ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

Б.А. Низомутдинов, А.С. Тропников,
А.Б. Углова

Разработка прогностической модели информационного образа пользователя с применением автоматизированных средств обработки данных из социальных сетей

Рекомендуемая форма библиографической ссылки

Низомутдинов Б.А., Тропников А.С., Углова А.Б. Разработка прогностической модели информационного образа пользователя с применением автоматизированных средств обработки данных из социальных сетей // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 532-540. — URL: http://keldysh.ru/abrau/2019/theses/82.pdf doi:10.20948/abrau-2019-82

Размещена также презентация к докладу

Разработка прогностической модели информационного образа пользователя с применением автоматизированных средств обработки данных из социальных сетей

Б.А. Низомутдинов[1], А.С. Тропников[1], А.Б. Углова[2]

[1] Университет ИТМО,
[2] РГПУ им. А.И. Герцена

Аннотация. В статье представлены результаты первого этапа исследовательского проекта, предполагающего разработку методики автоматизированного анализа характеристик информационного образа пользователей социальных сетей на базе социальной сети «ВКонтакте», а также взаимосвязи определенных характеристик с социально-психологическими атрбитами респондентов. В ходе пилотного исследования осуществлен сравнительный анализ автоматических методов анализа текстовых и визуальных данных, составляющих информационный образ пользователя, выделены основания и описана классификация компонентов информационного образа пользователя социальных сетей. Для выявления личностной опосредованности информационного образа был использован корреляционный анализ данных полученных с использованием методов парсинга и результатов психодиагностического исследования. Компоненты информационного образа образовали множественные взаимосвязи с социально-психологическими свойствами респондентов. На основе полученных взаимосвязей был сделан вывод о приемлемых прогностических возможностей автоматизированного анализа данных социального профиля.

Ключевые слова: ВКонтакте, социальные сети, автоматизированный анализ, профили

Developing Prognostic Models of User's Image with the Automated Social Networks Data Processing Methods

B.A. Nizomutdinov[1], A.S. Tropnikov[1], A.B. Uglova[2]

[1] ITMO University
[2] The Herzen State Pedagogical University of Russia

Abstract. This article presents a first phase's results of research project, dedicated to developing prognostic models of User's Image with the automated Social Networks data processing methods. A pilot study carried out a comparative analysis of automatic methods for analyzing textual and visual data, which constituting a User's informational image. We used a correlation analysis of data, which was obtained by using parsing methods and the results of psycho diagnostic research. As a result, multiple relationships with the socio-psychological characteristics of the respondents were identified. Based on obtained data, we have concluded about acceptable predictive capabilities of automated analysis.

Keywords: VK, social network site, personal traits, automated analysis

1. Введение

Аудитория социальных сетей расширяется с каждым годом. За период с 2008 по 2018 год, число пользователей одной только социальной сети Facebook увеличилось со 100 млн. до 2,3 млрд. [1]. Более 42% жителей Земли активно использует одну или несколько социальных сетей [2]. Пользователи оставляют в сети огромное количество общедоступных данных, которые активно используются в задачах политического и коммерческого маркетинга: для выявления общественного мнения, определения потребительских предпочтений, создания таргетированной рекламы и др.

Подобная информация может быть использована для выявления характеристик современной виртуальной коммуникации, определения факторов, опосредующих межкультурные и общественно-политические противоречия, выявления уязвимых социальных группы и т.д. [3] Использование автоматизированных систем поможет облегчить процесс получения подобной информации и упростить описание информационного образа пользователя, который может быть в ходе дальнейшего исследования соотнесен с реальным социально-психологическим портретом человека. Информационный образ понимается нами как целостная структура текстовых и визуальных компонентов сетевого облика человека, которая лежит в основе создания сетевой личности с набором определенных личностных черт. Можно говорить о том, что изучение сетевых данные пользователей позволит нам создать прогностическую модель для автоматизированного анализа данных личностных черт, реализуемых в реальной коммуникативной практике.

2. Методика и процедуры исследования

В ряде современных научных исследованиях рассматривается возможность разработки и использования новых методов для выявления социально-психологических характеристик пользователей на основе анализа информационного контента, которые могли быть также эффективны как и традиционные психодиагностические исследования [4,5]. Многие зарубежные и отечественные работы посвящены разработке различных методик, связанных с

выявлением взаимосвязей между данными из социальных сетей и информацией, полученной в ходе психологического тестирования [6,7].

Данное исследование включает в себя 2 подхода: классическое психодиагностическое исследование и автоматизированный сбор открытых данных в социальных сетях методом парсинга.

Целью данного исследования является теоретическое обоснование и экспериментальное изучение прогностической модели информационного образа на основе автоматизированной выгрузки массива различных данных из профилей социальных сетей.

На первом этапе исследования были решены следующие задачи:

1. Проведен анализ исследований, посвященных изучению взаимосвязи различных группах данных, получаемых из социальных сетей и личностных характеристик пользователей.
2. Осуществлен сравнительный анализ автоматических методов обработки текстовых и визуальных данных, составляющих информационный образ пользователя.
3. Выделены основания и описана классификация компонентов информационного образа пользователя социальных сетей.
4. Разработан комплекс автоматизированных методов выгрузки массива данных из профиля социальных сетей, составляющих информационный образ пользователя.
5. Выявлены взаимосвязи полученного массива данных с личностными свойствами испытуемых.

Пилотное исследование проходило в несколько этапов:

На первом этапе был осуществлен поиск и анализ иностранной и российской научной литературы, выявление используемых для решения поставленных задач информационных систем, произведен сравнительный анализ этих систем и тестирование специализированного программного обеспечения.

На втором этапе были выбраны профили испытуемых и произведена выгрузка всей необходимой информации, а также произведена обработка отсортированных данных с помощью различных сервисов. В ходе обработки данных и анализа выходной информации проведено отсеивание неподходящих или неточных сервисов.

На третьем этапе было проведено пилотное психодиагностическое исследование респондентов с использованием отобранных сервисов.

На завершающем этапе, используя обработанную информацию, соотнесение закономерностей и повторяющихся взаимосвязей данных с ранее проведенным психодиагностическим исследованием и создана таблица корреляции.

В дальнейшем исследовании запланировано, на основе данной таблицы разработать алгоритм, позволяющий автоматически обрабатывать массивы

выгружаемых данных и определять предполагаемые психоэмоциональные особенности пользователя.

В пилотном исследовании приняли участие 22 человека, относящихся к первому периоду зрелого возраста (19-32 лет). Средний возраст испытуемых 24,6 года. В исследовании участвовали 55 % (12) женщин и 45 % (10) мужчин. Участники исследования имели разный уровень образования: 6 человек (28%) с неоконченным высшим образованием и 16 (72%) с высшим образованием.

На основании разработанной модели исследования был выбран комплекс методов и методик с целью определения информационного образа пользователя социальных сетей. Были использованы следующие методы: психодиагностические методы, методы автоматизированного сбора информации и математической обработки. Психодиагностический комплекс методик содержит: авторскую анкету исследования поведения в Интернете; ценностный опросник С. Шварца; методику исследования самоотношения С.Р. Пантелеева; тест диагностики степени удовлетворенности потребностей (пирамида Маслоу). Отобранный психодиагностический инструментарий позволяет выявить систему ценностно-смысловых ориентаций испытуемых, а также особенности их мотивационно-потребностной сферы и самоотношения, которые лежат в основе формирования самопрезентации, формирования впечатления о себе для решения различных адаптационных задач.

3. Инструментарий

Основываясь на ряде проводимых исследованиях в области выявления взаимосвязей социальных сетей и личностных характеристик пользователей, можно выделить несколько направлений по определению взаимосвязей больших данных, полученных из социальных сетей и персональными качествами пользователей: обработка фотоизображений, семантический анализ текстовых пользовательских «постов» и анализ статистических данных. В данном исследовании использовались фотоизображения и статистические данные.

Для анализа выгруженных фотоизображений можно использовать один из нескольких инструментов машинной обработки изображений, способных выявлять отображаемые эмоции, подсчитывать количество людей на фотографиях, определять объекты на заднем фоне и рассчитывать положение предметов.

В качестве инструментария для анализа фотоизображения была выбрана платформа Azure от компании Microsoft. Данная вычислительная платформа может использоваться для выявления демонстрируемых человеком эмоций на его социальном «аватаре», вычислять количество лиц на снимках, а также их повторяемость. Данные о каких-либо перекрытых частях лица или необычных позах также могут дать информацию об особенностях характера человека.

В качестве выгрузки самих данных пользователей из социальных сетей может быть использован API. С его помощью разработчики приложений

напрямую могут обращаться к социальным сетям через выделенный интерфейс и выгружать необходимые данные, такие как имена, адреса, статусы, группы и т.д. Выгрузка такого массива информации в автоматическом режиме поможет ускорить процесс подбора материала для выявления психоэмоциональных особенностей пользователя

4. Извлечение информации автоматизированными средствами

По итогам анкетирования группы испытуемых была создана база, содержащая ссылки на страницы пользователей в социальной сети «ВКонтакте». На следующем шаге работы был применен метод автоматизированного сбора общедоступной информации со страниц профилей в социальной сети ВКонтакте.

В результате парсинга персональных страниц был получен массив данных, который представлял собой статистическую информацию о профиле пользователя социальных сетей, который включает в себя:

- ID аккаунта;
- ссылка на профиль;
- ФИО;
- кол-во интересных стр;
- кол-во видео;
- кол-во друзей;
- место работы;
- год рождения;
- дата рождения;
- ссылка на Инстаграм;
- статус;
- образование;
- город проживания;
- кол-во фотографий;
- ссылка на аватар.

На данной стадии исследования для автоматизированного сбора данных об участниках был использован метод парсинга контента страниц. Парсинг сайтов является эффективным решением автоматизации сбора и обработки информации. Сбор осуществляется специальным скриптом, настроенным на работу с vk.com. Процесс парсинга страницы профиля VK можно разделить на три шага.

1. Получение исходного кода html – страницы и копирование исходного кода страницы.
2. Извлечение из полученного кода нужной информации. Получив исходный код html - страницы, необходимо выполнить над ним обработку, т.е. отделить искомый текст от гипертекстовой разметки, выстроить иерархическое дерево элементов документа (DOM) и

извлечь из страницы искомую информацию. По заданным критериям выделить только основную информацию, которая представляет интерес.

3. Сохранение результата. После успешного извлечения данных страницы их необходимо сохранить в требуемом виде для дальнейшей обработки.

Данный способ имеет ряд ограничений, главное из них - скорость, и то, что для выгрузки отдельного рода информации необходима имитация действия пользователя, например, прокрутка. Но данный метод был использован для пилотного исследования, так как более прост, чем выгрузка по API. API в свою очередь предоставляет больше возможности для выгрузки, но более трудоемкий в настройках. На следующих этапах исследования планируется использовать выгрузку по API.

5. Автоматизированный анализ изображений

Для проведения экспериментального тестирования использовалась облачная платформа Azure от компании Microsoft, предоставляющая сервисы обработки изображений и анализа лиц. Используя предварительно проанализированные данные с более 350 тыс. изображений, данные сервисы способны эффективно и точно анализировать поступающие изображения на момент наличия определенных объектов, лиц, композиции и т.п. Всего же, облачный сервис обработки изображений способен выдавать информацию по более чем 30 критериям.

Используя данные, собранные по аватарам, была произведена выгрузка изображений в сервис обработки изображений и анализа лиц на платформе Azure с помощью модуля выгрузки, реализованного с использованием API.

Анализ изображений позволил получить информацию об их формате, размерах, наличии лиц, преобладающих цветах, имеющихся объектах и композиции. Каждому аватару был присвоен набор тэгов, классифицирующих изображение.

С помощью сервиса анализа лиц также были получены данные о возрасте, поле, лицевых атрибутах, аксессуарах и эмоциях людей, запечатленных на изображении, использующемся в качестве аватара.

6. Результаты пилотного исследования

На финальном этапе пилотного исследования, данные, полученные в ходе автоматической выгрузки, были объединены в более крупные категории, описывающие базовые компоненты информационного образа пользователя: информация об имени пользователя (самоименование), информация о возрасте, пространственная локализация, информация о близких отношениях, о профессиональном статусе, ассоциированность в сети, информация об образовании, конфиденциальности, информация о количестве друзей,

подписчиков, фотографий, видеозаписей, аудиозаписей, групп, подписок, а также эмоции, выявленные на фотографиях пользователей (удивление, печаль, нейтральное выражение, счастье, гнев, отвращение, презрение).

Для выявления личностной опосредованности информационного образа был использован корреляционный анализ. Все компоненты информационного образа образовали множественные взаимосвязи с личностными свойствами респондентов. Обратимся к интерпретации наиболее информативных фрагментов корреляционных плеяд. Для пользователей, состоящих в большом количестве групп, значимой ценностью является поддержка традиций (tradition maintenance) и социальная культура (culture specific values), у них напряжена потребность в безопасности, любви и социальной принадлежности (belongingness and Love needs). Для пользователей, выставляющих аватары, на которых распознаются такие эмоции как гнев и удивление, характерно испытывать потребность в безопасности и социальной принадлежности (belongingness and Love needs), а также у них в большей степени выражена внутренняя конфликтность и самообвинение.

Нейтральное выражение лица на аватарах характерно для тех пользователей, которым важна социальная культура, при этом счастливое выражение лица у них встречается на аватарах редко. Чем больше представлено информации о профессиональном статусе, тем больше для пользователя значимы такие ценности как зрелость и социальность. Респонденты, для которых значима ценность власти, чаще используют для аватаров фотографии, на которых присутствует эмоция презрения. При этом они часто указывают информацию о своих близких отношениях и партнерах. У респондентов, проводящих много времени в сети отмечается большее количество друзей и подписчиков, что служит подтверждению собственной самоценности и удовлетворяет потребность в самореализации.

7. Выводы и направления дальнейших исследований

В заключении можно сказать, что изучение взаимосвязей компонентов информационного образа и личностных особенностей показало довольно высокие прогностические возможности автоматизированного анализа данных социального профиля. Дальнейшее направление исследований будет связано с расширением выборки и анализом данных с учетом социально-демографических характеристик пользователей, а также особенностями их самопрезентации в сети.

Инструментарий, использованный в данной работе, был отобран на основании сравнительного анализа, основанного на ряде критериев, включая: максимальный объем обрабатываемых данных, скорость обработки данных и формат выгрузки данных.

Наш дальнейший интерес на последующих этапах будет сконцентрирован на проведении качественных исследований с целью выявления паттернов использования групп, а также публикуемого контента.

Литература

1. Number of monthly active Facebook users worldwide as of 3rd quarter 2018. - URL: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide
2. Digital in 2018: World’s Internet Users Pass the 4 Billion Mark. - URL: https://wearesocial.com/blog/2018/ 01/global-digital-report-2018
3. Kosinski M., Stilwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // Proc. the National Academy of Science of the United State of America. 2013. Vol. 110. P. 5802-5805. DOI: 10.1073/pnas.1218772110
4. Ross C., Orr E.S., Sisic M., Arseneault J.M., Simmering M.G., Orr R.R.: Personality and motivations associated with facebook use // Computers in Human Behavior. 2009. Vol. 25. P. 578-586. DOI: 10.1016/j.chb.2008.12.024
5. Крылова О.С., Власов Д.А., Шишков В.В., Алымов А.С., Ишин И.А., Колесников И.Е. Петров А.И. Описание информационного образа пользователя социальной сети с учетом его психологической характеристики // International Journal of Open Information Technologies. 2018. №4. С. 24-37.
6. Mairesse F., Walker M., Mehl M., Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text // Journal of Artificial Intelligence Research. 2007. Vol. 30. P. 457-500. DOI: 10.1613/jair.2349
7. Michal K. Mining Big Data to Extract Patterns and Predict Real-Life Outcomes // Psychological Methods. 2016. Vol. 21 (4). P. 493-506. DOI: 10.1037/met0000105
8. Introduction to VK API, https://vk.com/dev/first_guide

References

1. Number of monthly active Facebook users worldwide as of 3rd quarter 2018. - URL: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide
2. Digital in 2018: World’s Internet Users Pass the 4 Billion Mark. - URL: https://wearesocial.com/blog/2018/ 01/global-digital-report-2018
3. Kosinski M., Stilwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // Proc. the National Academy of Science of the United State of America. 2013. Vol. 110. P. 5802-5805. DOI: 10.1073/pnas.1218772110
4. Ross C., Orr E.S., Sisic M., Arseneault J.M., Simmering M.G., Orr R.R.: Personality and motivations associated with facebook use // Computers in Human Behavior. 2009. Vol. 25. P. 578-586. DOI: 10.1016/j.chb.2008.12.024
5. Krylova O.S., Vlasov D.A., Shishkov V.V., Alymov A.S., Ishin I.A., Kolesnikov I.E. Petrov A.I. Opisanie informacionnogo obraza pol'zovatelya social'noj seti s uchetom ego psihologicheskoj harakteristiki // International Journal of Open Information Technologies. 2018. Vol. 4. C. 24-37.

6. Mairesse F., Walker M., Mehl M., Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text // Journal of Artificial Intelligence Research. 2007. Vol. 30. P. 457-500. DOI: 10.1613/jair.2349
7. Michal K. Mining Big Data to Extract Patterns and Predict Real-Life Outcomes // Psychological Methods. 2016. Vol. 21 (4). P. 493-506. DOI: 10.1037/met0000105
8. Introduction to VK API, https://vk.com/dev/first_guide