Комплексный подход к маппингу профилей пользователей в социальных сетях

В.А. Белов[1], Д.С. Дроздов[1], Р.А. Шакуров[1], В.С. Мошкин[1], И.А.Андреев[1]

[1]Ульяновский государственный технический университет, ул. Северный венец 32, Ульяновск, Россия, 432027

Аннотация. В данной работе рассматривается комплексный подход по единоличной идентификации человека в нескольких разных социальных сетях посредством анализа не только данных анкет, но и слабоструктурированной информации со страниц соответствующих аккаунтов, а также графической информации. Также приводится описание программного сервиса, реализующего предложенный подход.

1. Введение

Активный рост аудитории социальных сетей привел к становлению этих ресурсов в качестве нового источника данных и знаний. На сегодняшний день существует множество социальных сетей и некоторые из них пользуются огромным успехом. Так существует 10 социальных сетей, количество пользователей которых достигает от 200 миллионов до 1.3 миллиарда пользователей. В России на данный момент пользуются наибольшей популярностью несколько социальных сетей, каждая из которых имеет свою направленность и специфику размещаемого контента. К таким ресурсам можно отнести «ВКонтакте», «Одноклассники», «Instagram», «Facebook», «Youtube», Twitter [1]. Многие интернет-пользователи имеют несколько аккаунтов в разных социальных сетях и публикуют в них различный или схожий контент. И найти человека в какой-либо из сетей становится проблематично.

Работа с социальными сетями может принести пользу при реализации функции системы управления персоналом компании, так как зачастую из социальных сетей о профессиональных и личностных качествах соискателя на конкретную должность можно узнать больше, чем из его резюме. В настоящее время сбор и/или содержательный анализ собранной в социальных сетях информации проводится вручную специалистами кадровых служб, что требует больших затрат времени и ограничивает объем обрабатываемой информации.

Таким образом, возникает потребность в разработке программной системы, позволяющей идентифицировать профиль человека в нескольких социальных сетях.

Такие разработки позволили бы агрегировать больше данных о пользователях для оценки степени выраженности их личностных особенностей. Данная работа направлена на решение задачи поиска комплексного подхода маппинга (сопоставления) профилей пользователей в различных социальных сетях на основе анализа структурированных данных, текстовой информации, а также графических материалов с целью дальнейшего анализа социального портрета пользователя.

2. Основные подходы к решению задачи идентификации пользователей

2.1. Методы и алгоритмы маппинга аккаунтов пользователей в социальных сетях

В настоящее время задача идентификации пользователей с использованием данных профилей социальных сетей решается различными способами [2-4].

В работах [5-8] описываются методы анализа данных профилей пользователей социальных сетей «MySpace», «StudiVZ», которые не столь популярны в России. Предлагаемые подходы предполагают построение векторов признаков, составляющих профили пользователя. К полученным векторам применяются методы точного, частичного и нечёткого сравнения.

В данных работах авторами были предложены признаки, являющиеся наиболее существенными при сравнении аккаунтов. Разработанные алгоритмы имеют эффективность порядка 80% на тестовой выборке аккаунтов пользователей.

В [7][8] приводятся методы маппинга профилей пользователей социальных сетей посредством анализа публикуемой неструктурированной (текстовой) информации. В [7] представлены выводы о том, что автора поста можно определить по уникальному стилю письма. В [8] используется метод, который учитывает не только текстовую информацию, публикуемую пользователем в посте, но и ассоциированную с ним метаинформацию: геолокация, время публикаций, хештеги и т.п.

2.2. Программные сервисы поиска пользователей в социальных сетях

FindFace-SearchFace. В настоящее время существует несколько сервисов для поиска профилей людей в социальных сетях. Большинство сервисов работают по принципу обычных поисковых систем - загружают все доступные открытые данные о профиле и сохраняют в локальную базу данных.

Одним из таких сервисов является система FindFace и ее бесплатный аналог SearchFace [9], которые позволяют находить профиль человека в социальной сети по его фотографии. Чтобы начать поиск, нужно выбрать фотографию, где отчетливо видно человеческое лицо, и загрузить снимок. Алгоритм найдет страницы с похожими фото и выложит ссылки на них с примерами изображений. Возле каждой ссылки будет стоять рейтинг от 0 до 1. Если показатель больше 0,67, то это значит, что система зафиксировала максимально полное совпадение. Разработанная нейросеть просканировала лица 500 миллионов пользователей социальной сети VKontakte. На данный момент SearchFace позиционируется как сервис для знакомств, который пока находится на начальном этапе развития.

Сервис «Поиск людей». Компания Яндекс создала сервис «Поиск людей» - это специализированный сервис, с помощью которого удобно и быстро находить размещенные в открытом доступе профили людей в социальных сетях. Система умеет группировать профили, принадлежащие одному и тому же человеку. Благодаря этому, поиск упрощается: на выдаче по запросу помещается больше результатов, и пользователь может выбрать, в какой социальной сети ему удобнее общаться с найденным человеком. Система индексирует в локальную базу открытые профили людей и проводит поиск по загруженным данным. [10]

Яндекс.Люди. В системе Яндекс.Люди для поиска используются текстовые данные, полученные из профилей социальной сети. Так из профилей человека выгружаются следующие данные:

- ФИО пользователя (или хотя бы один из параметров, позволяющих идентифицировать человека).
- Возраст пользователя.
- Место проживания или адрес пользователя.
- Место учебы или законченные учебные заведения.
- Место работы пользователя.

Сервис Яндекс.Люди работает согласно следующим принципам:

- Если профиль человека в соцсети скрыт настройками приватности, то информация о нем не будет отображаться в результате поиска.

- В выдаче отображаются только те сведения, которые доступны пользователям данной соцсети без регистрации.
- Аккаунты из разных соцсетей связываются Яндексом в группу, только есть точно видно, что они принадлежат одному и тому же человеку. Например, если из профиля одной соцсети была проставлена ссылка на профиль в другой.
- Сам сервис не хранит какой-либо информации, а действует только как поисковик по другим страницам Интернета (в данном случае — профиля социальных сетей, открытых для общего доступа). [11]

Загруженные профили формируются в похожие группы, что позволяет при идентификации пользователя в одной из сетей найти его аналоги в других сетях. Система Яндекс. Люди представлена на рисунке 1.

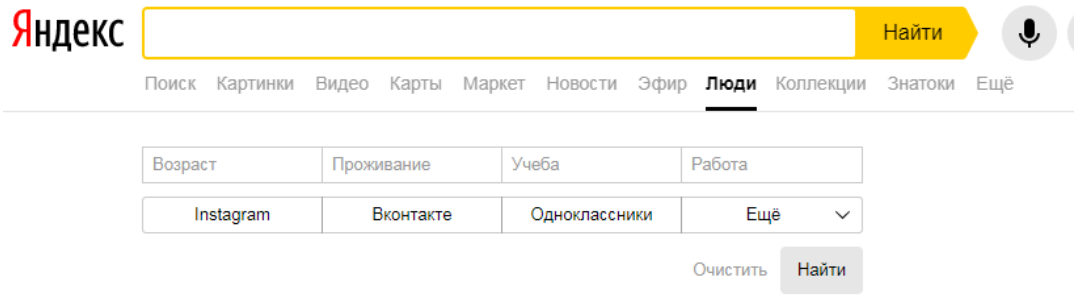

Рисунок 1. Система Яндекс.Люди.

Несмотря на наличие данных сервисов, решающих отдельные задачи поиска пользователей социальных сетей, в настоящее время не существует комплексных подходов и универсальных сервисов, позволяющих осуществлять сопоставление профилей пользователей в различны социальных сетях посредством анализа не только данных анкет, но и слабоструктурированной информации со страниц соответствующих аккаунтов.

3. Подход к маппингу профилей пользователей в социальных сетях

Входным значением для разработанного алгоритма маппинга профилей пользователей является один из профилей человека в социальной сети. Из данного профиля загружаются все возможные данные о человеке, и по этим данным формируется единая модель искомого профиля, которая включает следующие параметры:

1. Ссылка на профиль;
2. ФИО пользователя;
3. Дата рождения;
4. Место проживания;
5. Место рождения;
6. Список друзей пользователя;
7. Список постов;
8. Данные о месте работы, предыдущих местах работы;
9. Данные о месте учебы, предыдущих местах учебы;
10. Список контактов пользователя, ссылки на другие сайты, телефон;
11. Аватар профиля, а также фотографии из профиля.

Поиск схожих профилей происходит путем поиска по полям, загруженным из исходного профиля. По загруженным схожим профилям формируется рейтинг, по которому в последствии происходит сортировка. Рейтинг включает в себя следующие критерии:

- *Критерий наличия схожих фотографий и лиц на фотографиях.* Для определения лиц был разработан сервис на языке python, который работает с помощью библиотеки DLIB [12], которая позволяет находить лица на фотографиях и генерировать векторное представление найденного образа. В последствии происходит сравнение векторных норм матриц, и если

вектора совпадают с определенной долей отклонения, то профиль выводится в первую очередь.

- *Критерий наличия схожих контактов* определяется путем нахождения в профиле совпадающих ссылок.
- *Критерий наличия схожего места работы и места учебы.* Для вычисления данного показателя строки проходят предобработку (очищаются от знаков пунктуации, приводятся к нижнему регистру, затем строки лемматизируются), и вычисляется процент совпадающих лемм согласно следующей модели:

$$\frac{n}{\max(l1,\ l2)} \geq 0{,}55,$$

где n – количество попарно совпадающих лемм;
$l1$, $l2$ – количество лемм в 1 и 2 названии;

- *Критерий наличия схожих постов.* Данный показатель определяется с помощью нахождения расстояния Левенштейна (редакционное расстояние, дистанция редактирования) — минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую:

$$D(i,j)=\begin{cases} 0, i=0, j=0 \\ i, j=0, i>0 \\ j, i=0, j>0 \\ min\ \{(D(i,j-1)+1, D(i-1,j)+1, D(i-1,j-1)+m(S[i],S[j])\}, j>0, i>0 \end{cases}$$

В качестве второго метода нахождения схожих постов был реализован алгоритм шинглов. Данный алгоритм работает по принципу разбиения текста на шинглы, вычисления хэшей данных шинглов, попарное сравнение хешей. Визуальное представление алгоритма шинглов представлена на рисунке 2.

Рисунок 2. Алгоритм шинглов.

- *Критерий наличия схожих друзей.* Данный показатель вычисляется путем попарного сравнения имен и вычисления процента совпадений.

4. Реализация программной системы маппинга профилей пользователей в социальных сетях

Для проверки эффективности предложенного подхода была реализована программная система маппинга профилей пользователей в социальных сетях. Разработанная система представляет собой клиент серверной приложение, где сервер является web-сервисом на языке Java, разработанным с помощью программной платформы Spring Boot. Spring Framework — универсальный фреймворк с открытым исходным кодом для Java-платформы. Также существует форк для платформы .NET Framework, названный Spring.NET. [13]

Клиент является web приложением, разработанным средствами JavaScript библиотеки React. React — JavaScript-библиотека с открытым исходным кодом для разработки пользовательских интерфейсов. React разрабатывается и поддерживается Facebook, Instagram и сообществом отдельных разработчиков и корпораций. React может использоваться для разработки одностраничных и мобильных приложений. [14]

Система интегрируется с тремя самыми популярными социальными сетями в СНГ: «ВКонтакте», «Одноклассники» и «Facebook». Данные из социальной сети «ВКонтакте»

загружаются путем интеграции с бесплатным сервисом vk.api.[15]. Данные из социальных сетей «Одноклассники» и «Facebook» загружаются путем парсинга основных полей профилей.

В качестве входных данных система принимает ссылку на профиль в одной из социальных сетей. Из данного профиля загружаются все возможные данные о человеке, и по этим данным формируется единая модель искомого профиля. Поиск схожих профилей происходит путем поиска по полям, загруженным из исходного профиля. По загруженным схожим профилям формируется рейтинг, по которому в последствии происходит сортировка.

Схожие профили сортируются по полученному рейтингу и отображаются на web-форме. Пример работы системы представлен на рисунке 3.

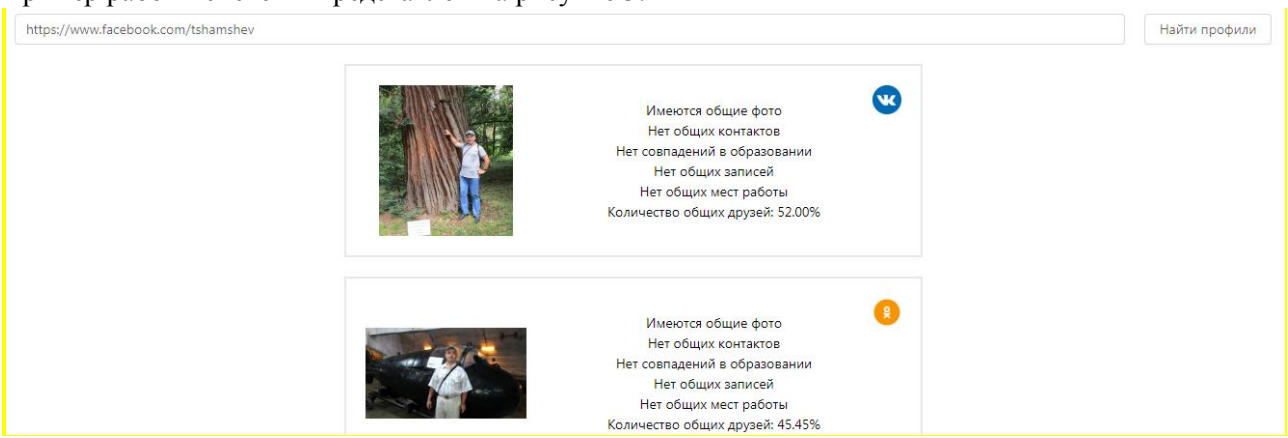

Рисунок 3. Результат работы программной системы.

5. Заключение

Таким образом, в рамках данной работы, был предложен комплексный подход по единоличной идентификации человека в нескольких разных социальных сетях посредством анализа не только данных анкет, но и слабоструктурированной информации со страниц соответствующих аккаунтов, а также графической информации.

В результате проделанной работы, была разработана программная система, выполняющая функцию поиска схожих профилей в социальных сетях. Приложение показало неплохие результаты и может быть использовано как открытый проект для поиска людей. Предложенная методика закладывает основу для дальнейшей работы по проведению соответствующих экспериментов, выработке новых алгоритмов.

6. Благодарности

Работа выполнена при финансовой поддержке РФФИ. Проекты № 18-47-730035 и 18-47-732007.

7. Литература

[1] Социальные сети в 2018 году: глобальное исследование [Электронный ресурс]. – https://www.web-canape.ru/business/socialnye-seti-v-2018-godu-globalnoe-issledovanie/ (Дата обращения: 21.12.2019).

[2] Рыцарев, И.А. Кластеризация медиа-контента из социальных сетей с использованием технологии BigData / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921-927.

[3] Павлыгин, Э.Д. Разработка программного комплекса для интеллектуального анализа социальных медиа / Э.Д. Павлыгин, А.Г. Подлобошников, Р.А. Савинов, Н.Г. Ярушкина, А.М. Наместников, А.А. Филиппов, А.А. Романов, В.С. Мошкин, Г.Ю. Гуськов, М.С. Григоричева // Автоматизация процессов управления. – 2019. – № 2(56). – С. 23-36.

[4] Ярушкина, Н.Г. Разработка программной системы семантического анализа контента социальных медиа / Н.Г. Ярушкина, В.С. Мошкин, А.А. Филиппов, Г.Ю. Гуськов, А.А. Романов, А.М. Наместников //Радиотехника. – 2018. – № 6. – С. 73-79.

[5] Gaewon, Y. SocialSearch: Enhancing Entity Search with Social Network Matching / Y. Gaewon, H. Seungwon, N. Zaiqing, W. Ji-Rong // EDBT/ICDT: Proceedings of the 14th International Conference on Extending Database Technology, 2011. DOI: 10.1145/1951365. 1951428.
[6] Motoyama, M. I Seek You – Searching and Matching Individuals in Social Networks / M. Motoyama, G. Varghese // WIDM '09: Proceeding of the eleventh international workshop on Web information and data management, 2009. DOI: 10.1145/1651587.1651604.
[7] Raad, E. User Profile Matching in Social Networks / E. Raad, R. Chbeir, A. Dipanda // 13th International Conference on Network-Based Information Systems (NBiS), 2010.
[8] Vosecky, J. User identication across multiple social networks / J. Vosecky, D. Hong, V.Y. Shen // Proceedings of First International Conference on Networked Digital Technologies, 2009.
[9] Система поиска клонов [Электронный ресурс]. – Режим доступа: http://searchface.ru/ (Дата обращения: 13.12.2019).
[10] Поиск людей [Электронный ресурс]: Режим доступа: https://yandex.ru/ support/peoplesearch/ (Дата обращения: 16.12.2019).
[11] Яндекс.Люди [Электронный ресурс]. – Как искать людей по социальным сетям. – Режим доступа: https://ktonanovenkogo.ru/web-obzory/yandeks-lyudi-poisk-lyudej-socialnyh-setyah. html (Дата обращения: 18.12.2019).
[12] Dlib [Электронный ресурс]. – Режим доступа: http://dlib.net/–DLibC++library (Дата обращения: 21.12.2019).
[13] Spring Boot [Электронный ресурс]. – Режим доступа: https://spring.io/projects/spring-boot (Дата обращения: 21.12.2019).
[14] React JS [Электронный ресурс]. – Режим доступа: https://ru.reactjs.org/ (Дата обращения: 21.12.2019)
[15] VK. API [Электронный ресурс]. – Режим доступа: https://vk.com/dev/methods – Описание методов API (Дата обращения: 21.12.2019).

An integrated approach to mapping user profiles on social networks

V.A Belov[1], D.S. Drozdov[1], R.A. Shakurov[1], V.S. Moshkin[1], I.A. Andreev[1]

[1]Ulyanovsk State Technical University, Severny Venets str. 32, Ulyanovsk, Russia, 432027

Abstract. This paper considers an integrated approach to the search for similar user profiles in social networks. The described method allows you to identify a person's profile in various social networks based on the analysis of structured, unstructured and graphical profile data.