ISSN 2782-2001
Системы анализа
и обработки данных
том 85, № 1, 2022, с. 59–72

http://journals.nstu.ru/vestnik
Analysis and data
processing systems
Vol. 85, No. 1, 2022, pp. 59–72

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND MANAGEMENT

УДК 004.415.2

DOI: 10.17212/2782-2001-2022-1-59-72

Заключительные 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 и социальных медиа более простым стал доступ к информации и ее распространение, ведь сами пользователи сети являются одновременно создателями и получателями различной информации. 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 медиа – влияние на восприятие, отношение и конечное поведение потребителей.

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 подробно отнестись к процессу проектирования системы.

В работе особое внимание обращено на подробное описание второй из основных частей системы (подсистемы анализа). Помимо этого, выделены полная архитектура и алгоритм функционирования. Полученные результаты будут в полном объеме использоваться при дальнейшей разработке. Работа над данной темой позволит облегчить процесс последующего тестирования и исследования системы.

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

* *Статья получена 22 июля 2021 г.*

ВВЕДЕНИЕ

Известно, что Интернет на сегодняшний момент – крупнейшая база данных во всем мире. Социальные сети являются одним из основных источников информации в Интернете [1]. Сами пользователи являются создателями контента, который доступен и прочим людям. Так как пользователей соцсетей достаточно много, образуется огромный массив различной информации – это открывает возможности для исследований в разных областях науки.

В связи с этим важное значение приобретает анализ размещенной в социальных сетях информации и возможность прогнозирования будущих событий. К сожалению, методы предиктивной аналитики на сегодня развиты не так сильно [2]. На основании этого очевидна актуальность применения данных методов в области анализа данных социальных медиа.

Целью настоящей работы является проектирование подсистемы анализа системы сбора и предиктивного анализа данных социальных медиа, а также разработка подробного алгоритма ее функционирования. Ранее в работе [3] рассматривалось подробное проектирование подсистемы сбора.

Для того чтобы описать алгоритм функционирования системы максимально подробно, требуется выделить основные модули подсистемы анализа. Основной проблемой при проектировании данной подсистемы можно назвать связи между модулями. Следует достаточно четко понимать, как части подсистемы взаимодействуют между собой, какие вычисления и методы производятся в ней.

Структура работы следующая: в первом и втором разделах представлены постановка задачи и структура подсистемы анализа. Подзадачи, которые решает подсистема анализа, описаны в разделах 3, 4 и 5. Алгоритм функционирования и архитектура подсистемы представлены в разделе 6. Полный алгоритм функционирования и архитектура всей системы описаны в разделе 7.

1. ПОСТАНОВКА ЗАДАЧИ

Необходимо разработать эффективную подсистему анализа. Целью функционирования подсистемы (как заключительной части целой системы) является представление пользователю интернет-источников, которые могут быть ему интересны в будущем по настоящим данным.

В подсистеме анализа при помощи алгоритмов кластеризации выделяется определенная часть ресурсов. Так как процесс кластеризации является основой подсистемы, к ней предъявляются следующие требования:

– определение повторяющихся ресурсов, переданных из подсистемы сбора;
– вычисление количества кластеров, которое должно получиться;
– выделение характеристик, по которым проводится кластеризация;
– структуризация характеристик кластеризации;
– нормализация характеристик кластеризации;
– вычисление центров кластеров;
– ограничение числа результатов.

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

1) определение числа кластеров;

2) определение характеристик кластеризации;

3) определение центров кластеров и желаемых результатов.

В свою очередь, каждой из подзадач соответствуют свои требования из общего перечня. Для определения числа кластеров необходимо определить повторяющиеся ресурсы и только после этого вычислить их количество; для решения второй подзадачи, помимо выделения характеристик для различных типов интернет-ресурсов, требуется структуризировать их и провести нормализацию. Решение третьей подзадачи предполагает определение элементов, которые будут являться центрами кластеров, и ограничение числа результатов.

2. СТРУКТУРА ПОДСИСТЕМЫ АНАЛИЗА

Схематично подсистема анализа с выделенными подзадачами представлена на рис. 1.

Рис. 1. Схематичное представление постановки задачи

Fig. 1. A schematic representation of the problem statement

До того как перейти непосредственно к подсистеме анализа, рассмотрим работу модуля ввода параметров. Как было сказано в [3], после сбора пользователь принимает решение, каким образом будет проводиться предиктивный анализ. Эти данные вводятся в промежуточном модуле (ввод параметров анализа). Он выполняет следующие функции:

– выбор количества желаемых результатов;

– выбор ресурсов, по которым производится анализ.

Для первого параметра количество задается в числовом виде либо не ограничивается. Если ограничений нет, результатом будут являться все собранные данные в зависимости от того, какие ресурсы будут выбраны.

Для второго параметра пользователь системы может как проводить анализ сразу по всем данным, так и запустить частную задачу. В этом случае возможен анализ, например, только по известным социальным сетям, по прочим интернет-ресурсам либо же только по определенным сайтам. Данные варианты должны быть доступны в интерфейсе системы. По результатам работы подсистемы сбора формируется общий список ресурсов, который можно разделить

на две части. Первая – социальные сети (данные, полученные через их API) и вторая – прочие ресурсы (данные, полученные при помощи парсинга).

Выделим следующие части подсистемы анализа:

– модуль вычислений;

– модуль сбора характеристик кластеризации;

– модуль кластеризации.

Рассмотрим решение каждой из заявленных выше подзадач отдельно.

3. ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ

При выборе параметров анализа пользователь системы указывает количество желаемых результатов. После сбора данных известно количество ресурсов, по которым проводится анализ. Учитывая это, количество кластеров будет равно количеству всех ресурсов, деленному на желаемое [4].

Однако сначала модулю вычислений необходимо провести нормализацию данных. Анализируется, нет ли повторяющихся интернет-ресурсов или сообществ в социальных сетях. Если в одном сообществе есть два упоминания, то оно должно рассматриваться один раз, но с наибольшим приоритетом (реализация будет описана ниже, после определения характеристики для кластеризации).

Учитывая повторяющиеся ресурсы, получаем следующую формулу определения числа кластеров:

$$k = \frac{A}{B}, \tag{1}$$

где A – количество результатов, полученное при сборе данных (с вычтенным числом повторений); B – количество желаемых результатов, заданное пользователем; k – число кластеров.

Количество кластеров, получаемое после использования формулы (1), не всегда будет равным целому числу. При дробной части, равной или большей, чем 0.5, задавать количество кластеров равным целой части $k + 1$ (округление в большую сторону). В противоположном случае количество кластеров принимается равным целой части k.

Используя это решение, можно столкнуться с некоторыми трудностями. Например, если будет 90 полученных при сборе результатов и задано 20 желаемых, $k = 4.5$, то это приведет систему к нахождению пяти кластеров. Можно предположить, что в каждом кластере будет примерно по 18 элементов, что не соответствует параметрам, которые задает пользователь. Это означает, что после первоначальной кластеризации системе необходимо определить, сколько ещё элементов потребуется выделить для представления результатов пользователю. При получении этого результата (в данном случае 2) необходимо разделить на него количество элементов в ближайшем кластере и провести еще одну кластеризацию. Данные используются непосредственно из этого кластера. В сложных ситуациях итерация может повториться еще несколько раз.

Аналогично проводится повторная кластеризация в случае количества данных, которое больше, чем было задано желаемых результатов. Однако операция проводится по получившемуся кластеру. Количество кластеров задается вычитанием количества желаемых результатов из числа уже имеющихся

элементов. Далее количество элементов кластера делится на получившееся число. Такая итерация может повторяться до получения необходимого количества желаемых результатов.

4. ОПРЕДЕЛЕНИЕ ХАРАКТЕРИСТИК КЛАСТЕРИЗАЦИИ

Когда количество необходимых кластеров определено, следует решить подзадачу характеристик кластеризации. Предполагается, что пользователю будут интересны ресурсы, где наибольшее количество раз встречаются ключевые слова, заданные при начальном поиске. В случае, если ресурсов с одинаковым количеством упоминаний несколько, наибольшее предпочтение отдается тем, которые имеют больший охват. Например, для социальных сетей подсчитывается количество подписчиков пользователя или сообщества. Для интернет-ресурсов, по которым проводится парсинг, определяется число посещений.

В случае социальных сетей для каждого подресурса (сообщество, канал или пользователь) модуль сбора характеристик анализа через API дополнительно получает информацию о количестве подписчиков. В случае обычных ресурсов возможно воспользоваться сторонним программным обеспечением – Яндекс.Метрика [5] или Google Analytics [6]. При помощи этих продуктов возможно определение прямых заходов на сайт и переходов из социальных сетей [7].

Помимо вышеупомянутых программных комплексов, для данной задачи можно использовать и платные сервисы – SimilarWeb, Serpstat, Alexa, Semrush [8] и Ahrefs [9]. Также есть и свободное программное обеспечение, например, ресурс 2*ip* [10]. Возможное применение различных плагинов для браузеров по типу RDS Bar [11]. Собственно, через RDS Bar можно получить рейтинг (без количества пользователей, посетивших сайт) в глобальной сети от Alexa или SemRush. Для задачи анализа только по сторонним интернет-ресурсам этот метод является достаточным.

Однако для общего анализа требуется конкретное количество посетителей, что возвращает нас к использованию других продуктов. Хорошим решением можно назвать онлайн-портал PR-CY [12]. Ресурс бесплатно предоставляет примерное количество посещений задаваемого сайта за день, месяц и год. Данные являются примерными, так как являются средним значением, полученным через вышеупомянутые ресурсы.

Для общего анализа некорректно сравнивать данные количества подписчиков социальных сетей и количества людей, которые посещают определенный сайт. К сожалению, статистика сообществ и каналов в социальных сетях доступна только их владельцам, что не позволяет в полной мере сравнить данные цифры. Возможно лишь предположение о том, что среднее количество посетителей за день примерно равно количеству подписчиков в социальной сети. В целом общая задача для анализа по всем результатам сбора является самой сложной задачей для системы.

Когда количественная характеристика для каждого результата сбора получена и сохранена, следует пересчитать ее для тех ресурсов, в которых должен быть назначен наибольший приоритет. Это просто реализовать, возвращаясь в модуль вычислений с уже известными данными. В случае, если

было два упоминания, количественная характеристика умножается на 2. Аналогично проводится операция и для другого количества упоминаний.

После этого производится нормализация данных. Существует несколько известных методов проведения нормализации кластеризуемых данных.

1. Min-max нормализация. В этом случае какой-либо отрезок (к примеру, от нуля до единицы) заполняется от одной границы к другой. Максимальное и минимальное количественные значения приравниваются к соответствующим значениям отрезка, после чего все остальные числа пересчитываются в его рамках. В итоге все данные имеют определенное значение из данного промежутка.

2. Нормализация стандартным отклонением. Используется приведение уже известного распределения к центрированному (с единичным отклонением).

3. Десятичное масштабирование. Этот метод одинаково смещает точку в десятичных числах. Все данные находятся в окрестности нуля [13].

По итогам рассмотрения этих методов становится понятно, что наиболее подходящими являются первый и третий. Второй метод применяется, если границы не известны. Предлагается использовать в разрабатываемой системе оба метода (min-max нормализацию и десятичное масштабирование).

Характеристика для кластеризации будет использоваться при расчете меры расстояния. Помимо этого, необходимо выбрать центры кластеров.

5. ОПРЕДЕЛЕНИЕ ЦЕНТРОВ КЛАСТЕРОВ И ЖЕЛАЕМЫХ РЕЗУЛЬТАТОВ

Из полученных данных для анализа в качестве центров будут выбраны наиболее удаленные значения. Это наибольшее и наименьшее количественные значения, рассчитанные в результате нормализации [14]. Если количество кластеров $k > 2$, определяем количество оставшихся центров кластеров:

$$m = k - 2, \tag{2}$$

где k – число кластеров; m – количество оставшихся центров кластеров.

Первый центр рассчитывается по следующей формуле:

$$n_1 = \frac{A}{m+1}, \tag{3}$$

где m – количество оставшихся центров кластеров; A – количество результатов, полученное при сборе данных (с вычтенным числом повторений); n_i – первый центр кластера.

Целая часть полученного числа становится новым центром (так же и в последующих формулах).

Если $k > 3$ и дробная часть $n_i > 0.5$, следующий центр вычисляется по формуле

$$n_i = n_1 \cdot i + 1, \tag{4}$$

где n_i – первый центр кластера, $i \in [1, m]$, m – количество оставшихся центров кластеров.

Возможно использовать чередование, в котором на каждой нечетной итерации добавляется единица, а на четной – не добавляется.

В случае, если дробная часть $n_i < 0.5$, следующий центр вычисляется по формуле

$$n_i = n_1 \cdot i, \tag{5}$$

где n_i – первый центр кластера, $i \in [1, m]$, m – количество оставшихся центров кластеров.

После нахождения всех центров начинается процесс кластеризации. Используется метод *k*-средних. При его использовании пространство данных разбивается на определяемое заранее количество *k* кластеров. Далее в исходном наборе выбирается количество объектов, равное *k*. Проводится вычисление для каждого из объектов – выбирается наиболее близкий центр. После образования новых кластеров центры пересчитываются и процесс продолжается до момента остановки [15]. Более подробно метод рассматривался в [2].

В результате получаем количество кластеров, соответствующее первоначальным расчетам. Пользователю интересен кластер, который включает в себя наибольшее количественное значение. В случае полного совпадения числа элементов кластера и количества результатов, заданного пользователем, этот кластер является результатом анализа. Если же он больше или меньше, проводится повторная кластеризация по той же процедуре.

6. ПРЕДСТАВЛЕНИЕ ИТОГОВЫХ РЕЗУЛЬТАТОВ. АЛГОРИТМ ФУНКЦИОНИРОВАНИЯ И АРХИТЕКТУРА ПОДСИСТЕМЫ

Можно предложить следующий общий вид подсистемы анализа (рис. 2).

Рис. 2. Подсистема анализа

Fig. 2. An analysis subsystem

Модуль вывода результатов анализа представляет итоговые данные пользователю. Показываются ссылки на социальные сети и их подразделы либо же интернет-ресурсы, на которых в будущем возможно появление интересующей его информации. Эти данные пользователь может сохранить в удобном для себя формате. Решено не хранить эту информацию в отдельном модуле, так как каждый процесс анализа с течением времени может выдавать различные результаты. Это сделано по причине того, что охват социальных сетей и интернет-ресурсов постоянно меняется.

Алгоритм работы подсистемы следующий.

1. Полученные при сборе ресурсы анализируются на уникальность модулем вычислений. Уникальные ресурсы получают приоритет $p = 1$. В противном случае выставляется $p > 1$ (в зависимости от количества). Количество кластеров рассчитывается с учетом введенных параметров анализа.

2. Происходит переход в модуль сбора характеристик кластеризации. Характеристики ресурсов для кластеризации собираются с использованием программных комплексов для определения охвата и запросов к API социальных сетей.

3. Производится возврат в модуль вычислений. Количественные характеристики пересчитываются для тех результатов, у которых $p > 1$. Осуществляется нормализация данных.

4. Происходит переход в модуль кластеризации. Кластеризация производится по уникальным ресурсам, будь то сообщество в социальной сети или обычный сайт. Для кластеризации необходимы постоянные обращения к модулю сбора. Итоговое количество результатов сравнивается с запрошенным количеством желаемых. Если данные числа не равны, операция повторяется для текущего или ближайшего кластера, чтобы отбросить или дополнить результаты.

7. АЛГОРИТМ ФУНКЦИОНИРОВАНИЯ И АРХИТЕКТУРА СИСТЕМЫ

Опираясь на ранее спроектированную подсистему сбора [3] и подсистему анализа, опишем полную архитектуру системы (рис. 3).

На рис. 3 X – входные данные для системы; X* – промежуточные данные, полученные в результате сбора; A* – ресурсы, на которых располагаются собранные данные; Y* – выходные данные системы.

Разработан алгоритм функционирования системы. Его блок-схема приведена на рис. 4.

Алгоритм работы системы с комментариями по каждому из блоков 1–19 следующий.

1. Ввод данных на сбор. Ключевые слова вводятся через модуль ввода данных на сбор.

2. Внесение данных в модуль планирования. Введенные ключевые слова вносятся в модуль планирования.

3. Сортировка запросов в модуле задач. Запросы делятся на две группы. Первая группа – запросы к классическим социальным сетям (использование API). Вторая группа – запросы по глобальной сети (использование парсинга). Запросы из первой группы вносятся в модуль очереди задач в первую очередь, далее вносятся запросы из второй группы.

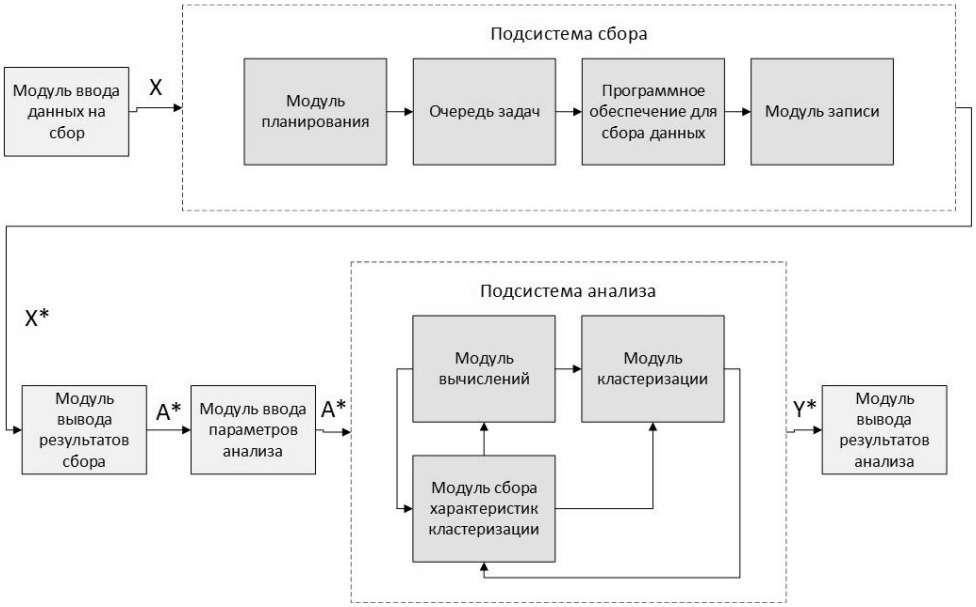

Рис. 3. Архитектура системы

Fig. 3. The system architecture

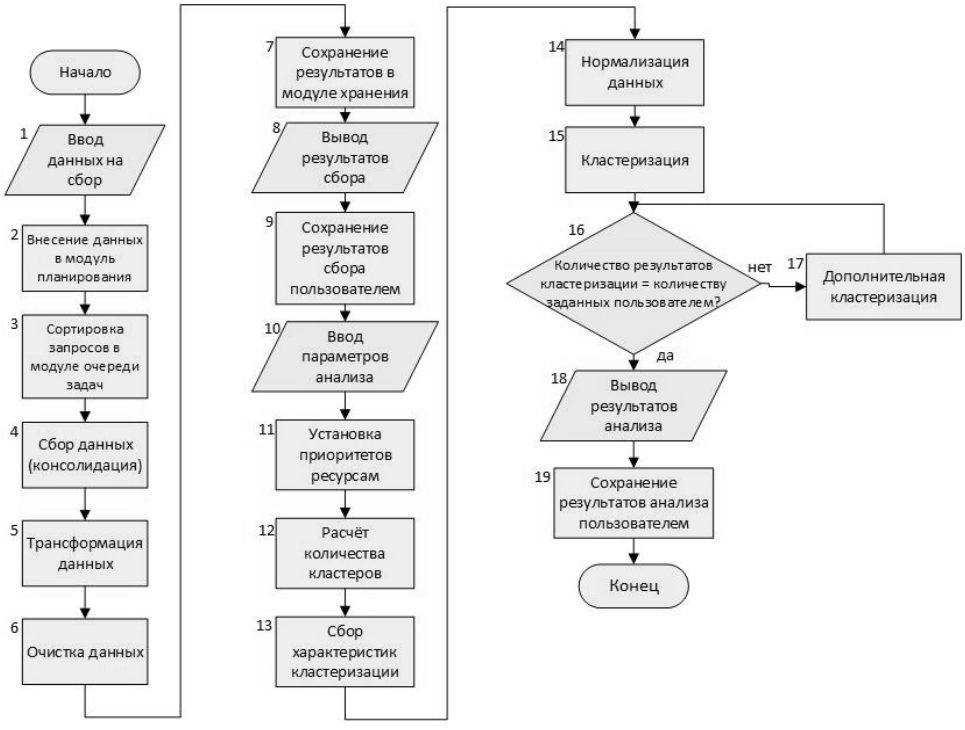

Рис. 4. Блок-схема алгоритма функционирования системы

Fig. 4. Block diagram of the system functioning algorithm

4. Сбор данных (консолидация). Данные собираются с использованием API и парсинга.

5. Трансформация данных. Данные трансформируются к общему списку. Система формирует список групп, каналов, сообществ и пользователей, на чьих страницах в социальных сетях расположена необходимая информация. Также вносятся в список и прочие интернет-ресурсы.

6. Очистка данных. Данные очищаются от информации, которая может быть отброшена (нахождение ботов, удаленных страниц).

7. Сохранение результатов в модуле хранения. Данные сохраняются в системе управления базами данных.

8. Вывод результатов сбора. Результаты представляются через модуль вывода результатов сбора.

9. Сохранение результатов сбора. Полученные результаты сбора сохраняются и выгружаются на локальный компьютер (по желанию пользователя).

10. Ввод параметров анализа. Желаемое количество результатов и желаемые ресурсы вводятся через модуль ввода параметров анализа.

11. Установка приоритетов ресурсам. Полученные при сборе ресурсы анализируются на уникальность. Уникальные ресурсы получают приоритет $p = 1$. В противном случае выставляется $p > 1$ (в зависимости от количества).

12. Расчет количества кластеров. Количество кластеров рассчитывается с учетом введенных параметров анализа.

13. Сбор характеристик кластеризации. Характеристики ресурсов для кластеризации собираются с использованием программных комплексов с целью определения охвата и запросов к API социальных сетей.

14. Нормализация данных. Количественные характеристики пересчитываются для тех результатов, у которых $p > 1$. Производится нормализация данных.

15. Кластеризация. Кластеризация производится по уникальным ресурсам, будь то сообщество в социальной сети или обычный сайт.

16. Сравнение количества результатов кластеризации и числа результатов, заданных пользователем. Итоговое количество результатов кластеризации сравнивается с запрошенным числом желаемых.

17. Дополнительная кластеризация. Если числа из шага 16 не равны, кластеризация повторяется для текущего или ближайшего кластера, чтобы отбросить или дополнить результаты (шаги 16 и 17 могут повториться несколько раз).

18. Вывод результатов анализа. Ресурсы, на которых в будущем возможно появление интересующей информации, представляются через модуль вывода результатов анализа.

19. Сохранение результатов анализа пользователем. Полученные результаты сбора сохраняются и выгружаются на локальный компьютер (по желанию пользователя).

ЗАКЛЮЧЕНИЕ

Предлагается оригинальное решение для архитектуры подсистемы анализа, отличающееся нелинейным построением. Модуль вычисления и модуль кластеризации постоянно обращаются к модулю сбора характеристик кластеризации. Такое решение позволяет четко распределить задачи по подсистеме.

При реализации системы возможно использование алгоритмов параллельной обработки данных для большего быстродействия ее работы.

Подробно описаны элементы и процессы подсистемы, рассматривается ее взаимодействие с пользователем. Предложены методы для подзадач нахождения числа кластеров, определения их центров, сбора характеристик кластеризации, а также возможные варианты нормализации данных.

По итогам проектирования представлены подробный алгоритм функционирования всей системы и полная архитектура системы. Предложенный алгоритм позволяет решить такие проблемы, как нахождение выходных результатов по параметрам от пользователя, выделение характеристик для кластеризации, нормализация данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Калытюк И.С.* Разработка и исследование алгоритма извлечения данных геолокации в социальных сетях // Научное сообщество студентов XXI столетия. Технические науки. – СибАК, 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. *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.

4. *Фролов В.В., Слипченко С.Е., Приходько О.Ю.* Метод расчета числа кластеров для алгоритма k-means // Экономика. Информатика. – 2020. – № 47 (1). – С. 213–225. – DOI: 10.18413/2687-0932-2020-47-1-213-225.

5. Яндекс.Метрика: web-сайт. – URL: https://metrika.yandex.ru/list? (дата обращения: 11.02.2022).

6. Google Аналитика: web-сайт. – URL: https://analytics.google.com/analytics/web/provision/#/provision (дата обращения: 11.02.2022).

7. *Молодецкая С.Ф., Шитова Т.Ф.* Оценка эффективности сайта на основе технологии нечеткого управления // Вопросы управления. – 2020. – № 2 (63). – С. 39–49.

8. *Кошик А.* Веб-аналитика 2.0 на практике: тонкости и лучшие методики. – М.: Диалектика, 2019. – 528 с.

9. Ahrefs – это инструменты и ресурсы SEO для роста вашего поискового трафика. – URL: https://ahrefs.com/ru/ (дата обращения: 11.02.2022).

10. 2ip: web-сайт. – URL: https://2ip.ru/analizator/ (дата обращения: 11.02.2022).

11. RDS Bar – расширение для SEO анализа сайта и страниц: web-сайт. – URL: https://www.recipdonor.com/bar (дата обращения: 11.02.2022).

12. PR-CY. Сервис самостоятельного продвижения сайта – онлайн-инструменты для вебмастеров, оптимизаторов и копирайтеров: web-сайт. – URL: https://pr-cy.ru/ (дата обращения: 11.02.2022).

13. *Бабичев С.* Оптимизация процесса предобработки информации в системах кластеризации высокоразмерных данных // Радиоэлектроника, информатика, управление. – 2014. – № 2. – С. 135–142. – DOI: 10.15588/1607-3274-2014-2-19.

14. Статистические алгоритмы кластеризации данных в адаптивных обучающих системах / С.А. Батуркин, Е.Ю. Батуркина, В.А. Зименко, И.В. Сигинов // Вестник РГРТУ. 2010. – № 1 (31). – С. 82–85.

15. *Hartigan J.A., Wong M.A.* Algorithm AS136: a K-means clustering algorithm // Applied Statistics. – 1979. – Vol. 28 (1). – P. 100–108.

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

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

Frantsuzova Galina A., D.Sc. (Eng.), professor, Department of Automation, NSTU. The main area of her scientific research is the synthesis of extreme regulation systems. E-mail: frants@ac.cs.nstu.ru

Gunko Andrei V., PhD (Eng.), associate professor at the Department of Automation, NSTU. The main area of his scientific research is development of automated systems for collecting and processing results. E-mail: gun@ait.cs.nstu.ru

DOI: 10.17212/2782-2001-2022-1-59-72

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

I.S. KALYTYUK[1], *G.A. FRANTSUZOVA*[2], *A.V. GUNKO*[3]

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 data. With the development of the Internet, as well as social media, it has become easier to access and distribute information because the network users themselves are both creators and recipients of varying information. The main type of social media is social networks. Facebook, VK, Instagram, YouTube, Twitter, Odnoklassniki, WhatsApp and Telegram messengers are among the most well-known ones. The most important functions of social media are to influence the perception, attitude and final behavior of consumers.

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.

In this paper, special attention is paid to the detailed description of the second of the main parts of the system (the analysis subsystem). In addition, the full architecture and

[*] *Received 22 July 2021.*

algorithm of operation are highlighted. The results obtained will be used in further development, and it is planned to use them in full. Working on this topic will facilitate the process of subsequent testing and research of the system.

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

REFERENCES

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

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 of predictive analysis of social media data]. *Avtomatika i programmnaia inzheneriia = Automatics and Software Enginery*, 2019, no. 4 (30), pp. 9–17. (In Russian).

3. 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.

4. Frolov V.V., Slipchenko S.E., Prikhod'ko O.Yu. Metod rascheta chisla klasterov dlya algoritma k-means [Clusters number calculating method for the k-means algorithm]. *Ekonomika. Informatika = Economics. Information technologies*, 2020, vol. 47 (1), pp. 213–225. DOI: 10.18413/2687-0932-2020-47-1-213-225.

5. Yandeks.Metrika [Yandex. Metrica]: website. Avaliable at: https://metrika.yandex.ru/list? (accessed 11.02.2022).

6. Google Analitika [Google Analytics]: website. Available at: https://analytics.google.com/analytics/web/provision/#/provision (accessed 11.02.2022).

7. Molodetskaya S.F., Shitova T.F. Otsenka effektivnosti saita na osnove tekhnologii nechetkogo upravleniya [Evaluation of the website effectiveness based on the technology of fuzzy control]. Voprosy upravleniya = *Management Issues*, 2020, no. 2 (63), pp. 39–49. DOI: 10.22394/2304-3369-2020-2-39-49.

8. Kaushik A. *Web analytics 2.0: the art of online accountability and science of customer centricity.* Indianapolis, Wiley, 2010 (Russ. ed.: Koshik A. *Veb-analitika 2.0 na praktike: tonkosti i luchshie metodiki*. Moscow, Dialektika Publ., 2019. 528 p.).

9. *Ahrefs – eto instrumenty i resursy SEO dlya rosta vashego poiskovogo trafika* [Ahrefs are SEO tools and resources for growing your search traffic]. Available at: https://ahrefs.com/ru/ (accessed 11.02.2022).

10. 2ip: website. (In Russian). Available at: https://2ip.ru/analizator/ (accessed 11.02.2022).

11. *RDS Bar – rasshirenie dlya seo analiza saita i stranits* [RDS Bar-extension for SEO analysis of the site and pages]. Available at: https://www.recipdonor.com/bar (accessed 11.02.2022).

12. *PR-CY. Servis samostoyatel'nogo prodvizheniya saita – Onlain instrumenty dlya vebmasterov, optimizatorov i kopiraiterov* [PR-CY. Self-promotion service – Online tools for webmasters, optimizers and copywriters]. Available at: https://pr-cy.ru/ (accessed 11.02.2022).

13. Babichev S. Optimization of information preprocessing in clustering systems of high dimension data. *Radio Electronics, Computer Science, Control*, 2014, no. 2, pp. 135–142. (In Russian). DOI: 10.15588/1607-3274-2014-2-19.

14. Baturkin S.A., Baturkina E.Yu., Zimenko V.A., Siginov I.V. Statisticheskie algoritmy klasterizatsii dannykh v adaptivnykh obuchayushchikh sistemakh [Statistical data clasterisation algorithms in adaptive training systems]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta = Vestnik of Ryazan State Radio Engineering University*, 2010, no. 1 (31), pp. 82–85.

15. Hartigan J.A., Wong M.A. Algorithm AS136: a K-means clustering algorithm. *Applied Statistics*, 1979, vol. 28 (1), pp. 100–108.

Для цитирования:

Калытюк И.С., Французова Г.А., Гунько А.В. Заключительные этапы проектирования системы сбора и предиктивного анализа данных социальных медиа // Системы анализа и обработки данных. – 2022. – № 1 (85). – С. 59–72. – DOI: 10.17212/2782-2001-2022-1-59-72.

For citation:

Kalytyuk I.S., Frantsuzova G.A., Gunko A.V. Zaklyuchitel'nye etapy proektirovaniya sistemy sbora i prediktivnogo analiza dannykh sotsial'nykh media [Final stages of designing a system for collecting and predictive analysis of social media data]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2022, no. 1 (85), pp. 59–72. DOI: 10.17212/2782-2001-2022-1-59-72.