УДК 004.042

МЕТОД СБОРА ИНФОРМАЦИИ НА ОСНОВЕ РАЗБИЕНИЯ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА ИСТОЧНИКА ПО ЧАСТОТЕ ОБНОВЛЕНИЯ РУБРИК

Мордасов В.А. [1], **Фастов Д.С.** [1], **Кученов Е.Н.** [1]
Научный руководитель – кандидат технических наук, доцент Платонов А.А. [1]
[1]Военно-космическая академия имени А.Ф.Можайского
vka@mil.ru

Введение.

Современные системы мониторинга, агрегирующие данные из множества открытых источников, сталкиваются с проблемой гетерогенности систем рубрикации. Анализ гетерогенных данных, основанных на различных правилах, представляет значительную сложность для информационных систем [1]. Каждый источник, будь то Bloomberg с финансовой таксономией или региональное СМИ, классифицирует контент по собственным правилам.

Анализ опыта платформ Meltwater, LexisNexis [3] и отечественных практик показывает, что простое слияние рубрикаторов ведет к информационному хаосу. Следствием становится комплекс проблем: потеря релевантности, семантическая избыточность, дублирование данных и невозможность построения корректных аналитических панелей. Также возникает необходимость «ручной переклассификации» и создаются препятствия для внедрения методов машинного обучения [4].

Таким образом, существует объективная научно-практическая потребность в разработке методов сбора, способных нивелировать разницу в рубрикаторах источников на этапе извлечения данных и минимизировать избыточность информационных потоков.

Основная часть

В работе предлагается оригинальный метод сбора информации, основанный на анализе частоты обновления рубрик внутри каждого информационного источника. В отличие от традиционных подходов, предполагающих равнозначную обработку всех разделов веб-ресурса, предлагаемый метод базируется на динамической приоритезации рубрик в зависимости от их реальной активности [2].

Ключевая гипотеза состоит в том, что распределение контента неравномерно. Рубрики делятся на четыре категории: часто (ЧОР, до 60 мин), средне (СОР, до 3600 мин), мало (МОР, свыше 3600 мин) и практически не обновляемые (НОР). Анализ показывает, что новости в ЧОР и СОР составляют лишь 15% от общего количества, тогда как в МОР и НОР сосредоточено 85% контента. Это означает, что основной объем информации генерируется небольшой группой рубрик

Наиболее рациональной стратегией сбора является концентрация усилий на часто и средне обновляемых рубриках, тогда как мало и не обновляемые разделы целесообразно обрабатывать по отдельному, значительно более редкому графику либо исключать из регулярного мониторинга. Такой подход позволяет сместить фокус вычислительных мощностей на действительно актуальный контент, минимизируя «холостые запросы» и дублирование данных, которое возникает из-за семантических пересечений между рубриками.

Для формализации задачи оптимизации вводится показатель оптимальности сбора E_i для i-источника, учитывающий три ключевых фактора: скорость сбора уникальных новостей (производительность), долю уникального контента в общем потоке (коэффициент уникальности) и сложность обработки, обратно пропорциональную числу отслеживаемых рубрик. Чем выше значение показателя, тем

эффективнее организован процесс сбора. С учетом весовых коэффициентов рубрик (ω_a, где 1 присваивается ЧОР, а 4 — НОР) показатель трансформируется таким образом, что основная задача сводится к минимизации суммарного веса отслеживаемых рубрик.

Предлагаемый метод включает следующие последовательные этапы:

1. **Первичный сбор данных** по всем рубрикам источника без исключения (длительность 1-3 суток) для накопления статистической информации.
2. **Классификация рубрик** на основе анализа временных рядов публикаций с отнесением каждой рубрики к категориям ЧОР, СОР, МОР, НОР и присвоением соответствующих весовых коэффициентов.
3. **Формирование рационального множества рубрик** для регулярного мониторинга (например, все ЧОР + оптимальная часть СОР).
4. **Итеративный сбор данных** с выбранным множеством рубрик и фиксация показателей эффективности.
5. **Расчет показателя оптимальности** и сравнение полученных значений для различных вариантов множеств с выбором наилучшего варианта, обеспечивающего максимальное значение E_{ω}^{i}.
6. **Переход в рабочий режим** с использованием оптимального набора рубрик.

Экспериментальная апробация метода подтвердила, что предложенный подход позволяет эффективно минимизировать дублирование, сократить общее время выполнения сбора и максимизировать охват уникального контента за счет поиска «точки насыщения» — момента, когда подключение новых рубрик перестает давать значимый прирост уникальной информации.

Выводы.

Предложенный метод сбора информации на основе разбиения информационного пространства источника по частоте обновления рубрик позволяет формализовать задачу оптимизации информационных потоков в условиях пересекающихся рубрикаторов. Использование классификации рубрик (ЧОР, СОР, МОР, НОР) и введенного показателя оптимальности сбора *E* дает возможность на практике выбрать обоснованный баланс между полнотой охвата информации и рациональным использованием вычислительных ресурсов.

Литература

1. Питькевич П.И. Методы агрегирования, уменьшения размера и обработки больших данных // Инженерный вестник Дона. — 2025. — № 3. — С. 45-52.
2. Платонов А.А., Мордасов В.А., Кученов Е.Н., Фастов Д.С. Автоматизированный сбор и анализ информации о событиях из открытых источников на основе тематических рубрикаторов // Методы и технические средства обеспечения безопасности информации: материалы 34-й научно-технической конференции, г. Санкт-Петербург, 2025 г. – СПб.: СПбГУ, 2025. – С. 143–145
3. Шилов М.А. Методы и модели мониторинга информации из открытых источников для бизнес-аналитики: магистерская диссертация. — Москва: НИУ ВШЭ, 2025. — 89 с.
4. TDWI Research. Best Practices for Data Integration in the Age of Big Data and AI / TDWI Research. — Renton, WA: The Data Warehousing Institute, 2023. — 35 с. — URL: https://tdwi.org/research/list/tdwi-best-practices-reports.aspx (дата обращения: 12.02.2026).