

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 XML-RPC интерфейса с сервером ЖЖ.

The screenshot shows the 'BlogMiner' application window. The main title is 'BlogMiner - Пользователь:[Администратор] база данных:[f_koltsov_db]'. The interface is in Russian and displays the 'Загрузка XML данных блоггеров ЖЖ (постов и комментариев)' (Loading XML data of bloggers from LiveJournal (posts and comments)) section.

Configuration parameters include:

- Опорный рейтинг: 11.04.2012
- Страницы рейтинга: начальная (1) - конечная (100)
- Позиция блоггеров: начальная (1) - конечная (2000)
- Период загрузки: [dropdown]
- Частота обновления статистики потоков, сек.: 1
- Частота обновления статистики рабочей группы, сек.: 10
- MaxPageNum = 221297
- UserPerPage = 20

Buttons: 'Начать загрузку', 'Остановить загрузку', 'Обновить статистику'.

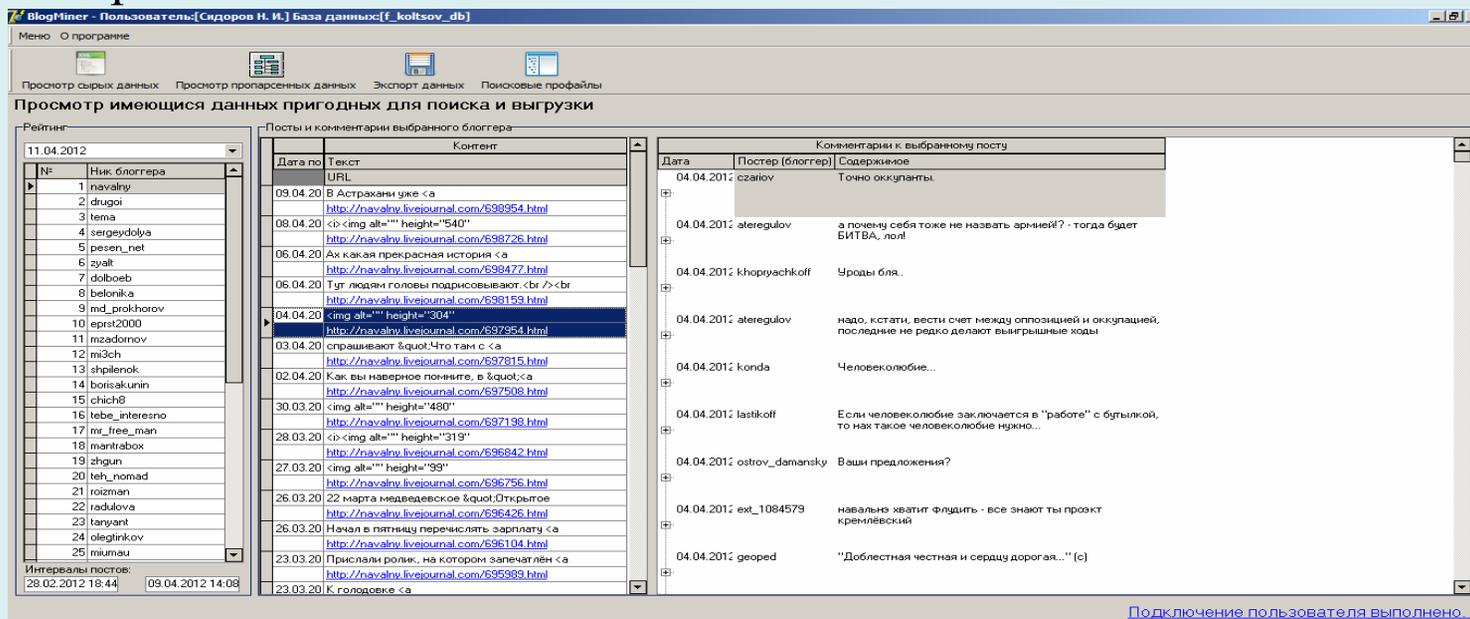
Below the configuration is a table titled 'Статистика рабочей группы' (Group statistics) with columns: №, Заголовок, Счетчик, and Примечание.

№	Заголовок	Счетчик	Примечание
0	Рейтинг. Загружено имен блоггеров:	2000	из 2000
1	Рейтинг. Загружено страниц:	100	«Кол-во имен блоггеров/UsersPerPage (всего стр.: 100)»
2	Рейтинг. Граничные позиции блоггеров		Min = 1; Max = 2000
3	Загружено профайлов блоггеров	1993	из 2000 (= 1 ... 2000)
4	Загружено блоггеров в указанном периоде	0	из 2000 (1 ... 2000)
5	Загружено постов (внутри рейтинга)	153528	(общее кол-во заранее не известно)
6	Из них, кол-во постов с комментариями	119756	
7	Из них, пропарсено постов (выделены страницы комментов)	118479	Max=[119756]
8	Выделено страниц (блоков) комментариев	118659	Min=[119756]

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.



BlogMiner - Пользователь: (Сидоров Н. И.) база данных: (f_koltsov_db)

Меню О программе

Просмотр сырых данных Просмотр пропарсенных данных Экспорт данных Поисквые профили

Просмотр имеющихся данных пригодных для поиска и выгрузки

Рейтинг: 11.04.2012

№	Ник блоггера
1	navalny
2	drugoi
3	tema
4	sergeydolya
5	pesen_net
6	zyalt
7	dolboeb
8	belonika
9	md_prokhorov
10	eprst2000
11	mzadornov
12	mi3ch
13	shpilenok
14	borisakunin
15	chich8
16	tebe_interesno
17	mr_free_man
18	mantrabox
19	zhgun
20	teh_nomad
21	roizman
22	reduleva
23	tanysant
24	oleglinkov
25	nikumau

Интервалы постов: 28.02.2012 18:44 09.04.2012 14:08

Посты и комментарии выбранного блоггера:

Дата по	Текст	Контент
09.04.20	В Астрахани уже <a	http://navalny.livejournal.com/598954.html
08.04.20	<img alt="" height="540"	http://navalny.livejournal.com/598726.html
06.04.20	Ах какая прекрасная история <a	http://navalny.livejournal.com/598477.html
06.04.20	Три льдины голыми подписываются <br	http://navalny.livejournal.com/598153.html
04.04.20	<img alt="" height="304"	http://navalny.livejournal.com/597954.html
03.04.20	спрашивают "Что там с <a	http://navalny.livejournal.com/597815.html
02.04.20	Как вы намерено поемте, в "<a	http://navalny.livejournal.com/597508.html
30.03.20	<img alt="" height="480"	http://navalny.livejournal.com/597198.html
28.03.20	<img alt="" height="319"	http://navalny.livejournal.com/596842.html
27.03.20	<img alt="" height="99"	http://navalny.livejournal.com/596756.html
26.03.20	22 марта медведовское "Открытие	http://navalny.livejournal.com/596426.html
26.03.20	Начал в пятницу перечислять зарплату <a	http://navalny.livejournal.com/596104.html
23.03.20	Прислали ролик, на котором запечатлен <a	http://navalny.livejournal.com/595993.html
23.03.20	К голодеке <a	

Комментарии к выбранному посту

Дата	Постер (блоггер)	Содержимое
04.04.2012	cszaiov	Точно оккупанты.
04.04.2012	aleregulov	а почему себя тоже не назвать армией? - тогда будет БИТВА, лол!
04.04.2012	khoryachkoff	Урады бя...
04.04.2012	aleregulov	надо, кстати, вести счет между оппозицией и оккупацией, последние не редко делают выигрышные ходы
04.04.2012	konda	Человеколюбие...
04.04.2012	lastik-off	Если человеколюбие заключается в "работе" с бутылкой, то как такое человеколюбие можно...
04.04.2012	ostrov_damansky	Ваши предложения?
04.04.2012	ext_1084579	навалюху хватит флидтить - все знают ты проклет кренлевский
04.04.2012	geored	"Доблестная честная и сердцу дорогая..." (с)

Подключение пользователя выполнено.

Экспорт 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 **Full Text Search Engine** (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.

The screenshot shows the BlogMiner application window with the following components:

- Menu:** Меню, О программе
- Navigation:** Просмотр сырых данных, Просмотр пропаренных данных, Экспорт данных, Поисковые профили
- Management:** Управление поисковыми профилями
- Profile Lists:**
 - Список профилей 1 (Table with columns: № для упоряд, Описание, Готов к экспорту, Есть данные)
 - Список профилей 2 (Table with columns: № для упоряд, Описание, Готов к экспорту, Есть данные)
- Export Dialog:** Экспорт результатов
 - Общие параметры выгрузки: Формат 1, Формат 2, Формат для SNA, XML
 - Сохранение постов в XML файле(ах)
 - все посты в одном файле
 - Путь для сохранения: G:\blogminer_v2_0_test\blogminer3\...
 - Префикс/имя файла: []
 - Buttons: Остановить, Выполнить
- Search Profiles:**
 - Оснoвные: Набор ключевых слов, Параметры для подго...
 - Список ключевых слов: "Дальнего", "Дальнего", "Дальнему", "Дальнему", "Дальний", "Дальний"
 - Список имен комментаторов: Назаров, Назарова, Назарове, Назарову, Назаровым
- Buttons:** Добавить, Сохранить, Отменить, Удалить, Создать, Сохранить, Отменить, Дублировать профиль, Удалить
- Footer:** Сору Keywords, Подготовить данные, Выгрузить результаты, Проверка FTS

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.