УДК 004.624

Разработка средств автоматизации поиска информации о человеке в открытых источниках сети Интернет

А. Ю. Самойлов, Е. Ю. Никитина
Пермский государственный национальный исследовательский университет
Россия, 614990, г. Пермь, ул. Букирева, 15
arksam@list.ru; +7 961 758-29-61

Рассмотрены сервисы, позволяющие получить информацию о человеке из открытых источников. Проанализированы функциональные возможности существующих средств поиска, в результате чего были выявлены их достоинства и недостатки. Разработан программный продукт для нахождения и обработки информации о человеке в открытых источниках сети Интернет.

Ключевые слова: *поиск информации; открытые источники; анализ информации; обработка информации.*

DOI: 10.17072/1993-0550-2020-1-74-79

Введение

На сегодняшний день социальные сети стали неотъемлемой частью жизни людей. Многие пользователи не представляют себе повседневную жизнь без них. Такие сервисы как "ВКонтакте", Одноклассники, Facebook и другие имеют огромный список того, какую пользу обществу они могут принести. Ежедневно люди листают ленту новостей, переписываются с другими людьми, слушают музыку. В социальных сетях пользователи продвигают бизнес, а также находят клиентов.

Однако социальные сети имеют и негативные стороны. В последнее время в социальных сетях распространяется запрещенный контент, создаются группы, занимающиеся противозаконной деятельностью. Полагаем, что распространители подобной информации, как минимум, должны находиться под пристальным контролем, который поможет избежать негативных последствий их деятельности.

Как известно, существующие сервисы социальных сетей предоставляют некоторые инструменты, позволяющие получать информацию о человеке. Эти инструменты предоставляют разные методы и форматы при работе с информацией – в результате возникает необходимость написания единой программы для различных сервисов, которая позволит использовать данные инструменты более эффективно, автоматизируя процесс поиска, сохранения и анализа полученных данных из разных информационных источников.

1. Сравнительный анализ

Интернет – пространство с огромным количеством пользователей. Некоторые пользователи могут представлять особый интерес для государственных органов, в силу чего необходим инструмент, который позволит получить и хранить информацию о таких пользователях. Необходимая для анализа информация о человеке представлена в целом ряде информационных источников, которые являются как закрытыми, так и открытыми (рис. 1).

На данный момент имеется возможность практически свободного получения такой информации из открытых источников данных, которые не только хранят информацию о людях, но и готовы поделиться ей. В результате возникла задача написания сервиса, позволяющего искать, накапливать и анализировать полученную информацию о человеке из открытых источников.

Рис. 1. *Виды источников информации*

Рассмотрим существующие сервисы с похожим функционалом. В данной работе будут рассмотрены сервисы, позволяющие получить какую-либо информацию о пользователях из открытых источников и анализировать ее. Среди них:

1) "Герда-бот";
2) "Инцидент менеджмент";
3) Поисковая система "СЕУС".

1.1. Сервис "Герда-бот"

В социальных сетях, в том числе во "Вконтакте", встречаются опасные сообщества (группы смерти, руферы), в них можно многое узнать про наркотики, зацепинг и прочие детско-подростковые искушения. Количество таких групп постоянно растет, названия и содержание меняются, поэтому пользователю не сразу открывается правда о том или ином ресурсе. Особую опасность такие ресурсы представляют для детей, которые по незнанию или из любопытства могут попасть в опасное сообщество, подвергнуться там запугиванию и будут бояться его покинуть.

Для родителей существует специальный сервис – "Герда-бот". "Герда-бот" – это бот, который анализирует группы с опасным контентом социальной сети "ВКонтакте" и сообщает родителям, если ребенок стал участником одной из них [1].

Достоинства:

1) имеют пополняемую базу опасных групп;
2) система в режиме реального времени проводит мониторинг опасных групп.

Недостатки:

1) поиск осуществляется только во "ВКонтакте";
2) результатом работы является информация, связанная с деятельностью пользователя в группах/сообществах.

1.2. Система мониторинга "Инцидент менеджмент"

Мнение граждан в социальных сетях может быть передано в соответствующие структуры в автоматическом режиме. Для этого в России используется система мониторинга "Инцидент менеджмент" позволяющая отслеживать реакцию населения на действия властей [2].

"Инцидент менеджмент" контролирует пять социальных сетей – "ВКонтакте", "Facebook", "Instagram", "Twitter" и "Одноклассники". Мониторинг осуществляется по ключевым словам.

Данная программа позволяет властям оценивать свои действия без всяких опросов и референдумов, опираясь на результаты мониторинга. В соответствии с полученными данными они будут предпринимать те или иные действия. Это поможет государству узнать реальные проблемы населения. С другой стороны, система "Инцидент менеджмент" – это контроль над всеми высказываниями в интернете.

Достоинства:

1) анализ в режиме реального времени;
2) поддержка нескольких социальных сетей;

Недостатки:

1) результатом работы является активность пользователя в определенной группе.

1.3. Поисковая система "СЕУС"

Система "СЕУС" предназначена для использования представителями правоохранительных органов с целью выявления информации, свидетельствующей о подготовке или осуществлении противоправной деятельности или агитации и пропаганде социально опасных явлений.

В ходе проведения мероприятий по противодействию информационному экстремизму "СЕУС" может быть использована для своевременного выявления фактов и источников размещения информационных материалов соответствующей тематической направленности с це-

лью предотвращения их дальнейшего распространения [3].

"СЕУС" осуществляет поиск информации в следующих сервисах: "ВКонтакте", "Facebook", "Instagram", "Twitter", "Одноклассники".

Достоинтва:

1) поддержка нескольких социальных сетей;
2) большое количество фильтров, по которым производится поиск.

Недостатки:

1) Нет поиска по группам.

1.4. Результат анализа систем

Рассмотренные выше сервисы не позволяют в полной мере выполнить задачу по сбору и анализу информации о человеке в открытых источниках сети Интернет. Основные факторы, препятствующие этому:

1) ограниченные источники получения информации. Каждая из систем позволяет осуществлять поиск информации только по ограниченному количеству источников.
2) решаемая системами задача собирает небольшую часть доступной информации – только крайне ограниченный набор параметров о пользователе.

Также эти сервисы обладают общими недостатками:

1) платное использование;
2) база данных хранится на стороннем ресурсе. Этот факт накладывает дополнительное ограничение при работе с данными.

Данные недостатки текущих сервисов подтверждают актуальность поставленной задачи.

2. Описание решения поставленной задачи

Для решения поставленной задачи на первом этапе в качестве информационного источника была использована социальная сеть "ВКонтакте".

2.1. Выбор языка программирования

В качестве языка программирования для реализации был выбран язык Java. Основные критерии, по которым он был выбран:

1) мультиплатформенность.

Java – мультиплатформенный язык программирования. Это значит, что программы, написанные на языке Java, можно выполнять на любой платформе, где установлена специальная исполняющая система Java – Java Virtual Machine (JVM).

2) наличие библиотек

Для данного языка разработано множество библиотек, которые упрощают работу и делают разработку более удобной.

2.2. Разработка архитектуры приложения

Для каждого сервиса, в котором необходимо осуществить поиск информации о человеке, нужно разработать свой модуль. Каждый модуль сервиса позволяет находить людей по введенным параметрам, и получать детальную информацию по конкретному человеку.

Сборщик сервисов позволяет работать со всеми модулями сервисов одновременно. Также он реализует функционал по сохранению детальной информации о человеке в базу данных и файл Excel.

Сборщик сервисов имеет в своем составе графический модуль, который представляет собой пользовательский интерфейс для взаимодействия со сборщиком сервисов, и состоит из трех панелей.

Левая панель содержит сервисы, по которым возможен поиск информации.

Центральная панель содержит параметры, по которым осуществляется поиск. Доступные поля для поиска:

- фамилия;
- имя;
- дата рождения;
- место работы;
- населенный пункт;
- идентификатор пользователя;
- название группы;
- идентификатор группы.

Правая панель отображает найденную информацию о человеке.

При заполнении полей и нажатии кнопки "Искать", выполняется поиск по сервисам с последующим построением отчета.

Для каждого сервиса отображается свой блок, в котором расположены блоки с информацией о людях: фамилия, имя, отчество, фото профиля. После выбора интересующего человека и нажатия кнопки "Сформировать отчет", выполняется поиск детальной информации об указанном человеке в сервисах, и строится детальный отчет.

Кнопка "Сформировать отчет xls" открывает проводник, который позволяет сохранить файл Excel с текущим детальным отчетом.

Кнопка "Сохранить в базу" сохраняет текущий детальный отчет в базу данных.

Описанная архитектура (рис. 2) позволяет добавлять модули поиска для любых открытых источников информации, расположенных в интернете, и работать с ними через единый интерфейс.

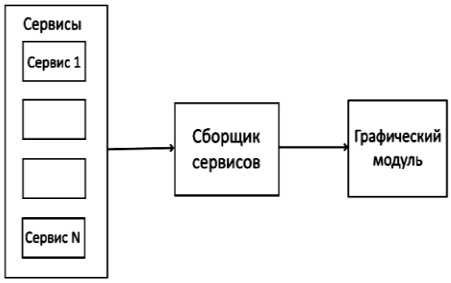

Рис. 2. *Архитектура поискового модуля*

Согласно поставленным требованиям, приложение для анализа найденной информации о человеке (рис. 3) должно иметь:

- набор полей для ввода, по которым может осуществляться поиск;
- набор полей, которые могут быть включены в отчет;
- средства для построения отчета на основе полученной информации.

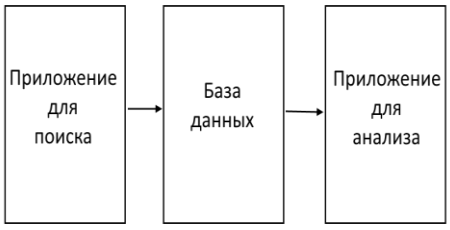

Рис. 3. *Взаимодействие приложений*

2.3. Выбор средства взаимодействия с социальной сетью "Вконтакте"

Для работы с социальной сетью "ВКонтакте" использовалось официальное API [4] (application programming interface), которое по запросу предоставляет необходимую информацию. Использовались следующие запросы:

1) users.search – поиск пользователей;
2) users.get – получение общей информации о пользователе;
3) database.getCities – поиск города;
4) wall.get – получение информации со стены;
5) friends.get – получение списка друзей;
6) photos.getAlbums – получение списка альбомов пользователя;
7) photos.get – получение фотографий из альбома;
8) groups.get – получение списка сообщетсв пользователя;
9) groups.search – поиск групп;
10) groups.getMembers – получение участников сообщества.

Для работы со всеми методами API необходимо передавать в запросе access_token – специальный ключ доступа. Он представляет собой строку из латинских букв и цифр и может соответствовать отдельному пользователю, сообществу или самому приложению.

Поддерживается три способа получения ключа доступа:

- Implicit flow – самый короткий и простой вариант. Ключ возвращается на устройство пользователя, где был открыт диалог авторизации (в виде дополнительного параметра URL);
- Authorization code flow – двухэтапный вариант с дополнительной аутентификацией сервера;
- Client credentials flow – авторизация по секретному ключу приложения. Этот подход необходимо использовать только для доступа к специальным secure-методам.

В данном модуле сервиса используется вариант Implicit flow. Время жизни токена – бессрочное. Для выполнения запросов также необходимо получить один из следующих ключей доступа:

- ключ доступа пользователя;
- ключ доступа сообщества;
- сервисный ключ доступа.

В модуле сервиса используется ключ доступа пользователя. Такой ключ требуется для работы со всеми методами API, за исключением методов секции secure. Ключ доступа – своего рода "подпись" пользователя в приложении. Он сообщает серверу, от имени какого пользователя осуществляются запросы, и какие права доступа он выдал приложению.

Получить ключ доступа пользователя можно с помощью Implicit flow. Текущие запрашиваемые права:

1) Offline – доступ к API в любое время (при использовании этой опции параметр

expires_in, возвращаемый вместе с access_token, содержит 0 – токен бессрочный);

2) Photos – доступ к фотографиям.

На данный момент API имеет следующие ограничения:

1) частотные ограничения [5].

К методам API ВКонтакте (за исключением методов из секций secure и ads) с ключом доступа пользователя или сервисным ключом доступа можно обращаться не чаще трех раз в секунду.

2) количественные ограничения.

Помимо ограничений на частоту обращений, существуют и количественные ограничения на вызов однотипных методов. Информация о точных лимитах не предоставляется.

После превышения количественного лимита доступ к конкретному методу может требовать ввода капчи, а также может быть временно ограничен (в таком случае сервер не возвращает ответ на вызов конкретного метода, но без проблем обрабатывает любые другие запросы).

Стоит отметить, что частотные ограничения можно обойти двумя путями:

- использование временных задержек между запросами;
- использование специального метода execute, позволяющего совершить до 25 обращений к разным методам в рамках одного запроса.

На данном этапе приложение позволяет доставать следующее количество информации в рамках детального отчета:

- получение друзей пользователей – 1000 человек;
- получение информации со стены – 100 записей;
- получение фото – 25 альбомов, до 1000 фото в каждом.

Что касается общего поиска пользователей, то существует следующее ограничение – API выдает информацию только о первых найденных 1000 пользователей. Данное ограничение на текущий момент времени не предоставляет возможности обойти при использовании API. Кроме того, некоторые пользователи вследствие нарушений правил сайта могут не попасть в данную выборку.

2.4. Выбор информации для обработки

Блоки информации, которые можно получить о пользователе из социальной сети ВКонтакте:

- общая информация;
- друзья;
- стена;
- фотографии;
- группы;
- подписчики;
- видеозаписи;
- подарки и т.д.

Блоки информации, которые обрабатываются приложением:

1) общая информация – содержит основную информацию со страницы пользователя;
   - идентификатор;
   - фамилию, имя, отчество;
   - время последнего посещения, текущий статус в сети;
   - количество фотоальбомов, видеозаписей, аудиозаписей, фотографий, заметок, друзей, сообществ, видеозаписей с пользователем, подписчиков, объектов в блоке "Интересные страницы";
   - дату рождения;
   - ссылку на фото профиля;
   - мобильный, дополнительный телефон;
   - поля "О себе", деятельность, карьера, место учебы;
   - статус на странице;
   - любимые фильмы, книги;
   - контакты в "Twitter", "Facebook", Livejournal, Instagram, Skype;
2) друзья – содержит информацию о друзьях
   - фамилию, имя, отчество;
   - ссылку на фото профиля;
3) стена (текст записи) – содержит текстовую информацию из записей со стены, в которых есть данная информация.
4) фотографии.

Выбор данных блоков информации объясняется тем, что они содержат большую часть информации среди прочих блоков. Блоки информации, которые можно получить о сообществе – "участники сообщества".

2.5. Ключевые используемые библиотеки

- com.vk.api – работа с SDK Java "ВКонтакте";
- com.google.code.gson – обработка ответов SDK, для которых нет необходимых объектов;
- org.apache.poi работа с файлами xls.

2.6. Реализация новых модулей сервисов

В составе разработанного программного продукта для автоматизации поиска информации о человеке в открытых источниках сети Интернет разработаны следующие модули:

- Service – интерфейс для разработки модулей сервисов;
- ServiceExecutor – сборщик сервисов;
- ServiceGUI – графический модуль;
- VkService – модуль сервиса ВКонтакте.

Чтобы ServiceExecutor или ServiceGUI могли работать с модулями сервисов, необходимо создать в директории данных модулей папку "plugins" и разместить в эту папку реализованные модули сервисов. В модуле Service описаны основные методы и классы, которые необходимы для разработки модуля сервисов:

- List<Fields> getAvailableFields – получение полей, по которым возможен поиск в данном сервисе;
- List<User> searchMainInfo – поиск пользователей внутри сервиса;
- DetailReport searchDetailInfo – получение детального отчета по пользователю;
- List<User> searchGroupInfo – поиск сообществ внутри сервиса;
- DetailReport searchGroupDetailInfo – получение детального отчета по группе.

Сервис также должен содержать файл с настройками config.properties, в котором описываются следующие параметры:

- service.id – уникальный id сервиса;
- service.name – имя сервиса;
- main.class – название класса, реализующего методы Service.

Заключение

На сегодняшней день не существует программного продукта для решения поставленной задачи в полном объеме. Были рассмотрены и проанализированы наиболее подходящие по функциональным возможностям аналоги, а именно: "Герда-бот", "Инцидент менеджмент", поисковая система "СЕУС". Данные сервисы имеют частично как необходимый функционал, так и недостатки.

Главным недостатком является то, что в силу специфичности решаемых вышеуказанными системами задач они не позволяют манипулировать тем объемом информации о человеке, который открытые источники предоставляют. Общими недостатками являются платное использование сервисов и хранение баз данных на стороннем ресурсе. Были описаны и обоснованы решения, принятые в ходе проектирования и разработки нового сервиса для автоматизации поиска информации о человеке в открытых источниках сети Интернет.

Список литературы

1) Gerda Bot – Мониторинг детей в социальных сетях, 2018–2019. URL: https://gerdabot.ru/ (дата обращения: 12.11.2019).
2) Медиалогия, 2003–2019. URL: https://www.mlg.ru/products/im/ (дата обращения: 12.11.2019).
3) *Руководство* пользователя ПС СЕУС: 2019. 59 с.
4) *Java* SDK: Социальная сеть "ВКонтакте", 2006–2019. URL: https://vk.com/dev/Java_SDK. (дата обращения: 20.10.2019).
5) *Выполнение* запросов к API "ВКонтакте": Социальная сеть "ВКонтакте", 2006–2019. URL: https://vk.com/dev/api_requests. (дата обращения: 12.11.2019).

Development of tools for automated search of information about a person in open web sources

A. Yu. Samoilov, E. Yu. Nikitina
Perm State University; 15, Bukireva st., Perm, 614990, Russia
arksam@list.ru; +7 961 758-29-61

The paper is concerned with services that allow one to get information about a person from open sources. The functionality of existing search tools has been analyzed, their advantages and disadvantages have been identified. A software product for finding and processing information about a person in open sources on the Internet has been developed.

Key words: *information search; open sources; information analysis; information processing.*