ИНТЕРНЕТ-ТЕХНОЛОГИИ В НАУКЕ, БИЗНЕСЕ И ОБРАЗОВАНИИ

АВТОМАТИЗАЦИЯ ПОИСКА ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ ПО ЗАДАННОЙ ТЕМАТИКЕ

Гордеев Александр Константинович
Сергеев Юрий Александрович
Финансовый университет при Правительстве РФ, Москва, студенты
s0meuser@yandex.ru, omgdef@gmail.com

Мы живем в век информационных технологий. Объемы информации, доступные в сети Интернет, скорость и количество участников информационного обмена растет с каждым годом. Чтобы эффективно использовать этот массив информации, необходимо обладать эффективными инструментами поиска и анализа информации.

Например:

- служба безопасности предприятия решает вопросы кадровой безопасности, заключающиеся, в том числе, в сборе данных о сотрудниках предприятия. Такие же сведения о сотрудниках конкурентов могут оказаться очень полезной информацией;
- маркетолог проводит маркетинговые исследования, собирает информацию о деятельности конкурентов и поведении потребителей, осуществляет мониторинг имиджа компании;
- аналитик собирает информацию для анализа;
- трейдер собирает новости о ситуации на рынке и в мире, систематизирует данные для анализа, оценивает настроение рынка, тенденции и прогнозы;
- журналист настраивается на необходимые источники информации, автоматически собирает информацию для статей, фильтрует и рубрицирует найденную информацию.

- менеджер по персоналу настраивается на необходимые кадровые агенства, специализированные порталы и форумы, автоматически собирает картотеку возможных кандидатур, извлекает и рубрицирует необходимые предприятию кандидатуры.

Поисковые системы отлично справляются с простыми однократными запросами. Однако если информационный поиск надо повторять постоянно или если предметная область сложна по структуре, то можно заметить, что:

- популярные поисковые системы сети Интернет перегружают вас тысячами бесполезных ссылок.
- поисковые системы не помнят, что вы уже видели, а что нет, и при следующем запросе принесут вам те же тысячи уже просмотренных ссылок.
- поисковые системы не умеют правильно сортировать полученную информацию и раскладывать ее по нужным рубрикам.
- Поисковые системы не всегда видят свежие тематические новости или события. Задержка в индексировании конкретного сообщения может доходить до двух недель.
- поисковая система сети Интернет выполняет поиск по конкретному запросу, а значит, нагружает вас повторяющейся рутинной работой.

Для оптимизации и автоматизации информационного поиска мы начали разрабатывать свой программный продукт – автоматизированную поисковую систему Dinase. В отличие от популярных поисковых систем сети Интернет, Dinase требует ручной настройки модели предметной области в виде списка источников и правил рубрикации. Правила рубрикации закрепляются за «умными папками». Каждая «умная папка» «знает», что в ней должно находиться и следит за своим наполнением. Сбором информации занимается специализированный поисковый робот, который постоянно работает на сервере или периодически запускается на локальном компьютере.

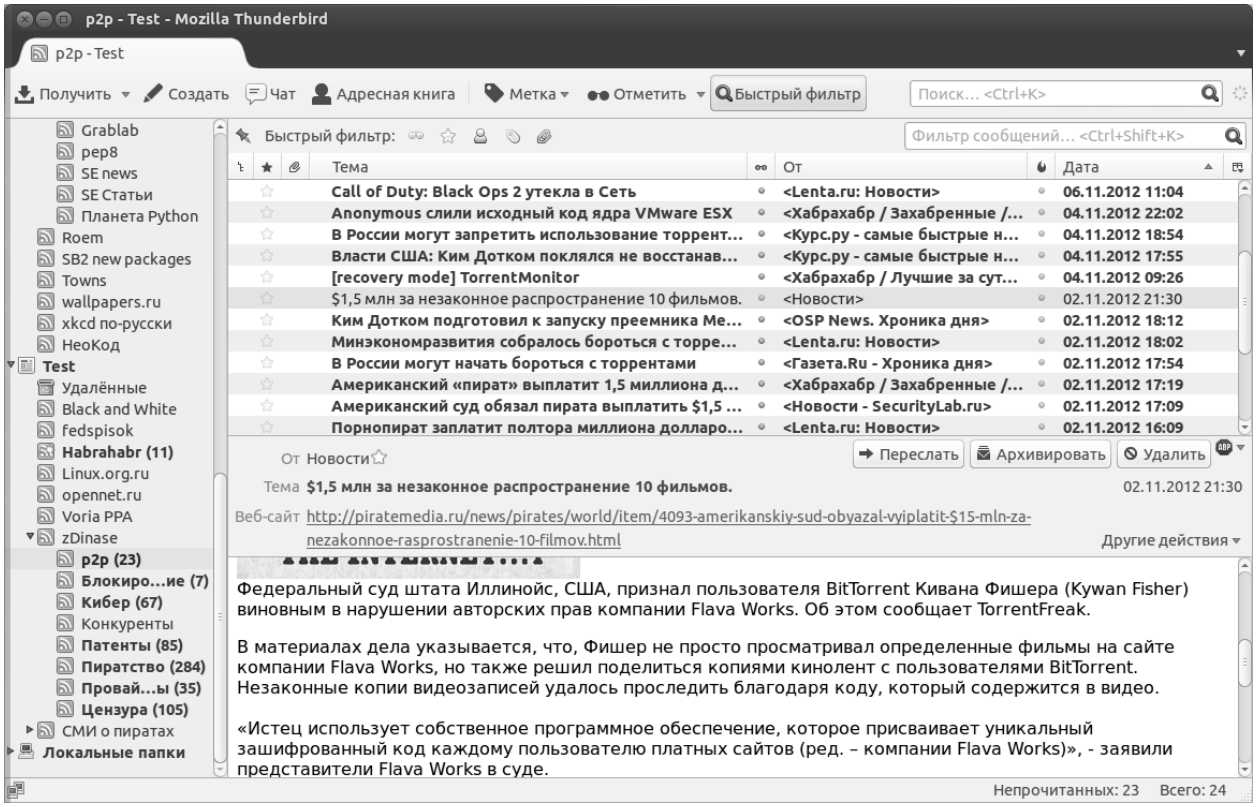

Рис. 1. Результат работы программы Dinase

Технология Dinase базируется на 7 шагах обработки информации:

По расписанию:

- сбор информации со всех указанных источников специальным роботом-пауком,
- разбор ресурсов до машиночитаемого состояния (парсинг),
- выделение «полезной» информации
- выбор новой информации,
- сохранение информации в базе данных,

По запросу «умной папки»:

- рубрицирование новой информации,
- формирование новостной ленты Atom, понятной для многих почтовых клиентов.

Пример работы программы приведен на рисунке 1.

Использование технологии Dinase особенно удачно и выигрышно в ситуации, когда поиском информации по определенной тематике приходится заниматься ежедневно. В этом случае единожды настроенная система способна

работать в автономном режиме и свести трудовые и временные затраты по поиску к минимуму.

Dinase на выходе предоставляет данные в машиночитаемом формате Atom. Благодаря этому программа не только легко интегрируется с почтовыми клиентами, которые предоставляют пользователю удобные средства управления сообщениями, но и со специальными программами постобработки, например программой эмоциональной оценки текста.

Техническая информация о программе Dinase:
Лицензия: GPL
Состояние: indev
Язык программирования: python
Окружение: паук: GNU/Linux, клиент: любое
СУБД: mongodb
Похожие коммерческие программные продукты:
Avalanche – http://www.tora-centre.ru/avl3.htm
Продукты компании RCO – http://www.rco.ru/
Продукты компании Медиалогия – http://www.mlg.ru/
Продукт X-Files компании АйТеко – http://www.i-teco.ru/xfiles.html