ISSN 2782-2001
Системы анализа
и обработки данных
том 81, № 1, 2021, с. 73–84

http://journals.nstu.ru/vestnik
Analysis and data
processing systems
Vol. 81, No. 1, 2021, pp. 73–84

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND CONTROL

УДК 004.415.2

DOI: 10.17212/2782-2001-2021-1-73-84

Начальные этапы проектирования системы сбора и предиктивного анализа данных социальных медиа*

И.С. КАЛЫТЮК[a], Г.А. ФРАНЦУЗОВА[b], А.В. ГУНЬКО[c]

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет

[a] *ivankalytyuk@yandex.ru* [b] *frants@ac.cs.nstu.ru* [c] *gun@ait.cs.nstu.ru*

В настоящей статье рассматривается проектирование системы сбора и предиктивного анализа социальных медиа. По мере развития сети Интернет, а также социальных медиа более простыми стали доступ и распространение информации, ведь сами пользователи сети являются одновременно создателями и получателями различной информации. Для получения новых знаний, которые могут быть полезны пользователям социальных медиа, возможно использование предиктивной (прогнозной) аналитики – комплекса методов статистического анализа, которые извлекают новую информацию из текущих и исторических данных. Такой метод анализа данных социальных медиа находится на стадии своего развития.

В основе предиктивной аналитики лежит автоматический поиск связей, аномалий и закономерностей между различными факторами. Для формирования прогнозной модели используется большой набор статистических методов моделирования, интеллектуальный анализ данных, машинное обучение, нейронные сети и другие механизмы. В совокупности с различными методами сбора информации с интернет-ресурсов, таких как парсинг и API социальных сетей, предиктивная аналитика может предлагать наиболее интересные для пользователя источники информации. Для того чтобы объединить методы предиктивного анализа и методы сбора данных, требуется внимательно отнестись к процессу проектирования системы.

В работе предложено формальное описание данных, которые использует будущая система. Помимо этого, выделены общая архитектура и алгоритм функционирования. Особое внимание обращено на подробное описание одной из основных частей системы (подсистемы сбора). Полученные результаты будут использоваться при дальнейшем проектировании, планируется рассмотрение подсистемы аналитики. Последующая работа над темой позволит детализировать архитектуру и алгоритм функционирования.

Ключевые слова: социальные медиа, предиктивный анализ, проектирование системы, архитектура, алгоритм функционирования, сбор данных, API, парсинг

* *Статья получена 24 августа 2020 г.*

ВВЕДЕНИЕ

Как известно, на сегодняшний день самым массовым и оперативным источником информации является Интернет, где существенное значение имеют социальные медиа, влияние которых на жизнь современного человека всё возрастает [1]. Основная разновидность социальных медиа – социальные сети. Из наиболее известных можно выделить следующие: Facebook, VK, Instagram, YouTube, Twitter, «Одноклассники», а также мессенджеры WhatsApp, Telegram. Важнейшие функции социальных сетей – влияние на восприятие, отношение и конечное поведение потребителей. Поскольку социальные сети – это огромная база данных, где представлена различная информация о миллионах людей, то это привлекает специалистов из самых разных областей.

В связи с этим важное значение приобретает анализ размещенной в социальных сетях информации и возможность прогнозирования будущих событий. Однако методы предиктивного анализа этих сетей находятся еще на стадии развития [2]. Отсюда очевидна актуальность внедрения таких методов для анализа данных социальных медиа, а также оценка применимости методов предиктивной аналитики.

Целью настоящей работы является проектирование системы сбора и предиктивного анализа данных социальных медиа на основе различных методов сбора данных из социальных сетей [3].

Основной проблемой при проектировании такой системы является разработка ее общей архитектуры и алгоритма функционирования. Следует достаточно четко понимать, какие подсистемы и модули должны быть в составе системы, иначе в будущем при разработке может возникнуть необходимость добавления или удаления некоторых частей.

Структура работы следующая: в первом и втором разделах представлена постановка задачи и формальное описание данных, которые будут использоваться в системе. Общий алгоритм функционирования и архитектура системы описаны в третьем разделе, а проектирование подсистемы сбора и связанных с ней модулей представлено в четвертом разделе.

1. ПОСТАНОВКА ЗАДАЧИ

Необходимо разработать эффективную систему сбора и предиктивного анализа данных социальных медиа. Целью функционирования системы является представление пользователю интернет-источников, которые могут быть ему интересны в будущем по настоящим данным.

К системе предъявляются следующие требования.

- Четкое описание входных данных.
- Необходимость структуризации собранных данных.
- Ограничение количества собранных данных для получения приемлемых результатов.
- Преобразование данных к виду, пригодному для анализа.
- Определение количества результатов.
- Четкое описание выходных данных.

Для достижения поставленной цели с учетом предъявляемых требований общую задачу разделим на ряд подзадач:

1) определение входных данных и структуризация;

2) определение ресурсов собранных данных;

3) определение выходных данных.

В свою очередь, каждой из подзадач соответствуют свои требования из общего перечня. Для определения входных данных и их структуризации необходимо иметь четкое описание входных данных; для решения второй подзадачи следует ограничить количество собранных данных с целью получения приемлемых результатов и преобразовать эти данные к виду, пригодному для анализа. Решение третьей подзадачи предполагает определение количества результатов и четкое описание выходных данных.

Схематично общая система предиктивного анализа с выделенными подзадачами представлена на рис. 1.

Рис. 1. Схематичное представление постановки задачи

Fig. 1. Schematic representation of the problem statement

Далее рассмотрим решение каждой из подзадач.

2. ОПИСАНИЕ ЗАДАЧ И ФОРМАЛЬНОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

Отметим, что входными данными для системы являются данные из первой подзадачи, которые формально можно описать следующим образом:

$$X = \{x_1, x_2, ..., x_i, ..., x_n\}, \quad (1)$$

где x_i – одно ключевое слово, которое будет использоваться для сбора данных; $i \in [1, n]$, n – количество ключевых слов. Входные данные могут быть единичными (поиск по одному ключевому слову, упоминанию пользователя) – в этом случае $X = x_1$. В случае массива учитывается только совместное появление в каком-либо месте интернет-ресурса.

Необходимо понимать, что сами по себе отдельные данные, которые будут собраны, не являются достаточными для дальнейшей работы. Для того чтобы использовать эту информацию, ее необходимо преобразовать в структурированный вид [4]. Можно привести различные методы для сбора данных социальных сетей, а именно методы получения данных через API и через

парсинг различных интернет-ресурсов. Существуют программные комплексы, использующие данные методы, которые помимо сбора данных сразу работают над их структуризацией [3].

Вторая подзадача предполагает определение тех ресурсов, на которых встречаются введенные пользователем ключевые слова. Так как после сбора данных будет производиться предиктивный анализ, имеет смысл выбирать только последние (по дате) ресурсы из сети Интернет. Предлагается до начала анализа ограничить количество ресурсов числом 3000.

Предполагается, что имеет место комбинированное использование разработки правил парсинга и изучения каждого из API социальных сетей. Возможны общие правила парсинга, однако всегда нужно анализировать попадающие системе ресурсы из сети Интернет и выбирать, в каком случае можно использовать запросы к API, а в каком – парсинг [5].

Представление промежуточных данных с учетом (1) имеет вид

$$X^* = \left\{ \begin{array}{l} A_1(x_1,\ x_2,...,\ x_i,...,\ x_n); \\ A_2(x_1,\ x_2,...,\ x_i,...,\ x_n); \\ \qquad ... \\ A_j(x_1,\ x_2,...,\ x_i,...,\ x_n); \\ \qquad ... \\ A_m(x_1,\ x_2,...,\ x_i,...,\ x_n) \end{array} \right\}, \qquad (2)$$

где A_j – ресурс, на котором упоминаются одновременно все ключевые слова; $j \in [1, m]$, m – количество полученных ресурсов (ограничено 3000).

Как видно, входные данные для анализа можно представить в виде вектора, включающего в себя интернет-ресурсы. После определения всех ресурсов, по которым был произведен сбор, исходные входные данные отбрасываются и оставляется только список полученных ресурсов. В результате (2) можно записать в виде вектора-столбца:

$$A^* = (A_1 A_2 \dots A_j \dots A_m)^T . \qquad (3)$$

Выходные данные системы – это данные третьей подзадачи, которые представляют собой вектор:

$$Y = \{y_1,\ y_2,...,\ y_i,...,\ y_n\}, \qquad (4)$$

где y_i – один результат анализа (новые знания, полученные в результате функционирования системы); $i \in [1, n]$, n – количество желаемых результатов. Предполагается, что количество желаемых результатов задает пользователь системы.

Поскольку результатом анализа системы должны быть ресурсы, на которых, возможно, в будущем появится интересная пользователю информация, то запись (4) по аналогии с (3) преобразуется к виду

$$Y^* = \left(A_1^* A_2^* \dots A_j^* \dots A_m^* \right)^T, \tag{5}$$

где A_j^* – один из ресурсов – результатов анализа по промежуточным данным вектора A^* (из него выбираются лишь некоторые, в зависимости от количества желаемых результатов, заданного пользователем); $j \in [1, m]$.

3. ОБЩИЙ АЛГОРИТМ ФУНКЦИОНИРОВАНИЯ И АРХИТЕКТУРА СИСТЕМЫ

При разработке архитектуры системы учтем, что основными ее модулями являются подсистемы сбора и анализа данных, схематично представленные на рис. 2.

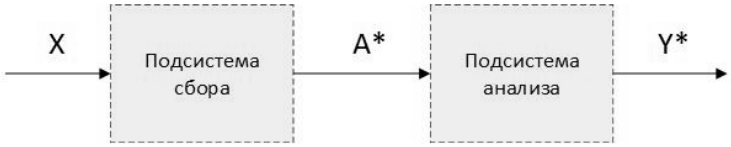

Рис. 2. Общая схема системы

Fig. 2. General schematic of the system

В системе кроме основных подсистем должно быть клиент-серверное приложение, которое осуществляет диалог пользователя и системы. Таким образом, интерфейс предполагает наличие следующих дополнительных модулей.

- Модуль ввода данных на сбор.
- Промежуточный модуль вывода результатов сбора.
- Промежуточный модуль ввода параметров анализа.
- Модуль вывода результатов анализа.

В результате расширенная схема системы принимает вид, показанный на рис. 3.

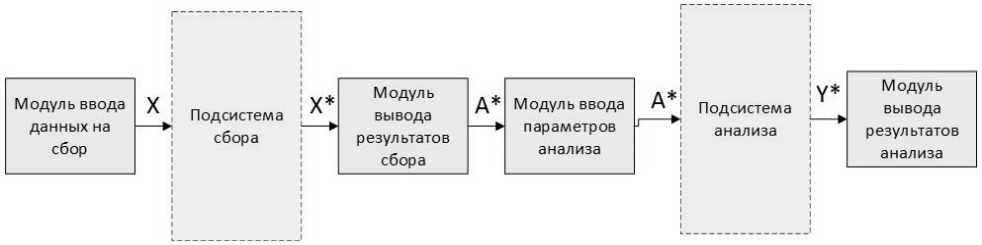

Рис. 3. Расширенная структура системы

Fig. 3. An extended structure of the system

Соответствующая предложенной структуре системы блок-схема алгоритма функционирования представлена на рис. 4.

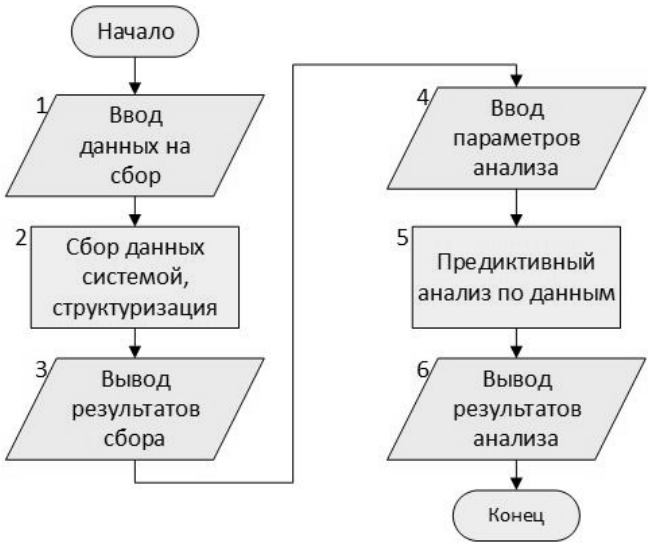

Рис. 4. Блок-схема общего алгоритма функционирования системы

Fig. 4. Block diagram of the general algorithm of the system functioning

Особенности функционирования системы иллюстрирует простой пример. Клиент (пользователь, исследователь) в интерфейсе системы выбирает некоторые конкретные данные, по которым будет производиться сбор данных и анализ. На основе этого система определяет ресурсы, с которых будут получены данные. После структуризации и прогнозирования пользователю предлагается несколько интересных для него вариантов. Например, если пользователь вводит данные по определенной фирме, то в результате будут выданы те группы (сообщества, каналы) в социальных медиа, где по результатам анализа предположительно могут быть новости и другая информация о данной фирме в будущем.

4. ПОДСИСТЕМА СБОРА И СВЯЗАННЫЕ С НЕЙ МОДУЛИ

Наиболее подходящими для сбора данных социальных медиа являются возможности программ, использующих API сайтов и парсинг. Потребуется комбинированное применение данных методов. Так как у многих сайтов API преимущественно свой [6], возможно следующее решение: под основные задачи, характерные для известных социальных медиа (поиск пользователей и их упоминаний / комментариев, тегов – ключевых слов), использовать API. В случае глобальных задач, таких как получение информационных статей, следует использовать парсинг.

Чаще всего архитектуры подсистем сбора описываются примерно одинаковым образом, как в [7]. Пользуясь этим, выделим следующие элементы для нашей подсистемы.

- Модуль планирования, который вносит в систему новые задачи на сбор.
- Очередь задач, определяющая порядок сбора данных.

- Программное обеспечение, сканирующее страницы сети Интернет для дальнейшего сохранения этих данных в базе данных и преобразовывающее информацию в структурированный вид.
- Модуль записи, который записывает структурированные данные на сервер (подсистема хранения).

Сначала пользователь вносит определенную информацию в систему, этим занимается модуль ввода данных на сбор. Входные данные переносятся в модуль планирования задач. Для всех основных ресурсов, по которым будет проводиться сбор, заранее формируются запросы. В данном модуле задачи никак не разбиваются по их сложности и времени выполнения, они создают список в зависимости от данных, введенных пользователем.

В ходе обработки создается порядок выполнения задач, он записывается в модуль очереди. В общем случае задача сводится к получению данных через API или парсинг по одному из интернет-ресурсов. Задачи могут выполняться параллельно, скорость обработки зависит от возможностей оборудования [8]. Порядок выполнения может быть сформирован по охвату социальной сети (количеству активных пользователей на данный момент). Обычно результат заранее не известен из-за постоянного обновления информации в Интернете [9], поэтому данный способ будет являться наилучшим для поставленной задачи.

Сбор информации включает в себя следующие этапы.

- Консолидация (извлечение данных).
- Трансформация (преобразование к структурированному виду).
- Очистка (удаление информации, которая может быть отброшена) [10].

Этапы консолидации и трансформации для сбора данных через API можно объединить. Этот метод чаще всего представляет информацию в структурированном виде [11]. Парсинг позволяет дополнить уже полученную информацию, используя не только социальные сети, но и обычные интернет-ресурсы.

Рассмотрим процесс консолидации. Сначала выбираются ресурсы, заданные в системе, как классические социальные медиа. К ним уже сформированы определенные стандартные запросы на поиск и сбор данных. Эти запросы относятся к API сайта. Такой метод разделения ресурсов позволяет облегчить задачу. Чаще всего на обычных новостных ресурсах более простым является использование стандартных алгоритмов парсинга, что нельзя сказать о социальных сетях, имеющих API.

В случае парсинга могут использоваться стандартные алгоритмы (например, библиотеки Python requests и BS4). Для создания правил парсинга данных в формате html и xml их может быть достаточно, но в случае формирования посредством JavaScript может понадобиться использование Selenium Web Driver, набора драйверов для работы с браузерами [12].

Также следует помнить об ограничениях, которые могут быть заложены создателями / администраторами интернет-ресурса:

– требование авторизации на сайте;

– блокировка большого количества запросов [13].

В первом случае решение может быть простым: предварительная авторизация с использованием POST-запросов. Однако регистрация на каждом из многочисленных ресурсов невозможна.

Во втором случае есть три известных метода решения:

1) установка времени между запросами;

2) имитация работы браузера при каждом запросе – эта возможность есть в вышеупомянутом Selenium Web Driver;

3) использование прокси.

Трансформация будет проводиться следующим образом:

1) выделение ресурсов, с которых получена информация;

2) для каждой социальной сети – получение списка источников: групп (каналов, сообществ) и пользователей, на чьих страницах расположены необходимые данные;

3) для каждого прочего интернет-ресурса – получение сайтов, в статьях / новостях которых расположена информация.

Очистка данных является третьим этапом сбора информации. Для данного этапа необходимо проводить проверку, является пользователь ботом или удален ли он. Это упрощает процесс анализа для подсистемы аналитики. В случае ботов основные социальные сети блокируют их в течение небольшого количества времени, что переносит их в разряд удаленных [14]. Ботов определить просто: это пользователи, зарегистрированные совсем недавно, состоящие в большом количестве сообществ. Учитывая это, данная проверка выглядит следующим образом: если пользователь зарегистрирован в течение предыдущих 24 часов и состоит более чем в 100 публичных страницах (группы, например, в социальной сети «ВКонтакте» можно скрыть), то пользователь вносится в список ботов и полученная информация отбрасывается без сохранения в базе данных системы.

Некоторые данные в социальных сетях пользователи делают скрытыми. Однако общая информация, такая как количество подписчиков (в Twitter, Instagram, YouTube и других) позволяет проводить очистку данных на их основе. В социальной сети «ВКонтакте» стоит ориентироваться именно на публичные страницы, хотя не все боты скрывают и группы.

Итоговый результат по каждой из задач необходимо сохранить, этим занимается модуль записи. При этом используется система управления базами данных. В дальнейшем подсистема предиктивного анализа обращается именно уже к записям базы данных.

Каждый запрос пользователя представляет собой отдельную обработку и анализ данных, приводящие к различным результатам. Чтобы не допустить излишней загрузки баз данных системы, предлагается очистка устаревших запросов (более месяца) с результатами. Пользователь может выгрузить уже имеющиеся результаты в одном из доступных форматов либо сохранить запрос. Следует учесть, что с течением времени данные становятся устаревшими. В любом случае это ведет к редактированию входных данных и полученных результатов для последующей аналитики [15].

По итогам работы подсистема сбора передает полученные результаты в модуль вывода результатов сбора. Он представляет результаты в удобном для пользователя виде. Далее данные можно сгруппировать для анализа по отдельным ресурсам, либо же аналитика проводится по всей полученной информации. Также можно указать желаемое количество результатов.

Рис. 5. Подсистема сбора

Fig. 5. Collection subsystem

Все действия в подсистеме сбора выполняются последовательно. В связи с этим можно предложить общий вид подсистемы сбора, показанный на рис. 5.

ЗАКЛЮЧЕНИЕ

Предъявляемые к системе требования распределены по подзадачам, совокупное решение которых позволит спроектировать эффективную систему сбора и предиктивного анализа данных социальных медиа. Предложено формальное описание входных, промежуточных и выходных данных системы, описаны особенности их представления и получения от пользователя. Представлена общая архитектура системы и алгоритм ее функционирования.

Подробно описаны элементы и процессы подсистемы сбора данных, рассматривается ее взаимодействие с пользователем. Предложены методы трансформации и очистки данных при сборе, а также возможные варианты обхода ограничений при парсинге данных и оптимизации хранения информации.

В дальнейшем планируется рассмотреть проектирование подсистемы анализа данных социальных медиа, которая является второй основной составляющей общей системы. Это позволит представить архитектуру системы и алгоритм функционирования в более детальном виде, что даст возможность окончательной разработки системы.

СПИСОК ЛИТЕРАТУРЫ

1. *Калытюк И.С.* Разработка и исследование алгоритма извлечения данных геолокации в социальных сетях // Научное сообщество студентов XXI столетия. Технические науки. – 2018. – № 11 (70). – С. 39–44.

2. *Калытюк И.С., Французова Г.А., Гунько А.В.* К вопросу выбора методов предиктивного анализа данных социальных медиа // Автоматика и программная инженерия. – 2019. – № 4 (30). – С. 9–17.

3. *Суханов А.А., Маратканов А.С.* Анализ способов сбора социальных данных из сети Интернет // International Scientific Review. – 2017. – № 1 (32). – С. 22–25.

4. Social media analytics – challenges in topic discovery, data collection, and data preparation / S. Stieglitz, M. Mirbabaie, B. Ross, C. Neuberger // International Journal of Information Management. – 2018. – Vol. 39. – P. 156–168.

5. *Низомутдинов Б.А., Тропников А.С., Углова А.Б.* Автоматизированный сбор данных социальных сетей для разработки факторной модели сетевой самопрезентации // International Journal of Open Information Technologies. – 2020. – Т. 8, № 1. – С. 64–71.

6. *Russell M.A.* Mining the social Web. Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more. – O'Reilly Media, 2013. – 448 p.

7. *Чесноков В.О.* Программное обеспечение сбора и анализа графов ближайшего окружения из онлайновых социальных сетей // Машиностроение и компьютерные технологии. – 2018. – № 8. – С. 34–44.

8. *Мельник Э.В., Клименко А.Б.* Применение концепции «туманных» вычислений при проектировании высоконадежных информационно-управляющих систем // Известия Тульского государственного университета. Технические науки. – 2020. – № 2. – С. 273–283.

9. *Райнова О.Д.* Решение задачи достижения наилучшего гарантированного результата поиска // Открытое образование. – 2006. – № 1. – С. 40–49.

10. *Гранаткин Д.С., Галиаскаров Э.Г.* Автоматизация сбора информации из открытых интернет-источников // Объектные системы. – 2016. – № 13. – С. 71–77.

11. *Турков Е.С., Степанов Ю.А.* Концептуальная модель модуля сбора данных о вакансиях для экспертной системы // Международный научно-исследовательский журнал. – 2020. – № 2-1 (92). – С. 75–78.

12. *Gojare S., Joshi R., Gaigaware D.* Analysis and design of selenium web driver automation testing framework // Procedia Computer Science. – 2015. – N 50. – P. 341–346.

13. *Thomas D.M., Mathur S.* Data analysis by web scraping using Python // 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA). – Coimbatore, India, 2019. – P. 450–454.

14. Система идентификации информационных угроз на основе открытых данных сети Интернет / Д.О. Маркин, С.М. Макеев, Н.В. Изотов, А.Ю. Андросов // Известия Тульского государственного университета. Технические науки. – 2020. – № 9. – С. 86–94.

15. Big Data: The management revolution / A. McAfee, E. Brynjolfsson, T. Davenport, D. Patil, D. Barton // Harvard Business Review. – 2012. – Vol. 90. – P. 66–67.

Калытюк Иван Сергеевич, аспирант кафедры автоматики Новосибирского государственного технического университета. Основное направление научных исследований – разработка и исследование систем сбора и анализа данных. E-mail: ivankalytyuk@yandex.ru

Французова Галина Александровна, доктор технических наук, профессор кафедры автоматики Новосибирского государственного технического университета. Основное направление научных исследований – синтез систем экстремального регулирования. E-mail: frants@ac.cs.nstu.ru

Гунько Андрей Васильевич, кандидат технических наук, доцент кафедры автоматики Новосибирского государственного технического университета. Основное направление научных исследований – разработка автоматизированных систем сбора и обработки результатов. E-mail: gun@ait.cs.nstu.ru

Kaytyuk Ivan S., postgraduate student at the Department of Automation in Novosibirsk State Technical University. The main field of his scientific research is development and research of data collection and analysis systems. E-mail: ivankalytyuk@yandex.ru

Frantsuzova Galina A., Doctor of Technical Sciences, Professor, Department of Automation, NSTU. The main field of her scientific research is synthesis of systems of extreme regulation. E-mail: frants@ac.cs.nstu.ru

Gun'ko Andrei V., Candidate of Technical Sciences, associate professor at the Department of Automation NSTU. The main field of his scientific research is development of automated systems for collecting and processing results. E-mail: gun@ait.cs.nstu.ru

DOI: 10.17212/2782-2001-2021-1-73-84

Initial stages of designing a system for collecting and predictive analysis of social media data[*]

I.S. KALYTYUK[a], *G.A. FRANTSUZOVA*[b], *A.V. GUN'KO*[c]

Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation

[a] *ivankalytyuk@yandex.ru* [b] *frants@ac.cs.nstu.ru* [c] *gun@ait.cs.nstu.ru*

Abstract

This article discusses the design of a system for collecting and predictive analysis of social media. With the development of the Internet, as well as social media, it has become easier to access and distribute information because network users themselves are both creators and recipients of diverse information. To gain new knowledge that can be useful to users of social media, it is possible to use predictive analytics – a set of statistical analysis methods that extract new information from current and historical data. This method of analyzing social media data is at the stage of its development.

Predictive analytics is based on automatic search for connections, anomalies and patterns between various factors. To form a predictive model, a large set of statistical modeling methods, data mining, machine learning, neural networks and other mechanisms are used. Together with various methods of collecting information from Internet resources, such as parsing and social network APIs, predictive analytics can offer the most interesting sources of information for the user. In order to combine the methods of predictive analysis and data collection methods, it is necessary to take a detailed approach to the system design process.

The paper proposes a formal description of the data that a future system uses. In addition, the general architecture and algorithm of functioning are highlighted. Special attention is paid to a detailed description of one of the main parts of the system (the collection subsystem). The obtained results will be used in further design, and it is planned to further study the analytics subsystem. Subsequent work on this topic will make it possible to detail the architecture and algorithm of functioning.

Keywords: social media, predictive analysis, system design, architecture, algorithm of functioning, data collection, API, parsing

REFERENCES

1. Kalytyuk I.S. Razrabotka i issledovanie algoritma izvlecheniya dannykh geolokatsii v sotsial'nykh setyakh [Development and research of an algorithm for extracting geolocation data in social networks]. *Nauchnoe soobshchestvo studentov XXI stoletiya. Tekhnicheskie nauki = Scientific community of students of the XXI century. Technical Science*, 2018, no. 11 (70), pp. 39–44.

2. Kalytyuk I.S., Frantsuzova G.A., Gunko A.V. K voprosu vybora metodov prediktivnogo analiza dannykh sotsial'nykh media [On the choice of methods for predictive analysis of social media data]. *Avtomatika i programmnaya inzheneriya = Automatics and Software Enginery*, 2019, no. 4 (30), pp. 9–17.

3. Sukhanov A.A., Maratkanov A.S. Analiz sposobov sbora sotsial'nykh dannykh iz seti Internet [Analysis of ways to collect social data from the Internet]. *International Scientific Review*, 2017, no. 1 (32), pp. 22–25. (In Russian).

4. Stieglitz S., Mirbabaie M., Ross B., Neuberger C. Social media analytics - challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 2018, vol. 39, pp. 156–168.

[*] *Received 24 August 2020.*

5. Nizomutdinov B.A., Tropnikov A.S., Uglova A.B. Avtomatizirovannyi sbor dannykh sotsial'nykh setei dlya razrabotki faktornoi modeli setevoi samoprezentatsii [Automated data collection of social networks for the development of a factor model of network self-presentation]. *International Journal of Open Information Technologies*, 2020, vol. 8, no. 1, pp. 64–71. (In Russian).

6. Russell M.A. *Mining the social Web. Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more.* O'Reilly Media, 2013. 448 p.

7. Chesnokov V.O. Programmnoe obespechenie sbora i analiza grafov blizhaishego okruzheniya iz onlainovykh sotsial'nykh setei [Software for collecting and analyzing immediate environment graphs from online social networks]. *Mashinostroenie i komp'yuternye tekhnologii = Mechanical engineering and computer technology*, 2018, no. 8, pp. 34–44. (In Russian).

8. Mel'nik E.V., Klimenko A.B. Primenenie kontseptsii "tumannykh" vychislenii pri proektirovanii vysokonadezhnykh informatsionno-upravlyayushchikh sistem [Application of the concept of "foggy" computing in the design of highly reliable information and control systems]. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = News of the Tula state university. Technical sciences*, 2020, no. 2, pp. 273–283.

9. Rainova O.D. Reshenie zadachi dostizheniya nailuchshego garantirovannogo rezul'tata poiska [Solving the problem of achieving the best guaranteed search result]. *Otkrytoe obrazovanie = Open Education*, 2006, no. 1, pp. 40–49.

10. Granatkin D.S., Galiaskarov E.G. Avtomatizatsiya sbora informatsii iz otkrytykh internet-istochnikov [Automation of information collection from open Internet sources]. *Ob"ektnye sistemy = Object Systems*, 2016, no. 13, pp. 71–77.

11. Turkov E.S., Stepanov Yu.A. Kontseptual'naya model' modulya sbora dannykh o vakansiyakh dlya ekspertnoi sistemy [Conceptual model of the module for collecting data on vacancies for the expert system]. *Mezhdunarodnyi nauchno-issledovatel'skii zhurnal = International Research Journal*, 2020, no. 2-1 (92), pp. 75–78. (In Russian).

12. Gojare S., Joshi R., Gaigaware D. Analysis and design of selenium web driver automation testing framework. *Procedia Computer Science*, 2015, no. 50, pp. 341–346.

13. Thomas D.M., Mathur S. Data analysis by web scraping using Python. *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2019, pp. 450–454.

14. Markin D.O., Makeev S.M., Izotov N.V., Androsov A.Yu. Sistema identifikatsii informatsionnykh ugroz na osnove otkrytykh dannykh seti internet [Information threat identification system based on open Internet data]. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = News of the Tula state university. Technical sciences*, 2020, no. 9, pp. 86–94.

15. McAfee A., Brynjolfsson E., Davenport T., Patil D., Barton D. Big Data: The management revolution. *Harvard Business Review*, 2012, vol. 90, pp. 66–67.

Для цитирования:

Калытюк И.С., Французова Г.А., Гунько А.В. Начальные этапы проектирования системы сбора и предиктивного анализа данных социальных медиа // Системы анализа и обработки данных. – 2021. – № 1 (81). – С. 73–84. – DOI: 10.17212/2782-2001-2021-1-73-84.

For citation:

Kalytyuk I.S., Frantsuzova G.A., Gunko A.V. Nachal'nye etapy proektirovaniya sistemy sbora i prediktivnogo analiza dannykh sotsial'nykh media [Initial stages of designing a system for collecting and predictive analysis of social media data]. *Sistemy analiza i obrabotki dannykh = Analysis and data processing systems*, 2021, no. 1 (81), pp. 73–84. DOI: 10.17212/2782-2001-2021-1-73-84.