Поисковые машины ближайшего будущего

Игорь Ашманов,
ЗАО «Поисковые технологии»

Постановка проблемы

Аутизм поисковиков как главная проблема

1. Поисковик как простой текстовый брокер
2. «Библиографическая лента» результатов поиска
3. Учёт только одной из трёх сил в сфере поиска
4. Безудержная портализация при сохранении аутичного поиска

1. Поисковик как формальный текстовый брокер

- Поисковики берут текстовую строку и возвращают её вхождения в тексты
- Поисковики не знают темы запроса и смысла запроса
- Поисковики не знают типа и темы возвращаемых документов
- Поисковики показывают не свои титулы и аннотации, а только то, что есть на сайте

2. «Библиографическая лента»

Результат работы «яйцеголовых»:

- Бесконечная лента: аналог списка литературы в научной статье
- Нечитаемость: невразумительный заголовок, аннотация, ненужные дата, URL, размер,
- Каша: разные типы информации в одной ленте

Мучения простого пользователя:

- Шарада: о содержании сайта нужно догадываться по URL и нечитаемой аннотации
- Метод проб и ошибок: перебор ссылок вслепую
- Программирование: подбор слов и операторов

3. Три силы в сфере поиска

Три силы с разными интересами:

- Разработчики поисковиков: поток пользователей и показ рекламы
- Пользователи: быстро найти нужный сайт
- Сайтовладельцы: первые места, поток пользователей, реклама

Поисковики:

- Дают пользователям кашу из аутичных, нечитаемых результатов
- Замешивают сайты в индекс, как неживую массу (и дают по голове каждому, кто пошевелился в жёлобе этой бетономешалки).

Результаты аутизма

- Высокое напряжение борьбы с вебмастерами, дорвейные войны
- Резкое падение качества результатов и замусоривание Интернета в целом
- Падение полноты (разнообразия) – даже при релевантной выдаче показ только одной коммерческой категории сайтов
- Рост недовольства пользователей, падение их лояльности одной поисковой машине

У Интернета есть желание и деньги улучшить поиск

- Каждый месяц появляются поисковые стартапы.
- Под «поиск» охотно дают деньги, а под «поиск с социальными сетями» – ещё более охотно.
- Крупные игроки резко замедлились и возятся с инфраструктурой, продажами, большими индексами, большим персоналом.
- Большие поисковики ориентированы на борьбу друг с другом; доминируют бизнес-идеи, в частности, война за desktop.

Как улучшить поиск?

Этапы работы поисковика

1. Выбор сайтов для обхода
2. Скачивание и индексация
3. Получение запроса от пользователя
4. Разбор запроса
5. Вычисление запроса (собственно поиск)
6. Показ результатов поиска

Улучшить поиск можно на каждом из этих этапов, и многочисленные стартапы это уже делают

1. Выбор набора сайтов

Выбор сайтов может решить проблему мусора и генерации дорвеев:

- Специальные поисковики (Dash, Аппликата, Новотека, Тындекс, iligent, Яндекс.Блоги и пр.)
- Выбор вебмастерами и пользователями (Персональный поиск Новотеки, Rollyo, пр.)
- Обмен размеченными списками сайтов (Del.icio.us etc.)

Большие поисковики пока этим пользуются мало, но есть множество стартапов

2. Выкачка и индексация

- Распознавание типа данных на этапе выкачки (форумы, блоги, товарные предложения, статьи, новости, описания товаров).
- Распознавание темы страницы (семантическое индексирование)
- Семантический разбор текстов, выделение объектов и фактов
- Разные индексы для разных типов сайтов.

Большие поисковики этим занимаются, но во вторую очередь, зато есть множество стартапов

3. Получение запроса

- Регистрация запросов/ответов вебмастерами
- Подсказка и уточнение запросов
- Программирование и обмен запросами между пользователями (MS)
- Персонализация, запоминание истории запросов (Yahoo, Google)
- Запрос на естественном языке (AskJeeves)

Ведутся активные работы в больших поисковиках, в множестве стартапов

4. Разбор запроса

- Распознавание темы запроса (каталог запросов)
- Распознавание типа запроса (анализ лексики)
- Разбор синтаксиса и семантики запроса
- Уточняющий диалог: итеративные запросы

Работы в больших поисковиках пока идут вяло, есть соответствующие стартапы

5. Поиск: вычисление запроса

- Повышение релевантности (улучшение алгоритмов)
- Учёт прошлых интересов пользователя
- Учёт поведения пользователей на поисковике

Почти всё уже сделано. Активно ведутся работы в больших поисковиках, стартапам труднее – нет базы текстов и логов запросов

6. Показ результатов

Здесь наивысшая плотность новых идей:

- Выдача по типам (большие поисковики, A9, Аппликата, Dash)
- Тематическая кластеризация (Clusty, Нигма, Квинтура)
- Графическая выдача и навигация (Vizzy, Квинтура, Тропа, etc.)
- Персонализация и настройка результатов (все)

Здесь ведутся бурные работы, ибо интерфейс – то, что в первую очередь видит и пользователь, и инвестор

Отдельная история: социальные сети поверх поиска

Сообществу можно поручить почти весь цикл настройки поисковика:

- Отбор сайтов (ведение каталогов)
- Создание названий и аннотаций сайтов
- Обмен индексами, группами сайтов
- Регистрацию и подбор запросов
- Оценку и разметку результатов поиска
- Обмен результатами поиска

Отдельная история: борьба за desktop

- Возможно, исход борьбы за поиск будет решён на поверхности рабочего стола
- Google и Microsoft, похоже, уже сделали на это ставку.
- Здесь основным преимуществом будет не функциональность, а гладкая совместимость с ОС и офисными приложениями
- Я бы поставил на выигрыш MS

Перспективы развития поисковиков

Всё вскипело:

- Очень много шумихи
- Очень много денег
- Очень много стартапов

Условия успеха:

- Удобство поиска на уровне DOS 1990 г.
- Нехватка рекламных площадей
- Падение качества и война с вебмастерами

Будем ждать революции. Кто даст миру «Windows для поиска»?

Будущие поисковые машины

- Каталоги сайтов: регистрация сайтов сообществами
- Каталоги запросов: регистрация и разметка запросов вебмастерами
- Структурированная выдача: по теме и типам документов.
- Читаемость выдачи: названия, аннотации, тэги от сообществ
- Понимание запроса: запросы на ЕЯ, распознавание темы, уточнение запроса
- Новые виды заработка на поиске: регистрация запросов, сайтов, ранжирование, хостинг поиска

У чего нет будущего

- «Библиографическая лента»
- Дальнейшие усилия по повышению традиционной «релевантности»
- Похвальба размером индекса
- Отношение к вебмастерам, как мёртвому материалу
- Ссылочное ранжирование (PageRank) в его текущем виде
- Автоматическая кластеризация результатов по темам
- Персонализация и настройка пользователем

О чём не сказано в этом докладе

- Мобильный поиск (борьба за смартфон)
- Локальный поиск и геотаргетинг (борьба за парикмахера)
- Блогопоиск и новостные поисковики (борьба за аффтара)
- Многоязыковый поиск и перевод (борьба с языковым барьером)
- Национальные проекты (борьба против Гугла)
- Ну и так далее .....

Спасибо за внимание!

Пишите: igor@ashmanov.com