УДК 004

РАЗРАБОТКА ПРОГРАММНОГО МОДУЛЯ ОБНАРУЖЕНИЯ АКТУАЛЬНЫХ ИСТОЧНИКОВ ИНФОРМАЦИИ

Волков В.А., студент, V курс
Научный руководитель: Спирин А.А., доцент
Федеральное государственное казённое военное образовательное учреждение высшего образования, подведомственное Федеральной службе охраны РФ «Академия ФСО РФ»
г. Орёл

Статья посвящена разработке программного модуля обнаружения актуальных источников информации в сети интернет. Предложенный алгоритм работы разработан на основе анализа существующих методов сбора, обработки и визуализации информации. Для разрабатываемого метода были сформулированы требования, предложен алгоритм работы модуля, оценена сложность алгоритма.

Ключевые слова: актуальные источники информации, автоматизированный сбор информации, программный модуль.

Введение

В настоящее время с постоянным развитием информационных технологий и Интернета доступ к информации стал безграничным и быстрым. Однако, огромное количество данных и ресурсов, доступных в сети, создает сложности при поиске актуальных источников информации. Поиск актуальных источников информации в современном информационном обществе является одним из важнейших аспектов работы исследователей, журналистов и многих других профессионалов. Необходимость оперативного и точного доступа к актуальным данным ставит перед аналитиками ряд сложных задач, связанных с определением источников, которые могут быть доверены и использованы для получения информации. [9]

Вот основные проблемы, с которыми сталкиваются пользователи при поиске актуальных источников:

– Информационный шум: с ростом объема информации в сети интернет возрастает и уровень "информационного шума". Поисковые системы могут возвращать множество результатов, несоответствующих потребностям пользователя.

– Устаревание информации: некоторые данные быстро устаревают. Поиск актуальных источников становится особенно сложным в случаях, когда необходимо получить информацию о текущих событиях или темах.

– Достоверность и авторитетность: существует множество неофициальных и ненадежных источников, что может привести к ошибкам в анализе и принятии решений. [2] [10]

Главной проблемой, с которой сталкиваются аналитики при поиске актуальных источников информации, является перегрузка информацией. В современном мире доступ к огромному объему данных и источников может привести к затруднениям в выделении релевантной и точной информации. Аналитики должны разрабатывать стратегии и использовать инструменты, которые позволяют фильтровать и извлекать наиболее актуальные и доверенные источники среди массы доступной информации, чтобы принимать информированные решения и проводить анализ на основе качественных данных. [5]

Целью исследования является выдвижение предложений по созданию автоматизированной системы поиска актуальных источников информации и разработке возможного алгоритма её работы с использованием современных методов и технологий.

Для достижения цели необходимо решить следующие задачи:

– Проанализировать существующие методы и подходы к поиску актуальных источников информации в сети интернет.

– Разработать алгоритм и модель, позволяющие оценивать актуальность источников информации на основе временных данных .

– Оценить применимость и перспективы использования автоматизированной системы в различных областях.

Цель и задачи исследования направлены на создание инновационной системы, способной улучшить процесс поиска актуальных данных, что имеет важное значение для различных областей деятельности, где актуальность и достоверность информации играют ключевую роль.

Обоснование актуальности

Сегодня время – один из самых важных ресурсов человека. Из-за обширной базы информационных источников информации данные системы зачастую могут обрабатывать запрос продолжительное количество времени, из-за чего процесс поиска информации может сильно увеличиться. Также существует проблема в постоянном росте и изменении источников информационных источников. На рисунке 1 представлена диаграмма зарегистрированных СМИ за последние 25 лет. [6]

Рисунок 1. Динамика увеличения количества, зарегистрированных СМИ.

По графику можно понять, что ситуация в социальных медиа начинает, развивается ещё более стремительно, популярность новостных пабликов на различных интернет-площадках возрастает ежедневно (На рисунке 2 показан рейтинг интернет-площадок России в 2023 г.), что приводит к созданию к ежедневному приросту количества каналов и пабликов.

Рисунок 2. Рейтинг интернет-площадок России в 2023 г.

Таким образом можно подвести итог, что в эпоху информационного перенасыщения, актуальность создания модуля поиска актуальных источников информации оказывается весьма значительной. С увеличением объема данных становится сложнее отличить достоверные и актуальные сведения от устаревших или искаженных. Широкое распространение дезинформации делают неотложным создание инструмента, способного фильтровать и предоставлять пользователям достоверные и авторитетные данные.

Борьба с дезинформацией и обеспечение легкого доступа к правдивой информации становятся ключевыми задачами в современном мире. Разработка подобного модуля не только соответствует вызовам современности, но и отвечает потребностям пользователей в надежных источниках данных, способствуя повышению общего уровня информационной грамотности и качества обмена знаниями.

Формирование требований к системе

Для решения задачи обнаружения актуальных источников информации в сети интернет был сформулирован ряд требований.

1. Функциональные требования:

1.1. Модуль должен иметь возможность собирать данные из различных открытых источников (новостные сайты, блоги, социальные сети и т.д.).

1.2. Модуль должен обеспечивать анализ и классификацию полученных данных на основе их актуальности и достоверности.

1.3. Модуль должен предоставлять фильтрацию источников по различным параметрам, таким как тематика, регион, авторитетность и др.

2. Производительность:

2.1. Модуль должен иметь высокую скорость обработки данных источников для быстрого доступа к актуальной информации.

2.2. Модуль должен быть оптимизирован для работы с большими объемами данных.

3. Интерфейс и удобство использования:

3.1. Модуль должен иметь интуитивно понятный интерфейс, удобный для пользователей.

3.2. Модуль должен обеспечивать возможность выбора и настройки фильтров для поиска актуальных источников.

4. Масштабируемость:

4.1. Модуль должен быть способен обрабатывать и хранить большое количество источников информации и обновлять их в режиме реального времени.

Проектные решения по функциональной структуре

Для алгоритма автоматизированного обнаружения актуальных источников информации можно выделить три основных этапа функционирования:

1. Сбор информации (наполнение базы данных существующими источниками информации для дальнейшего мониторинга и оценки актуальности публикаций). Для сбора информации в процессе сбора источников информации в социальных сетях, информационно-новостных порталов, размещённых в сети интернет, используются парсеры и ПО для мониторинга появления новых источников информации. [1]

2. Анализ информационных источников информации (оценка характеристик источника информации и публикаций, на основе которых строится рейтинг актуальных источников информации). Все собранные источники информации проходят через аналитическую составляющую системы, в которой составляется отчёт по заранее определённым показателям присущим источникам информации и публикациям. По результатам отчёта происходит присвоение весовых значений для каждого из показателей. Оцениваемые показатели источников информации и критерии их оценивания представлены в таблице 1.

Таблица 1

Показатели источников	Значение весового показателя						
	0	0,1	0,3	0,5	0,7	0,9	1
Количество подписчиков источника	Менее 200	200 – 10 тыс.	10 тыс. – 100 тыс.	100 тыс. – 500 тыс.	500 тыс – 1 млн..	1 мнн. – 3 млн.	Более 3 млн.
Средний охват аудитории на одной публикации	Менее 500	500 – 50 тыс.	50 тыс. – 300 тыс.	300 тыс. – 1мнн.	1 млн. – 3 млн	3 млн – 5 млн.	Более 5 млн.

Продолжение таблицы 1

Среднее количество публикуемых сообщений за сутки	Менее 1	1 – 1,5	1,5 – 3	3 – 5	5 – 10	10 – 15	Более 15
Официальность источника	Нет	-	-	-	-	-	Да
Ежемесячный прирост читателей	Менее 50	50 – 500	500 – 1 тыс.	1 тыс. – 5 тыс.	5 тыс. – 10 тыс.	10 тыс. – 20 тыс.	Более 20 тыс.
Длительность существования источника информации	Менее месяца	1 мес. – 6 мес.	6 мес. – 1 год	1 год – 2 года	2 года – 4 года	4 года – 10 лет	Более 10 лет

Данные о рассчитанных весовых показателях каждого информационного источника информации сохраняются в отдельный раздел базы данных.

3. Составление рейтинга источников информации по актуальности, визуализация данных и вывод результатов в пользовательском интерфейсе системы. Данный процесс, на основе сохранённых в базу данных весовых показателей информационных источников, выстраивает рейтинг от наиболее предпочтительных до менее предпочтительных информационных источников и выводит данные в виде списка. Также данный процесс позволяет более подробно изучить полученную на предыдущем этапе статистику в виде графиков изменения показателей каждого из источников информации.

Разработка алгоритма

Для разработки модуля обнаружения актуальных источников информации предлагается использовать следующие методы и подходы:

1. Сбор и анализ данных из различных источников – модуль должен иметь возможность сканировать и анализировать различные открытые источники.

2. Машинное обучение и анализ текста – модуль должен быть способен анализировать и классифицировать информацию на основе ее контекста и релевантности.[8]

3. Создание рейтинговой системы – модуль должен иметь возможность создавать рейтинговую систему для оценки актуальности и авторитетности источников информации. [3][4]

На рисунке ниже представлен общий алгоритм работы предлагаемого прототипа автоматизированной системы обнаружения актуальных источников информации.

Работа приложения начинается с запуска клиентского приложения.

В первую очередь для оценки актуальности источников информации необходимо собрать набор данных, который предстоит в последующем оценивать. В нашем случае на первом этапе работы приложения, необходимо сформировать базу данных из источников информации для оценки их параметров и выгрузки публикуемых сообщений. На графике процесс, называется «Обновление источников информации» обозначен как A1.

Второй этап работы алгоритма является основным в системе, на этом этапе происходит непосредственно оценка актуальности источника информации. Для этого информационная система выполняет ряд последовательных действий. В начале происходит сбор характеристик источника информации таких как количество подписчиков источника, средний охват аудитории на одной публикации в течении недели, количество публикуемых сообщений за сутки, проверка на официальность источника, ежемесячный прирост читателей, длительность существования источника информации. Каждая из этих характеристик будет оцениваться по шкале от нуля до единицы. После чего результат оценки будет сохранён в БД и по окончанию оценки системой всех информационных источников, будет составлен рейтинг актуальных источников информации. На графике это процесс A2.

На следующем этапе программа определяет наличие БД новостей в своем составе. Если такой базы данных нет, то программа автоматически генерирует файл в формате db, содержащий библиотеку БД SQLite. В который далее будут сохраняться для дальнейшего отображения публикаций источников, имеющихся в системе. После этого программа начинает процесс сбора сообщений(A3).

Если запуск программы происходит не в первый раз и уже имеется файл базы данных SQLite, то приложение определяет время последнего сообщения, рассчитывает период, за который необходимо произвести сбор сообщений, и переходит к этапу выгрузки сообщений (процесс A3).

После выгрузки сообщений происходит сохранение полученных результатов и демонстрация их пользователю на экране в рабочем поле программы. На этом этапе пользователь получает наиболее соответствующую текущему состоянию дел информацию.

При желании пользователь в самом приложении может ознакомиться с аналитическими данными, собранными в ходе работы алгоритма о каждом источнике информации, находящемся в БД, где будут отображаться изменение характеристик канала во времени.

На этом этапе работа с программой завершается.

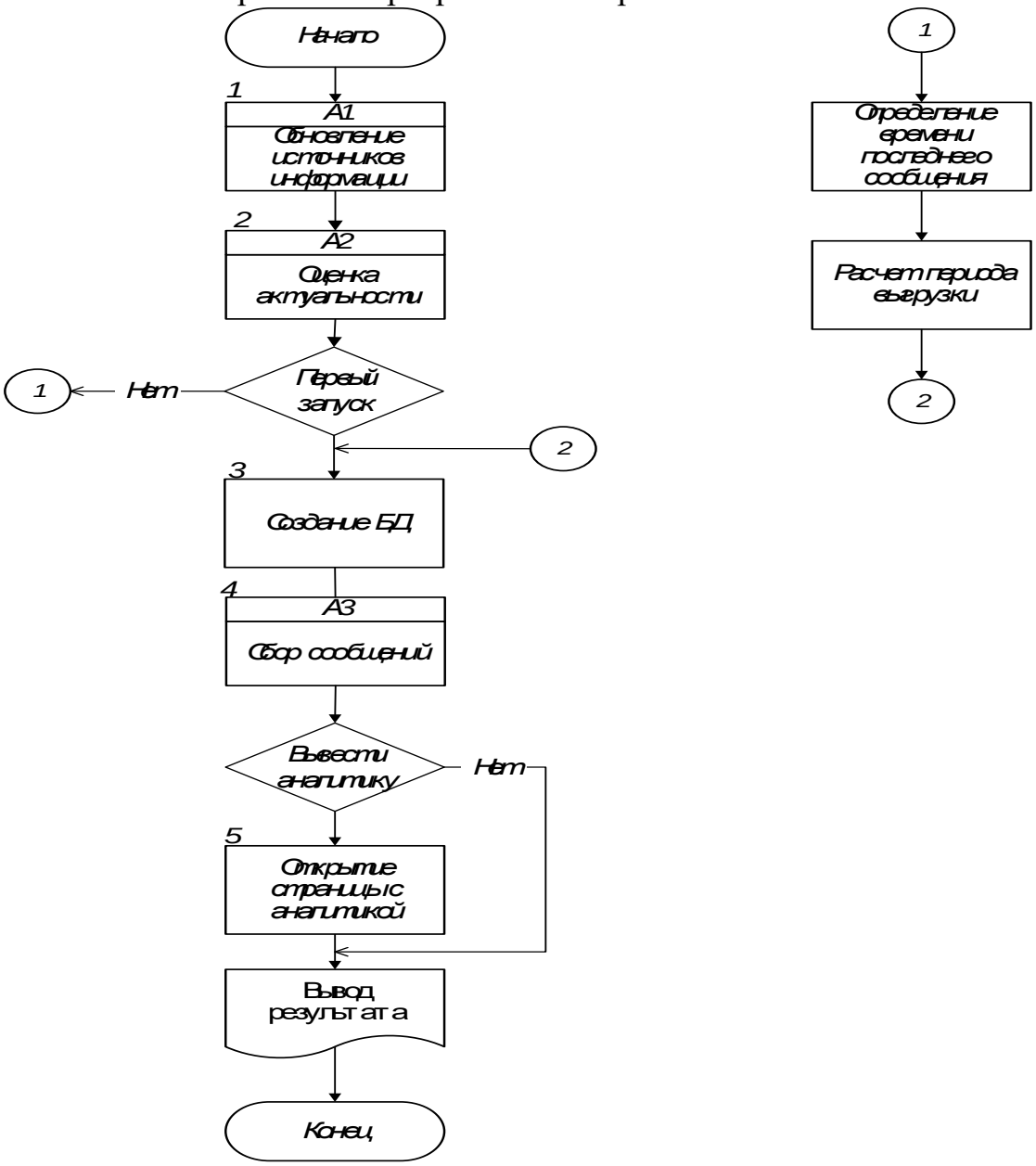

Рисунок 3. Общий алгоритм работы программы

Оценка сложности алгоритма

Под сложностью алгоритма подразумевается общее время выполнения данного алгоритма. Для расчета сложности алгоритма необходимо вычислить число шагов как конкретных процедур, так и всего алгоритма в целом. Обозначим переменной q источники информации, загружаемые автоматизированной системой. [7] [8]

Тогда для общего алгоритма работы приложения получим:

Таблица 2

№	Этап	Время исполнения	количеств	$T_{общее}$

		конкретного этапа	оитераций	
1	Процедура A1	T_1	n	$T_1 n$
2	Процедура A2	T_2	n	$T_2 n$
3	Проверка на первый запуск	T_3	n	$T_3 n$
4	Создание БД	T_4	n	$T_4 n$

Продолжение таблицы 2

5	Определение времени последнего обновления	T_5	n	$T_5 n$
6	Расчет периода выгрузки	T_6	n	$T_6 n$
7	Процедура A3	T_7	n	$T_7 n$
8	Вывод аналитических данных	T_8	n	$T_8 n$
9	Вывод результатов	T_9	n	$T_9 n$

Общая временная сложность алгоритма составит:

$$Q = \sum_1^9 \mathrm{T} * n \quad (1)$$

Однако, в ходе выполнения алгоритма наиболее ресурсоемким процессом являются процедуры A1, A2 и A3. Остальные этапы сравнительно малы и могут не учитываться. Поэтому сложность алгоритма равна:

$$Q(t) = T_1 + T_2 + T_7 \quad (2)$$

Список литературы

1. Шершенёв Н.О. — web-scraping как инструмент извлечения полезных данных из сети интернет [Электронный ресурс] / Промышленность. Информационные технологии – 2020, с. 137-138. Режим доступа: https://elib.psu.by/bitstream/123456789/35396/1/137-138.pdf

2. Замятин А.В. — Интеллектуальный анализ данных / Издательский Дом Томского государственного университета – 2020, с. 7-27.

3. Ющук Е.Л. — OSINT: что это, кому он нужен, какие методы сбора и типы информации использует? [Электронный ресурс] – 2020. – Режим доступа: https://yushchuk.livejournal.com/1451268.html

4. Меньшиков Я.С. — преимущества автоматического сбора данных в сети интернет над ручным сбором данных [Электронный ресурс] / Universum: технические науки: электрон. научн. журн. 2022. 10(103). Режим доступа: https://cyberleninka.ru/article/n/preimuschestva-avtomaticheskogo-sbora-dannyh-v-seti-internet-nad-ruchnym-sborom-dannyh/viewer

5. Vera Granikov, Reem El Sherif, France Bouthillier, Pierre Pluye — Factors and outcomes of collaborative information seeking: A mixed studies review with a framework synthesis [Электронный ресурс] / Journal of the Association for Information Science and Technology Volume 73, Issue 4 p. 542-560, 25 October 2021. Режим доступа: https://asistdl.onlinelibrary.wiley.com/

6. Левашов А.Н. — СМИ в России [Электронный ресурс] / Tadviser 2023.11.20. Режим доступа: https://www.tadviser.ru/index.php/

7. Керимов В.А., Гаджиев Ф.Г. — О методах оценки сложности алгоритмов / Вестник науки №4 (61) том 1, , С. 299 - 304. 2023 г.

8. Никитин П.В., Горохова Р.И., Бахтина Е.Ю., Долгов В.И., Коровин Д.И. — Алгоритмы извлечения информации из проблемно-ориентированных текстов / Вопросы безопасности. — 2023. - № 3. - С.1-10.

9. Алексеева Е.А. — Современная информационная среда: особенности работы с источниками информации / СМИ и массовые коммуникации — 2019.