несколькими причинами. Во-первых, данный продукт есть почти на всех компьютерах владельцев Windows, так как это стандартная программа пакета Microsoft Office. Во-вторых, это удобство и быстрота осуществления первичного анализа (вывод диаграмм, графиков). И в-третьих, для дальнейшего анализа каждый исследователь сможет выбрать такой статистический пакет, с которым ему комфортно работать. Для этого нужно просто скопировать и перенести всю информацию в выбранную программу. Результаты извлечения данных собираются в таблицу Exel, в которой каждая анкета представлена строчкой, а поля анкеты — столбцами. Поля таблицы представлены следующей информацией: ID, имя, фамилия, пол, дата рождения, страна, город, статус, семейное положение, школа, вуз, карьера, деятельность (чем увлекается/занимается пользователь), интересы, любимая музыка, любимые фильмы, любимые телешоу, любимые книги, любимые игры, сообщества, в которых состоит пользователь (группы по интересам). При более глубинном анализе этих данных, функционала Exel уже недостаточно, поэтому применяются специализированные статистические пакеты. Выбор таких программ достаточно велик, однако большое значение имеет удобство работы с пакетом, легкость его освоения, а также скорость произведения вычислений. Программа SPSS как раз подходит под все эти критерии.

В простейшем варианте, когда рассматриваются только анкетные данные пользователей, но не рассматриваются связи между ними, обработка

заключается, фактически, в агрегировании данных по каким-либо показателям. В проведенных нами исследованиях данные социальных сетей использовались для изучения интересов пользователей различных возрастных категорий из нескольких городов Саратовской области. Случайным образом были выбраны 3 наиболее крупных по численности города - Саратов, Энгельс и Балаково. В задаче исследования не стояло цели изучить только тех пользователей, у которых есть фотографии или тех, кто находится онлайн на момент сбора информации, поэтому были установлены следующие фильтры: Страна - Россия, Город - Балаково, Саратов, Энгельс, Возраст от - 14, Возраст до – 99.

При разработке программы я столкнулась с проблемой извлечения максимального числа анкет. Политика социальной сети Вконтакте не позволяла собрать информацию больше, чем с 1000 страничек пользователей за один раз. Поэтому для демонстрации работы приложения я ограничились таким числом. Если же необходимо большее число, то процедуру сбора можно провести несколько раз, а затем с помощью фильтрации в Exel удалить повторяющиеся анкеты. Кроме этого технические возможности программы не дают выбрать сразу несколько городов, поэтому я последовательно собрала информацию о 1000 жителях Балаково, затем 1000 Саратова и 1000 Энгельса. В общей сложности получилась база данных, состоящая из 3000 анкет.

Но несмотря на некоторые сложности работы с программой, все же главным ее достоинством является значительное сокращают времени проведения исследования. Так, например, чтобы собрать данные о 3000 пользователей сети «ВКонтакте» мне потребовалось не более 30 минут. Конечно, в приложение стоит внести некоторые корректировки: возможность фильтрации по нескольким городам или странам одновременно, также стоит продумать алгоритм самостоятельного внесения фильтров в зависимости от целей и задач исследователя. Но в целом – это готовая к работе программа, значительно упрощающая сбор информации из социальных сетей.

В разделе 3 «Результаты экспериментальной апробации авторского программного продукта VKtoExel. Социологическая интерпретация

полученных статистических данных» представлены итоговые результаты экспериментальной апробации программного продукта по сбору и анализу контента социальных сетей, а также описаны проблемы, возникающие при создании и использовании программного средства, а также последующей формализации и интерпретации данных. Извлеченная приложением информация была проанализирована по нескольким параметрам: 1) количественное соотношение мужчин и женщин в каждом городе; 2) количество пользователей различных возрастов в каждом городе; 3) количество участников популярных сообществ.

По итогу было получено следующее:

Доля мужчины и женщины в трех выбранных нами городах практически равна. Незначительное преобладание мужчин (51%) в г. Балаково, а в Саратове и Энгельсе - женщин (56% и 52% соответственно).

При анализе возрастного распределения пользователей возникли некоторые сложности: не все пользователи заполняют информацию о себе. Так, например, почти половина (46,5%) не указали свой год рождения, ограничившись только днем и месяцем. Следовательно, вычислить возраст было нельзя. В настоящей работе эти погрешности не устранялись, поскольку данные были собраны в тестовом режиме.

Поскольку мы не располагаем эмпирическими данными о том, насколько информация, публикуемая пользователями в профиле, соответствует их реальному представлению о тех или иных вопросах, то целесообразно исследовать их ценностные ориентации методом анализа наиболее популярных в сети пабликов (сообществ). Было получено, что популярные сообщества для всех трех городов практически идентичны. Пользователям интересны новости и мероприятия своего города. Кроме того респондентов прельщают группы, где с помощью несложных действий можно получить приз или какую-то вещь бесплатно. Также востребованными являются такие сообщества, где пользователи сами составляют новостной контент, например, пишут о том, где находятся посты ДПС - в каких районах, на каких улицах. Изучая различия

между сообществами мужчин и женщин, было получено, что особой популярностью у девушек пользуются паблики с красивыми картинками и стихами о любви и отношениях. Мужчин же привлекают группы со смешными картинками и интернет мемами.

Заключение. Развитие компьютерных технологий и внедрение их в повседневную жизнь открыли не только множество возможностей перед человечеством, но и поставили ряд проблем, для которых требуется найти решение. Большие объемы информации поступающей из Интернета и в частности из социальных сетей требуют специальной обработки. При этом она должна быть не только точной и качественной, учитывающей модель данных и их структуру, но и быстрой.

В мире больших данных мы можем проанализировать огромное количество информации, а в некоторых случаях – обработать все данные, касающиеся того или иного явления, а не полагаться на случайные выборки. Используя все данные, мы получаем более точный результат и можем увидеть нюансы, недоступные при ограничении небольшим объёмом данных. Также благодаря Big Data социологи больше не обязаны цепляться за причинность. Вместо этого можно находить корреляции между данными, которые открывают перед нами новые неоценимые знания. Зависимости не могут сказать нам точно, почему происходит то или иное событие, зато предупреждают о том, какого оно рода. И в большинстве случаев этого вполне достаточно. Однако основная трудность при работе с такими данными заключается в том, что не все социологи обладают соответствующими компетенциями написания программного кода или построения математических моделей для автоматизированного сбора и анализа больших массивов информации. В существующих же программах исследователю чаще всего не доступна модификация заложенных критериев, что не дает ему осуществить соответствующую статистическую интерпретацию, необходимую для проверки собственных гипотез или же объяснения имеющихся фактов.

В данной работе на основе проведенных исследований и выполненных разработок, решена актуальная научно-практическая задача, связанная с созданием универсального программного продукта, позволяющего извлекать данные из классических анкетных социальных сетей.

Проведенное исследование позволило составить представление об особенностях извлечения информации из социальной сети Вконтакте. Опыт работы с данным приложением позволил в короткие сроки собрать большой массив данных (суммарно было извлечено 3000 анкет пользователей), а также провести анализ собранной информации по нескольким параметрам: 1) количественное соотношение мужчин и женщин в трех выбранных нами городах Саратовской области; 2) количество пользователей различных возрастов в каждом городе; 3) количество участников популярных сообществ.

При интерпретации таких данных, стоит учитывать, что в Интернете, как и в реальном мире, люди объединяются в определенные социальные группы (социальные маски), которые между собой не сильно пересекаются. Основное глобальное деление происходит по проектам, внутри которых целевая аудитория разбивается на неформальные группы по интересам, возрасту и другим признакам. Причем таких социальных масок может быть несколько: днем человеку нужно деловое общение, вечером общение с друзьями и семьей, по выходным общение, например, связанное с хобби и т.д. У каждого набор масок будет свой, однако у каждой из них будут свои особенности, которые будут влиять на все поведение. Именно поэтому современный человек часто зарегистрирован в нескольких социальных сетях, в которых он удовлетворяет разные потребности и дает о себе разную информацию. Следовательно, для полноты проводимого анализа необходимо использовать информацию сразу из нескольких социальных сетей. В связи с этим в дальнейшем планируется разработка таких алгоритмов, которые позволяли ли бы динамично изменять параметры разработанного приложения, что сделало бы возможным собирать информацию по совершенно иным характеристикам, например, использование других социальных сетей.