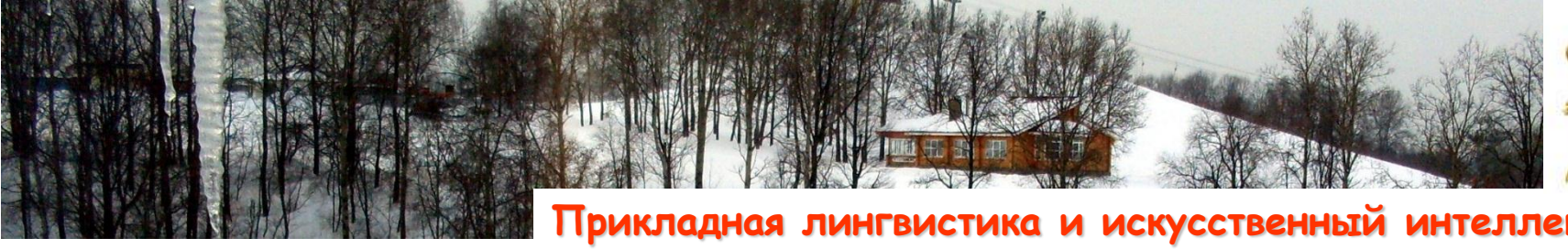


Автоматическая служба новостей – идеи, проблемы, решения

Александр АНТОНОВ, Станислав БАГЛЕЙ, Дмитрий ЛАНДЭ

**Корпорация «ГАЛАКТИКА», Москва,
Информационный центр «ЭЛВИСТИ», Киев,
Проект WebGround.su**

Москва-2012

Интеграторы новостей обеспечивают возможность доступа пользователей к материалам не всегда популярных веб-сайтов, которые иногда публикуют важную региональную или тематическую информацию.

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.

IPS 01.Охват данных в различных форматах

Проблема: Неоднородность средств представления в Интернете информации различной структуры, в различных форматах

Возможные решения:

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)…

IPS 02.Охват наибольшего количества необходимых источников. «Полнота»

Проблема: Необходимость соблюдения авторских и смежных прав, этических норм и т.п.

Возможные решения:

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...

IPS 03.Охват наибольшего количества необходимых источников. «Точность»

Проблема: Отбор качественных и оригинальных источников

Возможные решения:

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

IPS 04. Гибкость работы с контентом

Проблема: Изменение форм представления данных на ресурсах-источниках

Возможные решения:

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

IPS 05. Синхронизация интегратора с источниками

Проблема: Корректность ссылок на источники. Удаление информации с источников, переименование

Возможные решения:

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.

IPS 06. Оптимизация работы роботов

Проблема: объем трафика роботов интеграторов

Возможные решения:

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.

IPS 07. Юзабилити, улучшение навигации

Проблема: необходимость использования строки поиска, ввода неизвестных критериев поиска

Возможные решения:

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

IPS 08. Улучшение индексирования интегратора поисковыми системами

Проблема: перемещение части информации интегратора в категорию «скрытого веб»

Возможные решения:

1. RSS
2. Карта сайта
3. Другие вышеназванные средства улучшения навигации по веб-сайту интегратора.

IPS 09. Персонализация

Проблема: «Универсальная» информация для всех категорий пользователей

Возможные решения:

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

IPS 10. Аналитика

Проблема: Отсутствие инструментов для формирования нового знания

Возможные решения:

1. Определение тенденций
2. Определение связанных источников
3. Определение тональности
4. Выделение сущностей
5. Построение семантических сетей
6. "Прогнозирование новостей" на некоторый временной горизонт

IPS 11. Выявление новых сюжетов

Проблема: Традиционные технологии построения сюжетов дают информацию об уже всем известных событиях

Возможные решения:

1. Выявление аномальных сообщений в рейтинговых источниках
2. Резкое изменение превалирующей лексики
3. «Взрывное» появление дубликатов

IPS 12. Работа с данными на разных языках

Проблема: Неполнота охватываемой информации

Возможные решения:

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.

IPS 13. Визуализация результатов

Проблема: потеря полноты охвата/ точности при выборочной визуализации

Возможные решения:

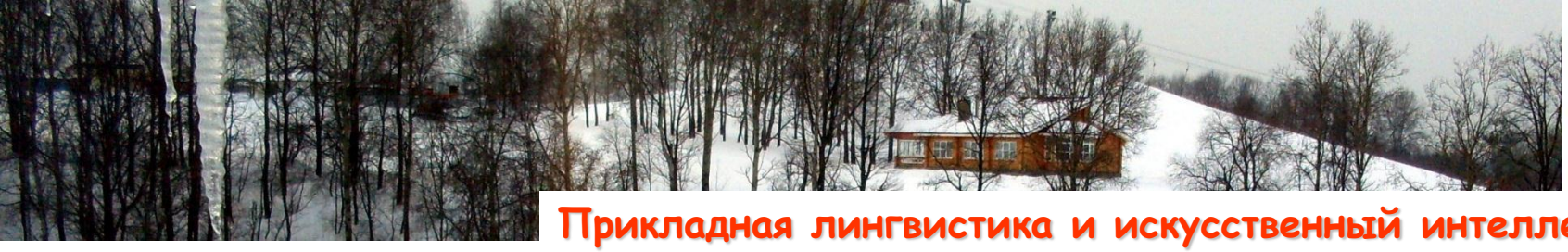
1. Java, флеш-технологии, HTML5
2. Построение удобных интерфейсов между средствами визуализации и аналитическими модулями.
3. Миграция на мобильные устройства, автомобильные и lcd-панели и т.п.

IPS 14. Только релевантная реклама

Проблема: уход от тематики, реклама не для людей

Возможные решения:

1. Классификация рекламы в соответствии с классификацией ресурсов
2. Взаимодействие с надежными рекламными службами
3. Целевая продажа тематической медийной рекламы



Спасибо за внимание!

**Александр АНТОНОВ, Станислав БАГЛЕЙ,
Дмитрий ЛАНДЭ**

http://WebGround.su