Особенности сбора данных в LiveJournal

Санкт-Петербург 2012

Цель работы:

- 1. Обеспечить полуавтоматический сбор данных из Интернета (посты с авторами и комментарии к ним с авторами этих комментариев).
- 2. Сформировать реляционную базу для хранения данных из интернета и интерфейс к БД с функциями.
- 2.1 Закачка данных из Интернета в базу.
- 2.2. Возможность формировать запросы и получать соответствующие отчеты.
- 2.3. Выгрузка результатов запросов в виде специфичных форматов, пригодных для обработки результатов выгрузки в различных сторонних ПО.

Проблемы получения данных

В документации ЖЖ было обнаружено средство, предусматривающее получение информации по выбранному пользователю и его постам. Это так называемый RSS запрос. Однако в ходе экспериментов было обнаружено, что данный запрос выдает только 25 последних постов.

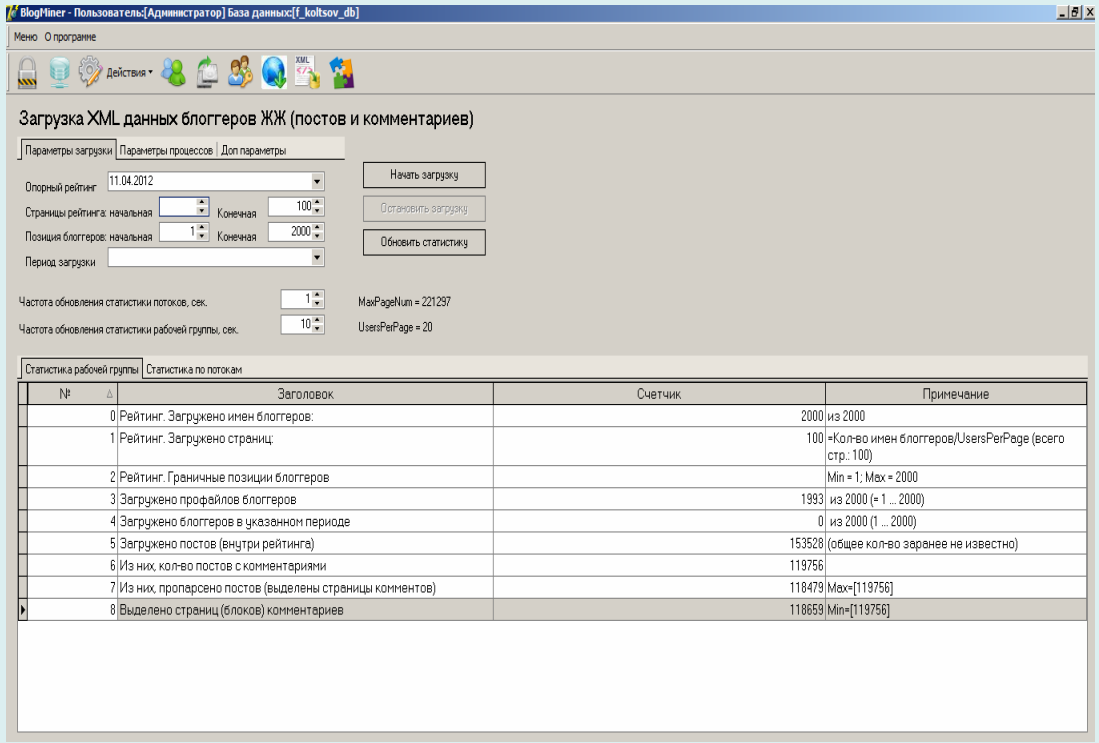
На основе экспериментов с сервером ЖЖ была найдена недокументированная функция получения дерева комментариев к любому посту пользователя.

Общий алгоритм загрузки данных из LiveJournal

1. Закачка данных пользователей, их постов и комментариев к постам основана на недокументированной XML процедуре.
2. Список закачиваемых пользователей берется из рейтинга пользователей ЖЖ.
3. Все вычисления и парсинг данных производятся offline.

Процедуры закачки:

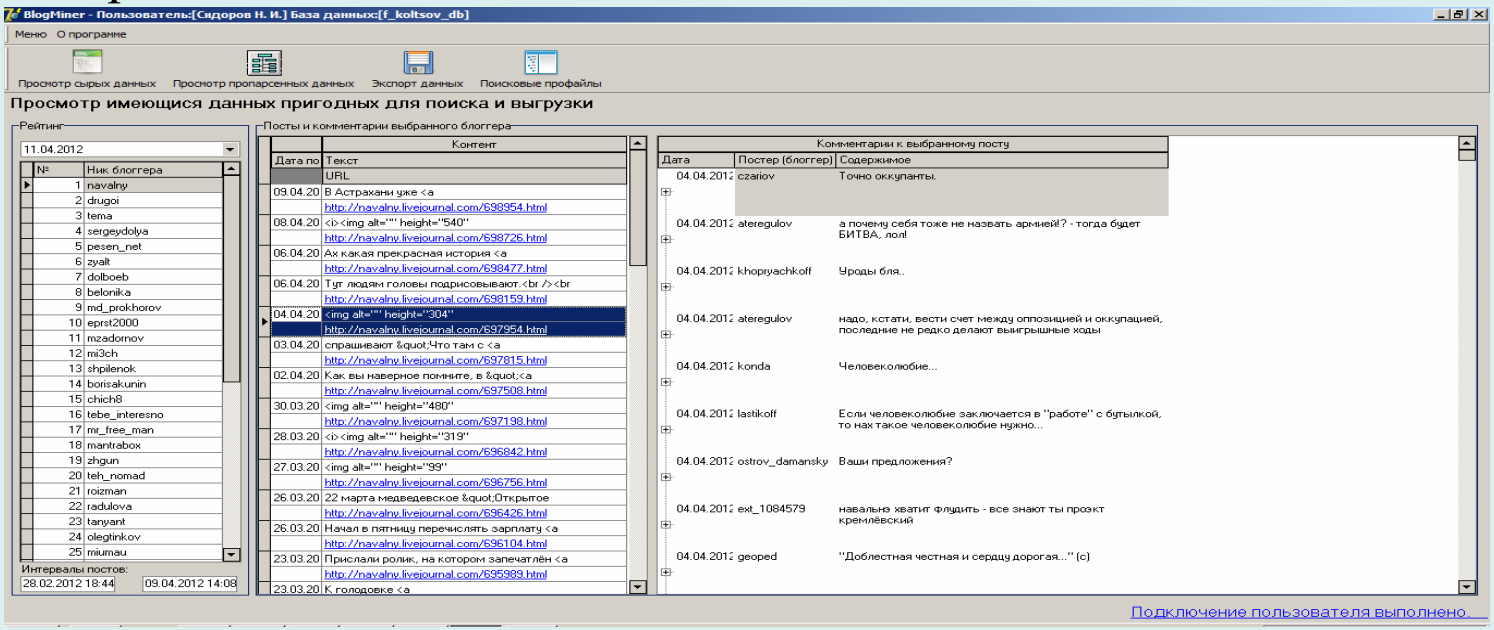
- Чтение страниц рейтинга,
- Парсинг имен пользователей LiveJournal
- Получение FOAF (информации о пользователях и отсечение групп – «Community»)
- Загрузка всех постов «пропарсенных» пользователей и комментариев к ним с помощью XMP-RPC интерфейса с сервером ЖЖ.



Парсинг закаченных данных

Загружаемые посты и комментарии представляют собой XML документы, поэтому необходимо выполнять извлечение (парсинг) текстов постов и комментариев из соответствующих XML документов, с сохранением взаимосвязей постов и комментариев.

Для хранения результатов парсинга созданы таблицы, позволяющие хранить результаты посты, комментарии и их взаимосвязь. Сохраняется дерево комментариев, когда другие блоггеры, дают комментарии на комментарии.

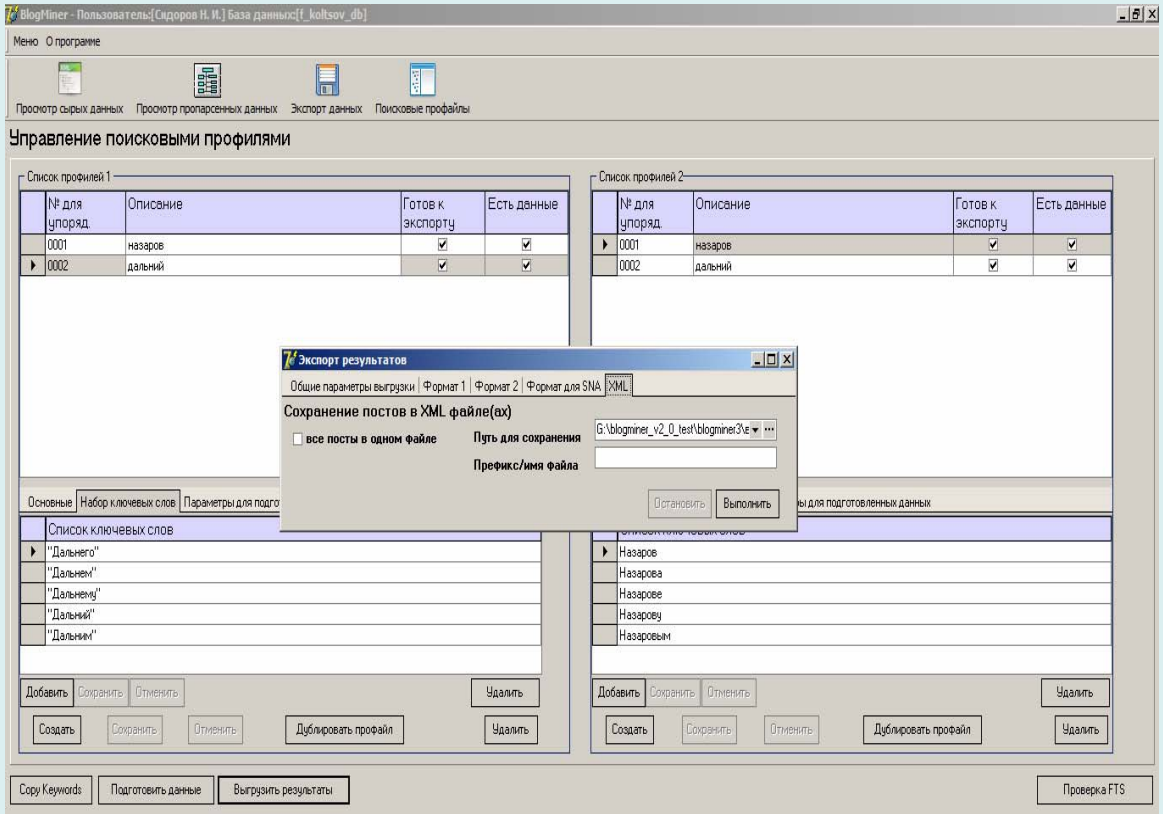


Эксопорт данных во внешние ПО

- Реализован экспорт данных по запросу, включающему ключевые слова
- Т.о. редуцируются данные, достигающие нескольких десятков килобайт.
- Поисковый инструмент: встроенный MS SQL механизм **F**ull **T**ext **S**earch **E**ngine (механизм полнотекстового поиска).

Виды экспорта (на основе FTSE).

1. **Stanford Topic Modelling Toolbox**. Один пост – один файл. Параметры для выгрузки: 1. период по дате. 2. количество постов. 3. наличие в посте ключевых слов. Прилагается файл с метаданными в формате CSV.
2. **GCLUTO**. Все посты в одном TXT файле.
3. **Выгрузка для сетевого анализа**. Файл 1: edge list. Файл txt, где в каждой строке пары обозначенных уникальными числами вершин.
Файл 2: список номеров вершин, соотнесенный с URL постов или именами комментаторов, формат txt.



Данные, используемые в работе:

В работе – топ 1400 блогеров:

1. 15.08-15.09.2011, 24 тыс. постов
2. 27.11-27.12.2011, 28 тыс. постов

Подготовлено – топ 2000 блогеров (после выборный период):

3. 04.03-10.04.2012, 119 тыс. постов, объем базы данных порядка 9Гб.