Автоматическая служба новостей – идеи, проблемы, решения

Тезисы доклада на конференции
«Прикладная лингвистика и искусственный интеллект 2012»

Антонов Александр, Баглей Станислав, Ландэ Дмитрий
alexa@galaktika.ru, baglei@galaktika.ru, dwl@visti.net

Сегодня в веб-пространстве существует большое количество онлайн-СМИ, социальных медиа, содержащих материалы новостного характера, интеграторов новостей.

Именно интеграторы новостей обеспечивают возможность доступа пользователей к материалам не всегда популярных веб-сайтов, которые иногда публикуют важную региональную или тематическую информацию. Кроме того интеграторы новостей выполняют функции «обратной связи», их информация перепечатывается (в худшем случае в режиме копипаста), обрабатывается аналитиками, обобщается, учитывается и, в конечном итоге, публикуется на аналитических веб-ресурсах, в виде аналитических отчетов и т.д.

Службы интеграции новостей сегодня известны всем, это Google News, Яндекс Новости, Рамблер Новости, Новотека, Мета Новини, UAport.net, webground.su …

При этом большинство названных служб интеграции новостей работают практически в автоматическом режиме, подчеркнем, не автоматизированном, с участием человека-оператора, а именно автоматическом. И хорошо, когда эти автоматы работают для людей, а не только для других автоматов, которые насчитывают деньги за «бесполезную» рекламу.

Автоматические интеграторы новостей зачастую стыкуются с информационно-аналитическими системами самого разного назначения, которые обеспечивают возможности контент-анализа текстов, извлечения информации, знаний. К таким системам, могут быть отнесены, например, Галактика Zoom, Медиалогия, Интегрум, InfoStream…

Именно проблемы, которые возникают при построении автоматических интеграторов новостей, пути их решения и некоторые идеи, возникающие при построении интеграторов и информационно-аналитических систем – предмет обсуждения в рамках данного доклада.

В предлагаемой таблице приведен краткий обзор идей, связанных с ними проблем и решений (IPS – Idea-Problem-Solutions).

IPS	Идея	Проблема	Решения
01	Охват данных в различных форматах	Неоднородность средств представления в Интернете информации различной структуры, в различных форматах	1. Введение ограничений (напр., охват только RSS) 2. Реализация метаязыков охвата любых текстовых форматов 3. Разработка/подключение конверторов из различных форматов 4. Распознавание графических изображений 5. Распознавание мультимедиа (звук → текст + признаки)
02	Охват наибольшего количества необходимых источников. «Полнота»	Необходимость соблюдения авторских и смежных прав, этических норм и т.п.	1. Использование новостей, не защищаемых законами об авторском праве. 2. Использование права «по умолчанию», зафиксированного на страницах ресурсов. 3. Заключение договоров о сотрудничестве с источниками. 4. Покупка информации с правами распространения.
03	Охват наибольшего количества необходимых источников. «Точность»	Отбор качественных и оригинальных источников	1. Многопараметрическое ранжирование источников, вычисление значений репутации 1.1. Цитируемость 1.2. Продуктивность 1.3. Периодичность 1.4. Популярность 1.5. Оригинальность 2. Краудсорсинг для отбора источников
04	Гибкость работы с контентом	Изменение форм представления данных на ресурсах-источниках	1. Не учитывать изменений, как в большинстве глобальных поисковиков. 2. Прямые договора с поставщиками с утверждением форматов, периодичности и т.п. 3. Создание комплексов мониторинга за состоянием источников. 4. Интеллектуальные автоматически настраиваемые парсеры.
05	Синхронизация интегратора с источниками	Корректность ссылок на источники. Удаление информации с источников, переименование	1. Не учитывать изменений. 2. Учет времени жизни публикаций на источниках при их включении в систему. 3. Мониторинг доступности отдельных документов 4. Создание комплексов мониторинга за состоянием источников.

06	Оптимизация работы роботов	Объем трафика роботов интеграторов	1. «Прозрачный» язык описания сценария работы робота. 2. Защита от зацикливания и др. возможных перегрузок. 3. Автоматизированная синхронизация времени сканирования с временем обновления источника. 4. Использование файлов типа sitemap.xml.
07	Юзабилити, улучшение навигации	Необходимость использования строки поиска, ввода неизвестных критериев поиска	1. RSS. 2. Карта сайта. 3. Иерархическая классификация документов и источников. 4. Кластеризация, выявление центроидов и новых рубрик. 5. Перевод в архивы наименее запрашиваемых (и наоборот – вывод из архива актуальных). 6. Отображение кластеров сниппетами из разных источников. 7. Автоматический сбор подкаста или видеовыпуска новостей из фрагментов.
08	Улучшение индексирования интегратора поисковыми системами	Перемещение части информации интегратора в категорию «скрытого веб»	1. RSS 2. Карта сайта 3. Другие вышеназванные средства улучшения навигации по веб-сайту интегратора.
09	Персонализация	«Универсальная» информация для всех категорий пользователей	1. Автоматическое формирование профиля по признакам → предсказание информационного интереса по текущей активности 2. Формирование страниц в зависимости от профиля (поискового запроса) 3. Общий аккаунт с другими сервисами. 4. Организация обратной связи, в т. ч. с социальными сетями
10	Аналитика	Отсутствие инструментов для формирования нового знания	1. Определение тенденций 2. Определение связанных источников 3. Определение тональности 4. Выделение сущностей 5. Построение семантических сетей 6. “Прогнозирование новостей” на некоторый временной горизонт
11	Выявление новых сюжетов	Традиционные технологии построения сюжетов дают информацию об уже всем известных событиях	1. Выявление аномальных сообщений в рейтинговых источниках 2. Резкое изменение превалирующей лексики 3. «Взрывное» появление дубликатов

12	Работа с данными на разных языках	Неполнота охватываемой информации	1. Развитие технологий автоматического потокового перевода 2. Выявление дубликатов и близких по смыслу документов на разных языках. 3. Учет дубликатов и подобия при построении аналитических отчетов.
13	Визуализация результатов	Потеря полноты охвата/ точности при выборочной визуализации	1. Java, флеш-технологии, HTML5 2. Построение удобных интерфейсов между средствами визуализации и аналитическими модулями. 3. Миграция на мобильные устройства, автомобильные и lcd-панели и т.п.
14	Только релевантная реклама	Уход от тематики, реклама не для людей	1. Классификация рекламы в соответствии с классификацией ресурсов . 2. Взаимодействие с надежными рекламными службами. 3. Целевая продажа тематической медийной рекламы.