Разработка платформы сбора, структурирования, хранения и поиска данных, публикуемых пользователями интернета

Студент: Щитинин Дмитрий, 545 группа
Научный руководитель: Изъюров А.Л.
Рецензент: Оносовский В.В.

9 июня 2011

Введение

Отзывы и рекомендации:

- Полезная информация при выборе товара или услуги
- Высокая степень доверия (по данным различных исследований)
- Подтвержается статистикой из потоков запросов к поисковым системам

Проблемы:

- Разнородность источников: по качеству, по представлению
- Воровство, дублирование, накрутки/реклама
- Сложность точного поиска

Задача

Спроектировать и разработать платформу, позволяющую собирать отзывы с различных источников в интернете, хранить, обрабатывать и осуществлять параметрический поиск по собранным данным.

Требования:

- Простота добавления новых типов отзывов
- Работа на высоких нагрузках
- Универсальный API для поиска и публикации отзывов
- Параметрический поиск данных

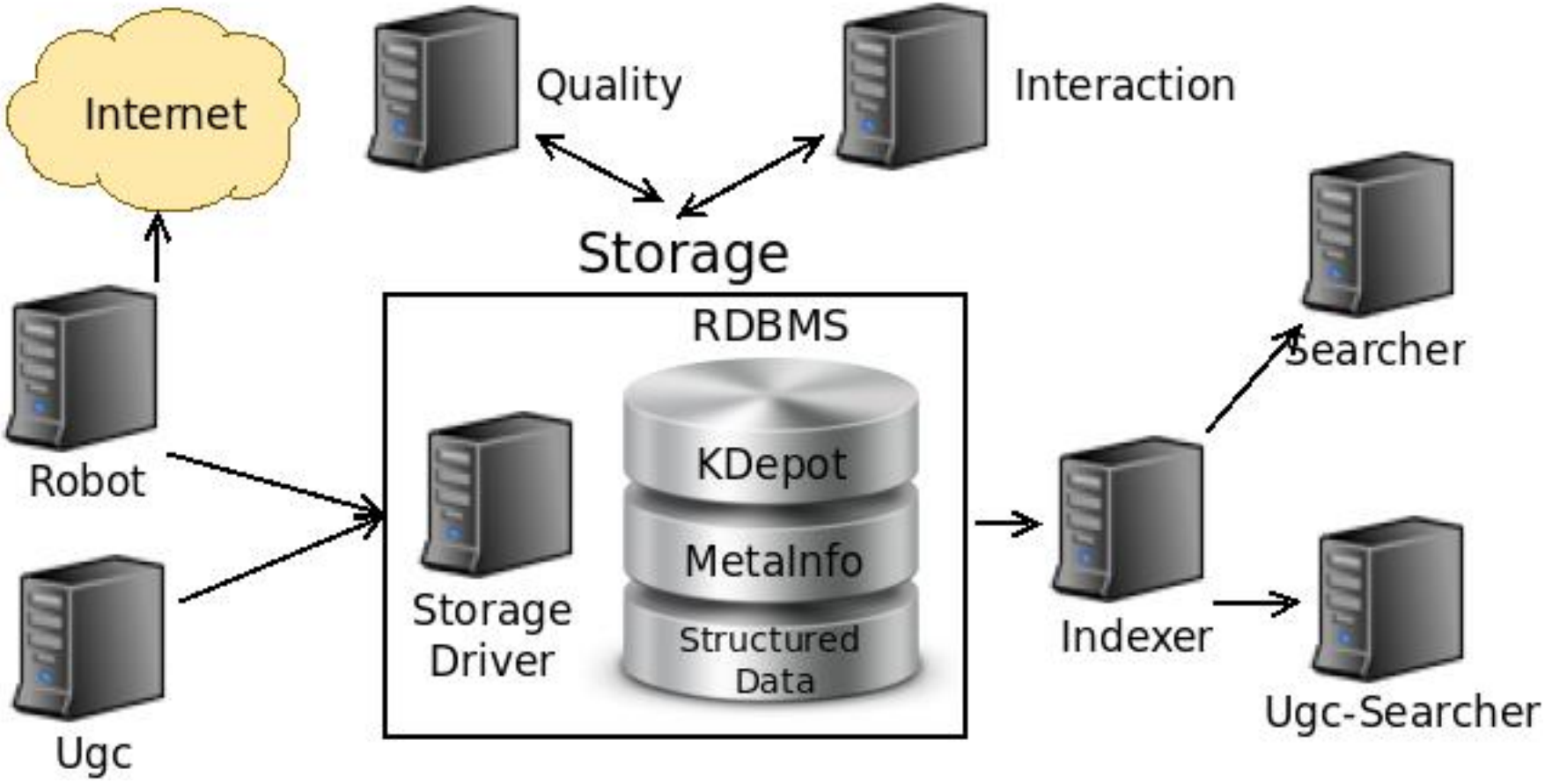
Особенности

- Отзывы различных типов имеют схожую структуру
- Новые отзывы появляются в интернете часто
- Большое количество дублей, спама, накруток
- Упрощение задачи понимания того, о чем отзыв, благодаря наличию тематического сервиса и его знаний в определенной предметной области

Реализация

- Сбор данных: микроформат hReview, местами ручной или полуавтоматический подходы
- Хранение: проприетарная технология KDepot для хранения полуструктурированных данных (реализация как метамоделированием - отображением в реляционную схему, так и нативная - на файловой системе)
- Обработка (обнаружение нечетких дубликатов): синтаксический алгоритм супершинглирования (supershingles)
- Поиск: подсистема, позволяющая гибко описывать правила индексации для различных типов отзывов. Ядро - библитека Apache Lucene.

Реализация


Internet
Quality
Interaction
Storage
RDBMS
KDepot
MetaInfo
Structured Data
Storage Driver
Robot
Ugc
Searcher
Indexer
Ugc-Searcher

Внедрение

- Использование в Яндекс.Авто
- Около 100 тысяч отзывов на автомобили
- 10-100 запросов в секунду
- Многократный запас по нагрузке
- На подходе другие сервисы

Заключение

- Спроектирована, разработана и внедрена платформа, позволяющая собирать отзывы с различных источников в интернете, хранить, обрабатывать и осуществлять параметрический поиск по собранным данным
- Простота добавления новых типов отзывов
- Работа на высоких нагрузках
- Универсальный API для поиска и публикации отзывов
- Параметрический поиск данных