УДК 004.89 *Научная статья* DOI: 10.18287/2223-9537-2024-14-4-569-581

Фокусированный сбор и обработка открытых данных социальных медиа

© 2024, И.О. Датьев ✉, А.М. Фёдоров, А.А. Ревякин

Институт информатики и математического моделирования им. В.А. Путилова Кольского научного центра РАН (ИИММ КНЦ РАН), Апатиты, Россия

Аннотация

Рассматривается развитие технологий сбора данных и осложняющие этот процесс особенности. Представлены методы фокусировки различного уровня: от управления границами сканирования до использования различных свойств веб-страниц. В данной работе термин «фокусировка» используется для более точной передачи специфических особенностей процесса целенаправленного сбора и обработки открытых данных социальных медиа. Описываемый процесс является многоступенчатым, и для его организации используются механизмы адаптивного управления, которые относительно заданной цели имеют разнонаправленный характер. В процессе управления задаваемые ограничения сужаются или расширяются, т.е. фокусируются на заданной цели. Представлен опыт проектирования архитектуры и программной реализации функций информационной системы, позволяющей производить автоматизированный фокусированный сбор и обработку открытых данных социальных медиа.

Ключевые слова: *фокусированный веб-сканер, социальная сеть, информационная система, интеллектуальный анализ, методы фокусировки сбора данных.*

Цитирование: *Датьев И.О., Федоров А.М., Ревякин А.А.* Фокусированный сбор и обработка открытых данных социальных медиа. *Онтология проектирования*. 2024. Т.14, №4(54). С.569-581. DOI:10.18287/2223-9537-2024-14-4-569-581.

Финансирование: Исследование выполнено в рамках государственного задания ИИММ КНЦ РАН Министерства науки и высшего образования РФ, темы НИР: «Методология создания информационно-аналитических систем поддержки управления региональным развитием, основанных на формирующем искусственном интеллекте и больших данных» (шифр темы FMEZ-2022-0007, номер государственного учёта 122022800551-0); «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределённости и риска» (шифр темы FMEZ-2025-0053).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Объём данных и скорость их роста в Интернете привели к трудности фокусированного сбора данных. Первой технологией, позиционированной для сбора данных из глобальной сети, является *web scraping* [1] — технология получения данных путём извлечения их со страниц веб-ресурсов, которая чаще всего представляет собой автоматизированный процесс выполнения программным кодом *GET*-запросов[1] к целевому веб-ресурсу.

Наиболее известными сканирующими виртуальными роботами (ботами) являются: *Xenon*, *BingBot*, *Googlebot*, *Yandex*, *ChatGPT* и др. Кроме того, веб-сканирование предлагается в виде услуги: программное обеспечение как услуга или данные как услуга. Эти услуги позволяют автоматически собирать любые общедоступные данные в Интернете. Примером использования веб-сканирования является мониторинг цен на рынках электронной коммерции, позволяющий клиентам отслеживать ценовые стратегии своих конкурентов. Сканиро-

[1] Методы *HTTP* запроса. https://developer.mozilla.org/ru/docs/Web/HTTP/Methods.

вание используется для агрегирования данных — процесса, позволяющего извлекать, преобразовывать, анализировать и визуализировать данные из нескольких источников.

Технологии сбора и обработки данных сети Интернет, а также некоторые особенности предоставления данных обсуждались в работах [2, 3]. К особенностям, осложняющим сбор данных, относятся динамически загружаемый и дублирующийся контент, а также защита от ботов. Некоторые веб-ресурсы используют для защиты от ботов методы обнаружения и блокировки с помощью определённых признаков [4]: несвойственная человеку скорость взаимодействия с элементами интерфейса; повторяющиеся однотипные действия; использование ссылок-приманок, которые содержатся только в коде веб-сайта и не видны обычным пользователям. Способы блокировки ботов состоят в следующем: запрет доступа к ресурсу с определённого *IP*-адреса; выдача страницы с сообщением об ошибке вместо страницы с запрошенным контентом; запрет доступа для идентификатора пользователя-злоумышленника при попытке авторизации на веб-ресурсе.

Дублирование веб-страниц в пределах домена бывает частичным или полным [3]. Полное дублирование связано с использованием инструментов управления данными на сайте и появлением документов-дублей, имеющих различный *URL*. Частичное дублирование встречается при применении инструментов управления данными на сайте (использование фильтрации и сортировок) - наиболее сложное для выявления, особенно если дублированные фрагменты текста перемешаны между собой или с фрагментами уникального текста.

1 Методы фокусировки при обработке веб-документов

Фокусированный сканер — это сканер, который собирает веб-страницы, удовлетворяющие определённому свойству, расставляя приоритеты на границе сканирования и управляя процессом исследования гиперссылок. Ограничения могут быть простыми и чёткими, например, сканировать страницы только определённого домена, или нечёткими, например, сканировать страницы о футболе или сканировать страницы с большими значениями рейтинга. Термин «фокусированный сканер» вместе с классификатором текста предложены в работе [5], а машинное обучение для определения границ сканирования применено в [6]. Машинное обучение и сбор данных об определённой предметной области (ПрО), в т.ч. с использованием графовых моделей, развивались в работах [7, 8].

Помимо, управления границами сканирования, для фокусировки могут использоваться и другие свойства веб-страниц. Темы страницы - важное свойство, которое привело к появлению термина «*тематический сканер*» [6, 9]. Сбор данных также может производиться по заданной эмоциональной окраске текста - тональности [10]. В [11] показана важность информации о пространственном расположении объектов на веб-странице. В [12] выделены три типа сегментации: визуальная, лингвистическая и денситометрическая. Визуальная сегментация [13] на основе алгоритма машинного зрения различает разделы веб-страницы. Лингвистическая сегментация основана на использовании языковых единиц (слов, слогов, предложений) в качестве статистических показателей для выявления структурных закономерностей в тексте. Денситометрическая сегментация [12] присваивает каждому веб-ресурсу плотность текста (определяется как результат деления количества токенов на количество строк). Денситометрическая сегментация работает хорошо, как визуальная сегментация, и быстро, как лингвистическая.

В *семантическом сканере* для фокусировки используется информация о семантике, чаще всего — онтологии ПрО для представления тематических карт и связывания веб-страниц с соответствующими онтологическими концепциями, что позволяет производить категоризацию веб-документов [14, 15]. Онтологии могут автоматически обновляться в процессе ска-

нирования [16]. Однако при использовании онтологии ПрО необходимо привлечение экспертов ПрО для формирования концептов и отношений. В работе [17] предложено вместо онтологии ПрО выполнять фокусировку с помощью схемы представления знаний, которая генерируется для каждого веб-документа и хранится в базе знаний [18]. Схема представления знаний менее выразительна, чем онтология ПрО (не определяет никаких правил или ограничений в отношении данных), но не зависит от ПрО и сохраняет преимущество использования технологий *Semantic Web*[2], таких как *Resource Description Framework* (*RDF*) *Schema* - набор классов и свойств для модели представления знаний, составляющий основу для описания онтологий с использованием расширенного *RDF*-словаря для структуры *RDF*-ресурсов[3].

Целью применения методов фокусировки является повышение объёма обладающих определёнными характеристиками собранных данных и сокращение времени сбора с учётом необходимости обхода блокировок со стороны администраторов веб-ресурсов. В фокусированных сканерах для повышения эффективности сбора данных всё чаще используются алгоритмы из области искусственного интеллекта[4].

2 Система фокусированного сбора открытых данных

Концептуальная схема технологии фокусировки сбора данных представлена на рисунке 1.

Разработанная информационная система (ИС) позволяет собирать данные с определённых веб-ресурсов (социальных медиа), генерировать отчёты различных типов, сравнивать данные друг с другом и масштабировать. ИС можно разделить на две части: серверную и интерфейсную.

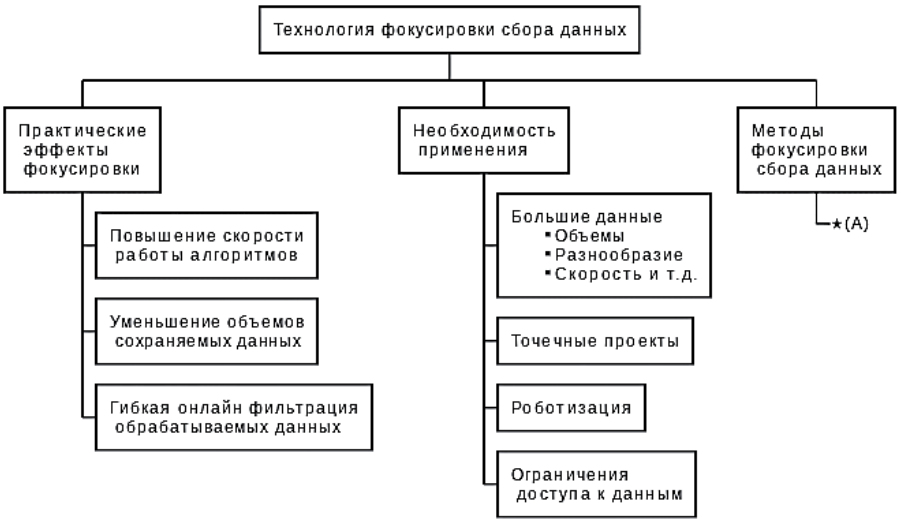

Рисунок 1 – Концептуальная схема технологии фокусировки сбора данных (*(A) обозначает переход к детальному представлению методов фокусировки данных – показана на рисунке 5)

2.1 Архитектура ИС

ИС представляет собой несколько взаимодействующих через сеть компонентов. Каждый компонент выполнен в виде отдельного докер-контейнера[5], что позволяет размещать компоненты на различных физических серверах. Все используемые контейнеры запускаются на одном сервере. Положительным эффектом от использования докер-контейнеров является возможность оперативно развернуть систему на любом физическом сервере, имеющем до-

[2] *Semantic Web Activity*. http://www.w3.org/2001/sw.
[3] *Resource Description Framework (RDF) Schema Specification*. http://www.w3.org/TR/2000/CR-rdf-schema-20000327.
[4] 70 *Top AI Web Crawler Tools*. https://topai.tools/s/web-crawler.
[5] Докер-контейнер — стандартизированный, изолированный и портативный пакет программного обеспечения.

статочно оперативной и дисковой памяти. Исключение составляет база данных (БД) *MongoDB*[6], расположенная на сервере ИИММ КНЦ РАН.

Формирование каждого контейнера задаётся инструкциями в *Dockerfile*, который определяет структуру и конфигурацию контейнера. Этот файл содержит все необходимые инструкции для создания и настройки образа контейнера, включая установку зависимостей, настройку среды выполнения и другие параметры. Такой подход обеспечивает изоляцию и независимость каждого компонента системы, а также упрощает процесс их разработки, тестирования и серверного размещения.

В настоящее время система размещается на двух отдельных физических серверах. На основном сервере развернуты все компоненты системы и организован доступ для администраторов верхнего уровня. На резервном сервере организован доступ для работы операторов системы, в функции которых не входит конфигурирование системы, а только работа с потоками собираемых данных. Различная функциональная компоновка серверов обеспечивается запуском нужного подмножества докер-контейнеров.

2.2 Основные компоненты ИС

Серверная часть приложения реализована на основе библиотеки для создания веб-серверов. В процессе работы автоматически генерируется и разворачивается инструмент документирования и тестирования.

База данных. В ИС использована СУБД *MongoDB*. БД содержит следующие коллекции:

- *Users* – список зарегистрированных пользователей;
- *GlobalSettings* – глобальные настройки системы;
- *Tasks* – список задач;
- *Dialogs* – список диалогов (чатов) из социальных сетей;
- *DialogsHistory* – список историй взаимодействия с коллекцией *Dialogs*;
- *Posts* – список постов из социальных сетей;
- *Messages* – список исходящих сообщений, отправленных через систему;
- *Notifications* – список оповещений для пользователей (*Users*);
- *Themes* – список тем для поиска ключевых слов / фраз внутри постов (*Posts*).

Все коллекции представляют собой набор данных, связанных друг с другом с помощью уникальных идентификаторов.

Selenium – инструмент для автоматизации действий веб-браузера, используется для автоматизированного тестирования приложений. *Selenium* разворачивает виртуальный браузер *Firefox* на сервере, позволяет сохранять данные сессии, эмулировать действия реального пользователя посредством управления курсором и клавиатурой. В результате из веб-документа выделяется необходимый контент и записывает в БД.

Клиентская часть представляет собой оптимизированное веб-приложение для отображения интерфейсов системы.

Варианты использования ИС представлены на рисунке 2. В ИС разработаны и программно реализованы следующие роли пользователей: Админ, Суперадмин, Наблюдатель.

Админ доступны все инструменты для конфигурирования мониторинга социальных сетей. Роль Суперадмина расширена относительно Админ управлением пользователями – создание, удаление, назначение ролей пользователям, а также конфигурирование мониторинга каждого из них. Роль Наблюдателя позволяет просматривать результаты мониторинга, сконфигурированного для Наблюдателя пользователем Суперадмин.

[6] *MongoDB* — документоориентированная система управления базами данных (СУБД) с открытым исходным кодом.

2.3 Интеграция ИС с социальными медиа

При проектировании программного модуля эмуляции действий пользователя для сбора данных социальных сетей выделены следующие этапы (рисунок 3):

1) *Авторизация пользователя* в мессенджере. После ввода пароля может быть получен *QR*-код или запрошен код безопасности. После успешной авторизации создаётся и сохраняется сессия веб-браузера с авторизованными данными пользователя.
2) *Получение списка диалогов*. Загружаются все диалоги пользователя, для каждого диалога извлекается его название и добавляется в список.
3) *Получение сообщений из диалогов*. Извлекаются текстовые сообщения и добавляются в список для каждого диалога.
4) *Завершение сессии*.

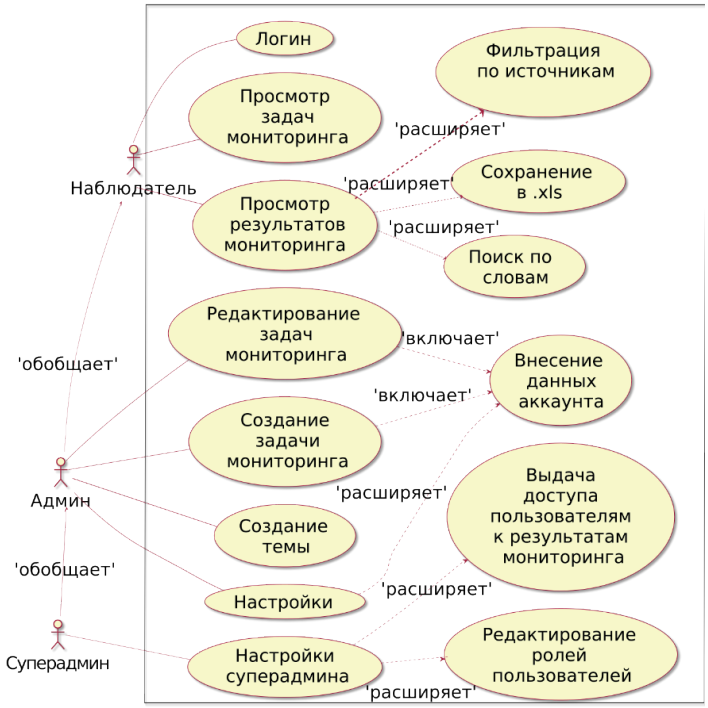

Рисунок 2 – Диаграмма использования системы фокусированного сбора и обработки открытых данных

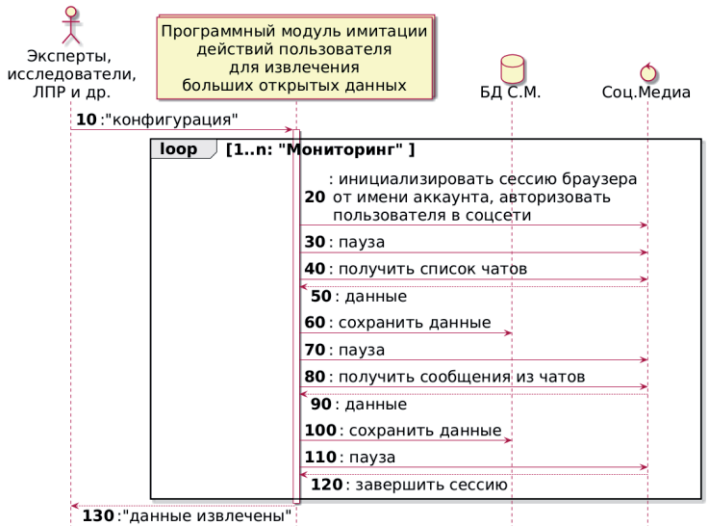

Рисунок 3 – Диаграмма последовательности. Логика работы модуля эмуляции действий пользователя

Программный модуль эмуляции пользователя производит временные задержки между различными действиями пользователя. Это необходимо для учёта технических особенностей реализации социальных сетей. Разные этапы каждой задачи занимают разное количество времени в зависимости от характеристик и текущего состояния Интернет-соединения, производительности и загруженности серверов социальной сети, количества задач, производительности и загруженности серверов, на которых развернута ИС. Основные изменяемые модельные параметры эмуляции соответствуют паузам на рисунке 3. Варьирование значениями этих параметров позволяет повысить процент успешно собранных данных и избежать блокировок со стороны социальной сети. Дополнительная сложность сбора данных с помощью эмуляции действий пользователя заключается в периодической подмене кодовых имен объектов, размещённых на веб-странице, производимой социальной сетью. Эта особенность преодолевается посредством использования неизменных названий объектов для последующей идентификации вспомогательных веб-объектов.

Пример интерфейса ИС в виде фрагмента ленты сообщений социальных сетей представлен на рисунке 4 .

3 Проектирование и реализация алгоритмов фокусировки

Разработанная ИС предназначена для организации мониторинга информационного пространства различных веб-ресурсов. Модульная архитектура системы позволяет при необходимости нарастить её потенциал и расширить границы обрабатываемого информационного пространства. Возможности ИС нацелены на работу с определённым множеством ресурсов.

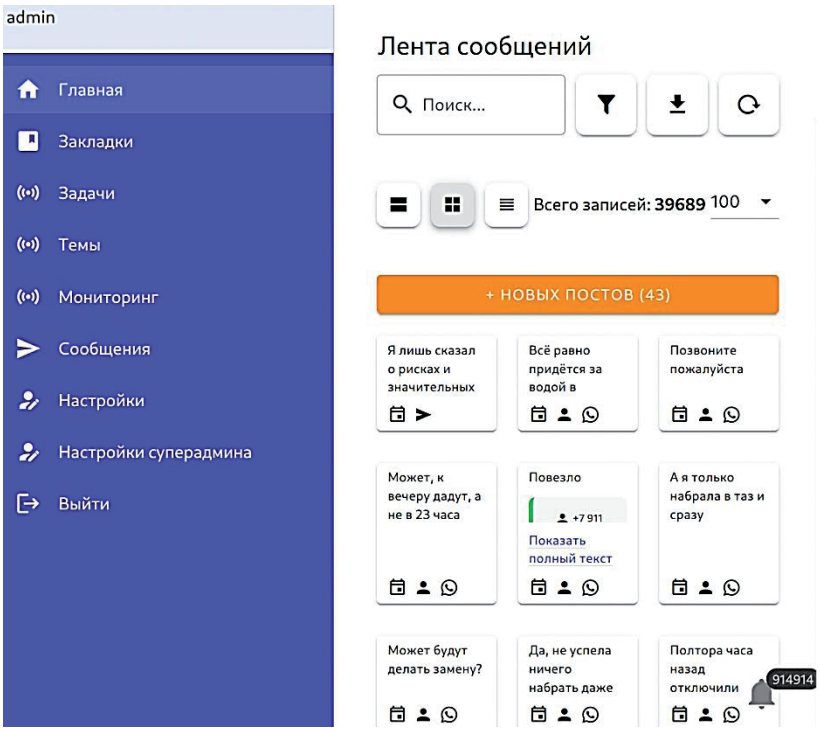

Рисунок 4 – Фрагмент клиентского приложения ленты сообщений социальных сетей

На рисунке 5 представлены методы фокусировки сбора данных (см. рисунок 1).

Классификация алгоритмов фокусировки. Для конфигурации описываемой ИС применены следующие варианты фокусировки:

1 ограничение области поиска: выбранное подмножество социальных сетей;
2 ограничение объёма извлекаемых данных: только открытые источники;
3 анализ адресов (ссылок): работа только с подходящими адресами;
4 тематическая фокусировка: ключевые слова и тематические группы слов.

Используемые варианты обеспечивают ограничение обрабатываемого информационного пространства. Первые два - концептуальные - задают общие ограничения, которые на этапе проектирования системы влияют на выбор потенциально используемых в работе подходов, инструментов и технологий. Третий и четвёртый уровни - оперативные - предполагают получение эффектов от фокусировки в процессе непосредственной работы ИС, сконфигурированной с теми или иными параметрами. Перечисленные способы фокусировки используются в данной работе и представлены на рисунке 6 в виде *UML*-диаграммы вариантов использования.

Фокусировка на открытых данных социальных медиа определяется на этапе проектирования архитектуры системы. Фокусировка на заданной тематической повестке (например, «ЖКХ» или «здравоохранение») или на заданном уровне тональности (например, «позитив») определяется конкретной конфигурацией модулей системы, непосредственно обрабатывающих данные.

Эффекты от фокусировок можно использовать для уменьшения объёмов сохраняемых данных, увеличения скорости их обработки и для фильтрации данных перед их размещением в разных хранилищах или для обработки разными агентами.

Первая фокусировка – *социальные сети*. В рамках решаемых задач запланирован тематический анализ высказываний пользователей сети Интернет, которые они оставляют в виде реплик в чатах и комментариев к публикациям. Существует много виртуальных площадок, на которых пользователи могут обмениваться сообщениями. Наилучшими источниками для получения таких данных являются социальные сети и мессенджеры. Для отработки технологий мониторинга сформированы структуры данных, программные и пользовательские интерфейсы, позволяющие организовать единый подход к различным источникам.

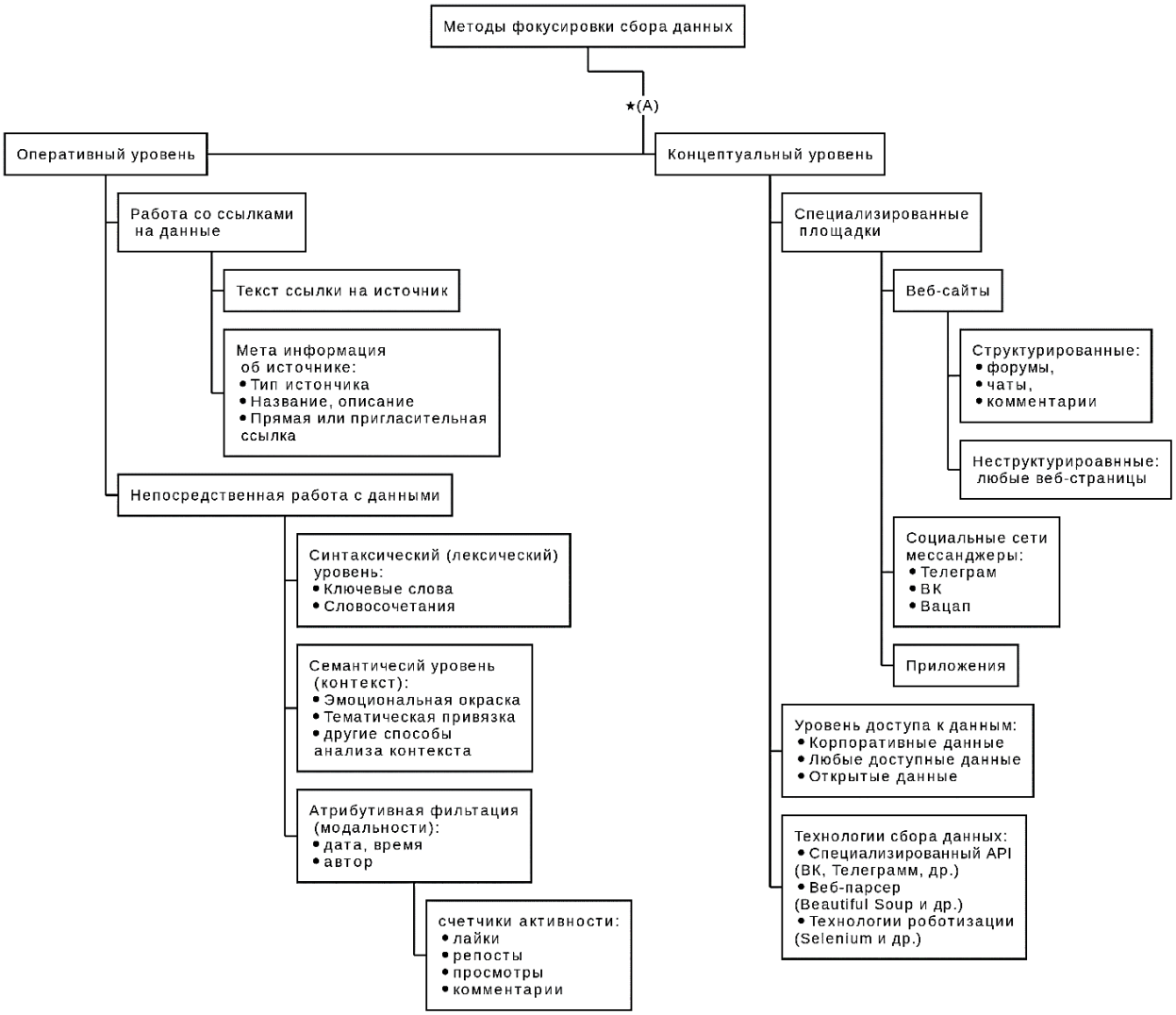

Рисунок 5 – Методы фокусировки сбора данных

Вторая фокусировка – *открытые источники*. Задача по исследованию социального дискурса на основе сообщений в социальных сетях имеет свою специфику. Такие исследования ориентированы на то, что сообщения в виртуальном пространстве социальных сетей отражают реальные настроения в обществе. Однако в общем случае виртуальное информационное пространство предполагает наличие в нём, помимо людей, виртуальных акторов (ботов или технических аккаунтов), действующих по заранее заданным алгоритмам и преследующих определённые цели. Количество и активность этих виртуальных сущностей может формировать повестку дискурса, статистически влияя на голоса реальных пользователей. Потенциально подвержены большему влиянию виртуальных акторов открытые площадки.

Наличие процедур проверки подлинности участников дискуссии сокращает влияние виртуальных факторов на реальный дискурс. Организация работы виртуальной площадки невозможна без работы средств автоматизации, поэтому степень открытости виртуальной площадки пропорциональна степени потенциального влияния на неё этих средств.

Фокусировка на открытых источниках позволяет направить работу ИС на извлечение и обработку данных из этих источников, и не тратить силы и средства на преодоление систем защиты и организацию других действий по проникновению на закрытые площадки.

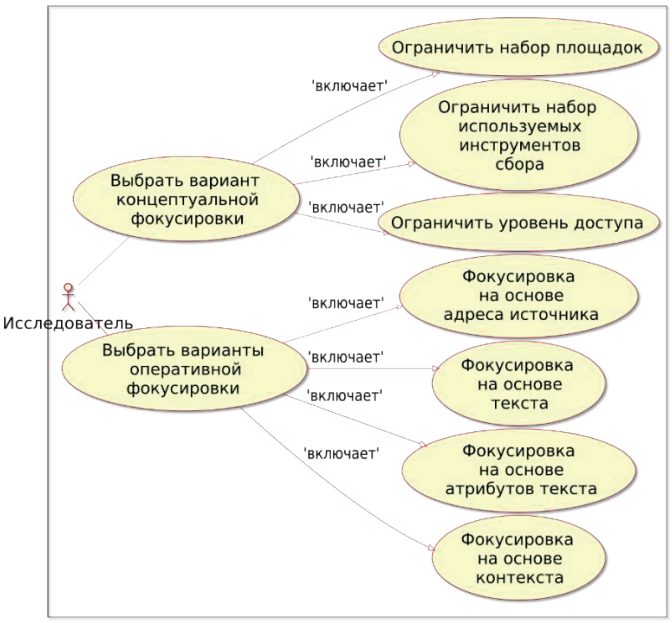

Рисунок 6 – Диаграмма использования.
Конфигурирование фокусировки сбора данных

В дополнение к концептуальной фокусировке используются оперативные алгоритмы фокусировки. В частности, предусматривается проверка источников из заданного списка на предмет соответствия ряду условий. В простом случае — это заранее сформированный перечень ссылок. В ИС также используется задание списка источников в виде сохранённых в аккаунте подписок на эти источники, который со временем может изменяться. Оперативные алгоритмы фокусировки учитывают эту особенность следующими способами.

Проверка типа ресурса. Типология ресурсов определяет размещаемый пользовательский контент. В случае с изучением социального дискурса интерес представляют ресурсы, на которых пользователи имеют возможность от своего имени публиковать собственные тексты и писать к ним комментарии.

Проверка прямой ссылки на ресурс. На некоторых виртуальных площадках доступ к определённым ресурсам осуществляется только через прямое приглашение или по пригласительной ссылке. Проверка сочетания указанных факторов позволяет провести фокусировку мониторинга на таких источниках.

Проверка названий ресурсов. В рамках задачи по исследованию дискурса на виртуальных площадках ставятся подзадачи, связанные с тематической фильтрацией обрабатываемых источников. В большинстве случаев названия таких источников отражают их тематическую направленность. В зависимости от первоначального списка источников данный способ фокусировки позволяет потенциально сократить время обработки данных.

Основным эффектом от применения описанного вида фокусировки является сокращение количества фактически обрабатываемых ресурсов, что приводит к экономии времени, затрачиваемого на их мониторинг.

Третья фокусировка *– тематическая фильтрация*. Данный способ фокусировки основан на лексическом анализе текстов. На базовом уровне темы задаются с помощью ключевых слов и их групп – тематических наборов. Наличие одного из слов в анализируемом тексте относит его к соответствующей теме. Одной из сложностей в данном случае является задание как можно более полного множества словоформ, отражающих отслеживаемую тему. На основе тематической окраски текстов производится дальнейшая их обработка - сохранение в БД, запись в ленту сообщений и оповещений, визуальная подсветка при отображении и др.

В общем случае, аналогично тематической фильтрации, можно организовать фокусировку по любым другим поставленным в соответствие тексту атрибутивным данным. Одной из особенностей текстов социальных сетей является их тесная связь с дополнительной информацией: автор, дата публикации, счётчики активности (комментарии, просмотры и др.), мультимедийные приложения и др.

Компоненты текстов и их атрибуты в различных сочетаниях широко используются для кластеризации, классификации и других способов дифференциации текстов. Например, некоторые алгоритмы тематического моделирования используют в своей работе модальности текста, т.е. их сопутствующие атрибуты [19]. В результате финальные тематические распределения вычисляются на основе текстов и на основе их атрибутов в соответствии с заданными пропорциями. В данной работе представленная фокусировка применялась только в отношении тематической фильтрации текстов социальных сетей. Полученные в данном исследовании результаты представлены на рисунке 7.

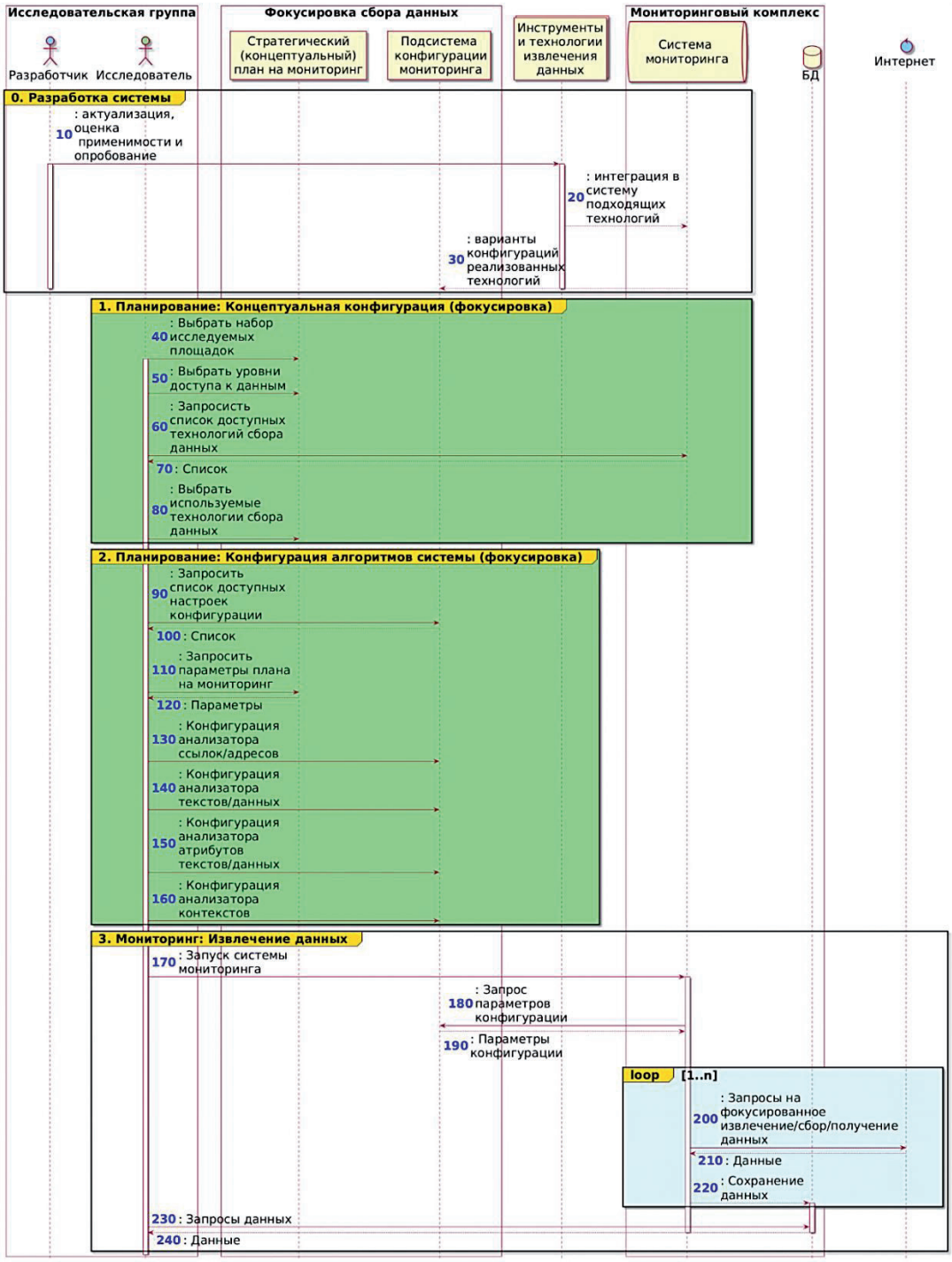

Рисунок 7 – Схема фокусировки сбора данных

Заключение

Представлен опыт проектирования ИС фокусированного сбора открытых данных онлайновых социальных сетей. Разработаны и реализованы алгоритмы фокусированного сбора данных, которые представлены в многоуровневой форме.

Список источников

[1] ***Boeing G., Waddell P.*** New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings // Journal of Planning Education and Research. 2016. DOI:10.1177/0739456X16664789. arXiv:1605.05397.

[2] ***Кулешов С.В., Зайцева А.А., Левашкин С.П.*** Технологии и принципы сбора и обработки неструктурированных распределенных данных с учетом современных особенностей предоставления медиа-контента // Информатизация и связь. 2020. № 5. С.22-28. DOI 10.34219/2078-8320-2020-11-5-22-28. EDN FMQNTT.

[3] ***Кулешов С.В., Зайцева А.А.*** Феноменологическое описание процессов сбора и обработки интернет-документов // Изв. вузов. Приборостроение. 2023. Т.66, № 12. С.1002-1010. DOI:10.17586/0021-3454-2023-66-12-1002-1010.

[4] ***Москаленко А.А., Лапонина О.Р., Сухомлин В.А.*** Разработка приложения веб-скрапинга с возможностями обхода блокировок // Современные информационные технологии и ИТ-образование. 2019. Т.15, №2. С.413-420. DOI: 10.25559/SITITO.15.201902.413-420.

[5] ***Soumen Chakrabarti***. Focused Web Crawling, in the Encyclopedia of Database Systems. Dynamic topic models // In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA, ACM, 2006. P.113–120. DOI:10.1145/1143844.1143859.

[6] ***Soumen Chakrabarti, Martin van den Berg, Byron Dom.*** Focused crawling: a new approach to topic-specific Web resource discovery // Computer Networks, Volume 31, Issues 11–16, 1999, P.1623-1640. DOI: 10.1016/S1389-1286(99)00052-3.

[7] Using Reinforcement Learning to Spider the Web Efficiently / Jason Rennie and Andrew McCallum. ICML 1999.

[8] ***Diligenti M., Coetzee F., Lawrence S., Giles C.L., and Gori M.*** (2000). Focused crawling using context graphs Archived 2008-03-07 at the Wayback Machine // In Proceedings of the 26th International Conference on Very Large Databases (VLDB). P.527-534, Cairo, Egypt.

[9] ***Taylan D., Poyraz M., Akyokuş S. and Ganiz M.C.*** Intelligent focused crawler: Learning which links to crawl // 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey. 2011. P.504-508. DOI: 10.1109/INISTA.2011.5946150.

[10] ***Tianjun Fu, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen.*** 2012. Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers // ACM Trans. Inf. Syst. 30, 4, Article 24 (November 2012), 30 pages. DOI: 10.1145/2382438.2382443.

[11] ***Yu Y.B., Huang S.L., Tashi N., Zhang H., Lei F., Wu L.Y.*** A Survey about Algorithms Utilized by Focused Web Crawler // J. Electron. Sci. Technol. 2018, 16, 129. DOI:10.11989/JEST.1674-862X.70116018.

[12] ***Kohlschütter C., Nejdl W.*** A densitometric approach to web page segmentation // Proceedings of the 17th ACM conference on Information and knowledge management, New York. 2008. P.1173-1182.

[13] ***Sun Y., Jin P., Yue L.*** A Framework of a Hybrid Focused Web Crawler // Future Generation Communication and Networking Symposia, 2008. FGCNS '08. Second International Conference, Sanya, 2008. P.50-53.

[14] ***Hassan T., Cruz C., Bertaux A.*** Ontology-based Approach for Unsupervised and Adaptive Focused Crawling // In Proceedings of the International Workshop on Semantic Big Data, Chicago, IL, USA, 19 May 2017. ACM: New York, NY, USA, 2017. P.21–26.

[15] ***Boukadi K., Rekik M., Rekik M., Ben-Abdallah H.*** FC4CD: A new SOA-based Focused Crawler for Cloud service Discovery // Computing 2018, 100, P.1081-1107. DOI:10.1007/s00607-018-0600-2.

[16] ***Dong H., Hussain F.K.*** SOF: A semi-supervised ontology-learning-based focused crawler // Concurrency and Computation: Practice and Experience. 25(12). (August 2013). P.1623-1812.

[17] ***Hernandez J., Marin-Castro H.M., Morales-Sandoval M.*** A Semantic Focused Web Crawler Based on a Knowledge Representation Schema // Applied Sciences. 2020; 10(11):3837. DOI:10.3390/app10113837.

[18] ***Krótkiewicz M., Wojtkiewicz K., Jodłowiec M.*** Towards Semantic Knowledge Base Definition // In Biomedical Engineering and Neuroscience / Hunek, W.P., Paszkiel, S., Eds.; Springer International Publishing: Cham, Switzerland, 2018. P.218–239.

[19] ***Датьев И.О., Федоров А.М.*** Аддитивная регуляризация при тематическом моделировании текстов сообществ онлайновых социальных сетей. *Онтология проектирования*. 2022. Т.12, №2(44). С.186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.

Сведения об авторах

Датьев Игорь Олегович 1981 г. рождения. Окончил Кольский филиал Петрозаводского государственного университета (2004). К.т.н. (2011). В ИИММ КНЦ РАН старший научный сотрудник, учёный секретарь. Автор более 100 научных работ в области разработки моделей и технологий для региональных информационно-коммуникационных систем. Author ID (РИНЦ): 180256; Author ID (Scopus): 56070103900; Researcher ID (WoS): J-1839-2018. *i.datyev@ksc.ru.* ✉

Федоров Андрей Михайлович 1978 г. рождения. Окончил Кольский филиал Петрозаводского государственного университета (2000). К.т.н. (2005). В ИИММ КНЦ РАН ведущий научный сотрудник, заместитель директора по научной работе (с 2017 г.). Доцент кафедры информатики и вычислительной техники в филиале Мурманского арктического университета (МАУ) в г. Апатиты. Область научных интересов сосредоточена на разработке моделей и технологий информационной поддержки для регионального управления. Author ID (RSCI): 4285-9780; Author ID (Scopus): 57203929412; Researcher ID (WoS): D-5859-2016. *a.fedorov@ksc.ru.*

Ревякин Андрей Андреевич 1992 г. рождения. Окончил Запорожский национальный технический университет (2014). Магистрант первого курса филиал МАУ в г. Апатиты по специальности 09.04.02 Информационные системы и технологии. Программист ИИММ КНЦ РАН. Руководитель отдела *Frontend* разработки в компании *Happy Job* (Москва). Область научных интересов сосредоточена в разработке информационных систем с большими массивами данных, открытыми источниками информации. Author ID (RSCI): Author ID (ORCID): 0009-0006-3170-3990. *andrewreviakin@yandex.ru.*

Поступила в редакцию 08.07.2024, после рецензирования 02.10.2024. Принята к публикации 28.10.2024.

Scientific article DOI: 10.18287/2223-9537-2024-14-4-569-581

Focused collection and processing of open social media data

© 2024, I.O. Datyev ✉, A.M. Fedorov, A.A. Reviakin

Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Center RAS (IIMM KSC RAS), Apatity, Russia

Abstract

The article addresses the development of data collection technologies and the complexities that challenge this process. Methods for focusing at various levels are discussed, ranging from managing scanning boundaries to leveraging diverse properties of web pages. Here, the term "focusing" is used to accurately reflect the specific characteristics of targeted data collection and processing of open social media data. This process is multi-stage, employing adaptive control mechanisms that adjust dynamically toward the specified objective. During control, these defined constraints are either narrowed or broadened to align with the target goal. The article also presents insights from the design of an information system's architecture and software, enabling automated, focused collection and processing of open social media data.

Keywords: *focused web crawler, social network, information system, intelligent analysis, data collection focusing methods.*

For citation: *Datyev IO, Fedorov AM, Reviakin AA.* Focused collection and processing of open social media data [In Russian]. *Ontology of designing.* 2024; 14(4): 569-581. DOI:10.18287/2223-9537-2024-14-4-569-581.

Financial Support: The work is supported by the Ministry of Science and Higher Education of the Russian Federation. Topic title: Methodology for creating information and analytical systems to support the management of regional development based on formative artificial intelligence and big data (reg.n. 122022800551-0). Subsequent topic title: Methods

and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specifics under conditions of uncertainty and risk (FMEZ-2025-0053).

Conflict of interest: The authors declare no conflict of interest.

List of figures

References

[1] ***Boeing G, Waddell P.*** New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. Journal of Planning Education and Research. 2016. DOI:10.1177/0739456X16664789. arXiv:1605.05397.

[2] ***Kuleshov SV, Zaitseva AA, Levashkin SP.*** Technologies and principles of collecting and processing unstructured distributed data taking into account modern features of providing media content [In Russian]. Informatization and Communication. 2020; 5: 22-28. DOI 10.34219/2078-8320-2020-11-5-22-28. EDN FMQNTT.

[3] ***Kuleshov SV, Zaitseva AA.*** Phenomenological description of the processes of collecting and processing Internet documents [In Russian]. Izv. universities Instrumentation. 2023; 66(12): 1002-1010. DOI:10.17586/0021-3454-2023-66-12-1002-1010.

[4] ***Moskalenko AA, Laponina OR, Sukhomlin VA.*** Development of a web scraping application with the ability to bypass blocking [In Russian]. Modern information technologies and IT education. 2019; 15(2): 413-420. DOI: 10.25559/SITITO.15.201902.413-420.

[5] ***Chakrabarti S***. Focused Web Crawling, in the Encyclopedia of Database Systems.Dynamic topic models. In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA, ACM. 2006. P.113–120. DOI:10.1145/1143844.1143859.

[6] ***Chakrabarti S, van den Berg M, Dom B.*** Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 1999; 31(11–16): 1623-1640. DOI: 10.1016/S1389-1286(99)00052-3.

[7] Using Reinforcement Learning to Spider the Web Efficiently. Jason Rennie and Andrew McCallum. ICML 1999.

[8] ***Diligenti M, Coetzee F, Lawrence S, Giles CL, Gori M***. Focused crawling using context graphs Archived 2008-03-07 at the Wayback Machine. In Proceedings of the 26th International Conference on Very Large Databases (VLDB), Cairo, Egypt. 2000. P. 527-534.

[9] ***Taylan D, Poyraz M, Akyokuş S, Ganiz MC.*** Intelligent focused crawler: Learning which links to crawl. 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey. 2011. P. 504-508. DOI: 10.1109/INISTA.2011.5946150.

[10] ***Tianjun Fu, Ahmed Abbasi, Daniel Zeng, Hsinchun Chen.*** 2012. Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers. ACM Trans. Inf. Syst. 2012; 30, 4, Article 24. 30 p. https://doi.org/10.1145/2382438.2382443.

[11] ***Yu YB, Huang SL, Tashi N, Zhang H, Lei F, Wu LY.*** A Survey about Algorithms Utilized by Focused Web Crawler. J. Electron. Sci. Technol. 2018; 16, 129. DOI:10.11989/JEST.1674-862X.70116018.

[12] ***Kohlschütter C, Nejdl W.*** A densitometric approach to web page segmentation. Proceedings of the 17th ACM conference on Information and knowledge management, New York. 2008. P. 1173-1182.

[13] ***Sun Y, Jin P, Yue L.*** A Framework of a Hybrid Focused Web Crawler. Future Generation Communication and Networking Symposia 2008 (FGCNS '08). Second International Conference, Sanya. 2008. P. 50-53.

[14] ***Hassan T, Cruz C, Bertaux A.*** Ontology-based Approach for Unsupervised and Adaptive Focused Crawling. In Proceedings of the International Workshop on Semantic Big Data, Chicago, IL, USA, 19 May 2017. ACM: New York, NY, USA. 2017. P. 21–26.

[15] ***Boukadi K, Rekik M, Rekik M, Ben-Abdallah H.*** FC4CD: A new SOA-based Focused Crawler for Cloud service Discovery. Computing. 2018; 100. P.1081–1107. DOI:10.1007/s00607-018-0600-2.

[16] ***Dong H, Hussain FK.*** SOF: A semi-supervised ontology-learning-based focused crawler. Concurrency and Computation: Practice and Experience. 2013; 25(12): 1623-1812.

[17] ***Hernandez J, Marin-Castro HM, Morales-Sandoval M.*** A Semantic Focused Web Crawler Based on a Knowledge Representation Schema. Applied Sciences. 2020; 10(11):3837. DOI:10.3390/app10113837.

[18] ***Krótkiewicz M, Wojtkiewicz K, Jodłowiec M.*** Towards Semantic Knowledge Base Definition. In Biomedical Engineering and Neuroscience. Hunek, W.P., Paszkiel, S., Eds.; Springer International Publishing: Cham, Switzerland. 2018. P. 218–239.

[19] ***Datyev IO, Fedorov AM.*** Additive regularization in topic modeling of texts from communities of online social networks [In Russia]. *Ontology of designing*. 2022; 12, 2(44): 186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.

About the authors

Igor Olegovich Datyev (b. 1981) graduated from the Kola branch of Petrozavodsk State University in 2004. Candidate of Technical Sciences (2011). At the IIMM KSC RAS, he is a senior research and scientific secretary. Author of more than 100 scientific papers in the field of developing models and technologies for regional information and communication systems. Author ID (RSCI): 180256; Author ID (Scopus): 56070103900; Researcher ID (WoS): J-1839-2018. *i.datyev@ksc.ru* ✉.

Andrei Mikhailovich Fedorov (b. 1978) graduated from the Kola branch of the Petrozavodsk State University (2000). Candidate of Technical Sciences (2005). At the IIMM KSC RAS, he is a leading researcher, deputy director for research work (since 2017). Associate Professor at the Department of Informatics and Computer Engineering at the Murmansk Arctic University (MAU) branch in Apatity. Research interests focus on the development of models and technologies of information support for regional management. Author ID (RSCI): 4285-9780; Author ID (Scopus): 57203929412; Researcher ID (WoS): D-5859-2016. *a.fedorov@ksc.ru.*

Andrey Andreevich Reviakin (b. 1992) graduated from Zaporozhye National Technical University (2014). First-year master's student of the Murmansk Arctic University branch in Apatity, specialty 09.04.02 Information systems and technologies. Programmer at the IIMM KSC RAS. Head of Frontend Development Department at Happy Job (Moscow). The area of scientific interests is concentrated in the development of big data information systems, and open sources of information. Author ID (RSCI): Author ID (ORCID): 0009-0006-3170-3990. *andrewreviakin@yandex.ru*.

Received July 08, 2024. Revised October 2, 2024. Accepted October 28, 2024.