Search Based Applications at the Confluence of Database and Search Engine Technology

Gregory Grefenstette
Old World View

**Structured Data**
- Databases
- Transaction Logs
- Spreadsheets
- LDAP Directories
- ERP Exports
- Application Data

**Unstructured Data**
- Web Pages
- Email
- Office Documents
- RSS Feeds
- PDFs
- Blogs & Forums
Two types of information, two ways to find information

**DATABASES**

Structured Data
Transaction
All tuples
Safe, Precise, SQL
Slow

**SEARCH ENGINES**

Text
Similarity Ranking
Intuitive
Fast
Partial
Organisational Data *all over the place*

**Source Systems**
- JDE
- SAP
- Siebel
- Scala
- Oracle
- Access
- PS

**Delivery Systems**
- CRM
- BI
- ODS
- RCK
- ODS
- Off-line analysis

20+ types of systems
6% with 50+ ERP systems alone

*Source: Leveraging Search to Improve Contact Center Performance*
Richard Snow, VP & Research Director, Customer & Contact Center Management
March 2009
Search Based Applications Goals

**Users**
- Limited number of users
  - Usage complexity, production costs
- Large number of users
  - Ease of use, traffic scalability

**Interface**
- Functional
  - Heavy one-shot development
- Agile applications
  - Simple data access, use of standard web technologies

**Querying**
- Dedicated resources
  - Datamarts, additional hardware
- Generic data layer
  - Real time data, high performance querying

**Data Source**
- Structured data
- All data
  - Connectors, structuration of data
Search Engine now handle rich semantics

- **Text fields**
  - “... the certification test is ...
- **Numerical fields**
  - 3.14159265
- **Date**
  - 16/11/1957
- **GPS coordinates, real world**
  - 48.451065619, 1.4392089
- **Categories**
  - Top/Animals/Pets/Dog
- **Metadata**
  - Source: dailymotion
Advantages of Search Based Applications

Usability
360°, Google-like, interactif, conversational

Agility
Flexible, Agile, Days vs Months

Performance
Real time, millions of end-users, Terabytes of information
# How Search Based Applications Work

## Connect
- Get real-time flows of data
- Get and maintain security information

## Process
- Manage file formats (PDF, office, drawings, XML…)
- Ability to understand free text to relate text to business objects

## Access
- A highly scalable data repository
- Full text search, navigation and reporting capabilities (the Index)

## Interact
- A Framework to create search oriented web applications at the speed of light
- Create virtual feeds of information (RSS, etc…) and associate widgets

---

Packaged as one, easy to deploy software

Built-in cluster architecture, for high-availability and scalability

---

Complete DIY

Perfect SBA
When to SBA and not to SBA

- Beware that a SBA is not
  - A replacement for transactional applications
    - A SBA won’t manage your workflows, lifecycles, and won’t modify the existing systems
  - A good excuse to drop your business intelligence software
    - A SBA goal is not produce the pixel perfect highly complex final report you have to submit to the SEC
  - Replace all your complex, historical business systems
    - A SBA goal is not to reproduce all the business logic of existing applications. It’s to simplify it for information access

- A SBA addresses critical business issues by enabling easy Search & Discovery into key data by key users
Four Types of Search
Four Types of Search

Form Based Search (overlay database)

Traditional search on database is complicated:

- Several fields to fill in
- Need to know field values
Four Types of Search

Unique Search Box

![Cuil Search Engine](image-url)
Four Types of Search

Faceted Search
Four Types of Search

Map Search (GPS, mobile)
Four Types of Search

- In the past ten years, people have learned to use the unique search box

<table>
<thead>
<tr>
<th>Rank</th>
<th>Google</th>
<th>Bing</th>
<th>Ask</th>
<th>Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26.79%</td>
<td>46.76%</td>
<td>49.90%</td>
<td>54.15%</td>
</tr>
<tr>
<td>2</td>
<td>23.39%</td>
<td>18.81%</td>
<td>13.03%</td>
<td>18.11%</td>
</tr>
<tr>
<td>3</td>
<td>18.72%</td>
<td>15.92%</td>
<td>16.09%</td>
<td>12.31%</td>
</tr>
<tr>
<td>4</td>
<td>12.78%</td>
<td>8.40%</td>
<td>6.72%</td>
<td>7.08%</td>
</tr>
<tr>
<td>5</td>
<td>8.23%</td>
<td>5.23%</td>
<td>6.42%</td>
<td>3.73%</td>
</tr>
<tr>
<td>6</td>
<td>4.55%</td>
<td>1.94%</td>
<td>3.77%</td>
<td>2.47%</td>
</tr>
<tr>
<td>7</td>
<td>2.76%</td>
<td>1.40%</td>
<td>0.71%</td>
<td>0.97%</td>
</tr>
<tr>
<td>8</td>
<td>1.36%</td>
<td>0.71%</td>
<td>2.24%</td>
<td>0.68%</td>
</tr>
<tr>
<td>9</td>
<td>1.02%</td>
<td>0.77%</td>
<td>0.81%</td>
<td>0.33%</td>
</tr>
<tr>
<td>10</td>
<td>0.41%</td>
<td>0.06%</td>
<td>0.31%</td>
<td>0.18%</td>
</tr>
</tbody>
</table>

- People have learned to use facets on shopping sites
- People are *learning* to do map search
- Forms are still boring
Faceted Search and Semantics

› Facets are semantic dimensions
› They are visualisations of semantics that users can understand
› Semantics can come from databases
› Semantics can come from text
› Common semantic dimensions link together structured and unstructured data
› Search Based Applications are possible because of semantics
Semantics

Type

Equality

Relation
Databases are “structured”

Database semantics comes from row:column

- Column defines *Type* of entity
  - Examples: client, supplier
- Row defines *Relation* between entities
  - *Primary key* - main entity
- As for *Equality*
  - This is the whole problem of Master Data Management

<table>
<thead>
<tr>
<th>Name</th>
<th>User</th>
<th>Phone</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graff</td>
<td>rgraff</td>
<td>392-3900</td>
<td>Pharmacy</td>
</tr>
<tr>
<td>Harris</td>
<td>bharris</td>
<td>392-5555</td>
<td>Medicine</td>
</tr>
<tr>
<td>Ipswich</td>
<td>zipswitch</td>
<td>846-5656</td>
<td>PHHP</td>
</tr>
</tbody>
</table>

**Immediate Federated View**

<table>
<thead>
<tr>
<th>Name</th>
<th>User</th>
<th>Phone</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erik</td>
<td>McClain</td>
<td>Address 1</td>
<td>Billing</td>
</tr>
<tr>
<td>McClain</td>
<td>Erik</td>
<td>Email</td>
<td>CRM</td>
</tr>
<tr>
<td>Erik</td>
<td>McClain</td>
<td>Birth Date</td>
<td>Support</td>
</tr>
<tr>
<td>Erik</td>
<td>McClain</td>
<td>Phone</td>
<td>ERP</td>
</tr>
</tbody>
</table>
Text is “unstructured”
  ❖ Semantics not explicit
    • Resources and processing needed

Natural Language Processing
  ❖ **Typing** entities/things
    • Rules, lists, ontologies
  ❖ **Equality**
    • Linguistic variants, morphology, stemming, synonyms
  ❖ **Relations**
    • Parsing, co-occurrence (related terms), Linked Open Data

---

Google has acquired social search service Aardvark, says a source that has been briefed on the deal, for around $50 million. We first reported on the discussions between the two companies ...
Semantics in databases and text

Database
- **Type** == column
- **Relation** == row
- **Equality** == ???

Text
- **Type** == ontologies
- **Relation** == parsing
- **Equality** == morphology, synonyms

- Search Based Applications
  - Structure from databases
  - Linguistic variation from text
    - **Types** == facets
    - **Relation** == fields
    - **Equality** == text processing
Semantics in Search Engine
Database semantics is imported into search engine facets.
Semantics is extracted from unstructured text by NLP

autosuggestion
language identification
related terms
related queries
spell checker
faceted search
local search

lemmatisation
synonyms
phonetic
cross language
Ontology Matcher

INDEX

phonetic
lemmatisation
stopwords
sentiment analysis
named entity

QueryMatcher
FastRules
HTMLRelevantContextExtractor
Categorizer
Clusterer

Data Sources
Search Pipelines

Understand the user query & enhance results

**Language Detection**
- 54 languages supported (used to choose the right tokenizer)

**Parsing (Tokenizer)**
- Split words, detect end of phrase, ...

**Synonym Expansion**
- Expanded query with user-defined synonym

**Lemmatization / Phonetic**
- Determine lemma and stem of the word. Apply Phonetization algorithms

**Query Rewriting (Regexp, …)**
- Rewrite special expressions such as word*

**Transformation into index query**
- Query is rewritten to be comprehensible by the index

**Index**

**Return results**
- Index return results set matching user query and security rights

**Ranking**
- Rank results using density, text scoring and ranking formula

**Related Terms**
- Extract Related Terms from the result set

**Summary**
- Determine Summary to be displayed for each hit

**Highlighting**
- Highlight the words that are matching the user query

Search Side

**Understand the user query & enhance results**

**Language Detection**
- 54 languages supported (used to choose the right tokenizer)

**Parsing (Tokenizer)**
- Split words, detect end of phrase, ...

**Synonym Expansion**
- Expanded query with user-defined synonym

**Lemmatization / Phonetic**
- Determine lemma and stem of the word. Apply Phonetization algorithms

**Query Rewriting (Regexp, …)**
- Rewrite special expressions such as word*

**Transformation into index query**
- Query is rewritten to be comprehensible by the index

**Index**

**Return results**
- Index return results set matching user query and security rights

**Ranking**
- Rank results using density, text scoring and ranking formula

**Related Terms**
- Extract Related Terms from the result set

**Summary**
- Determine Summary to be displayed for each hit

**Highlighting**
- Highlight the words that are matching the user query
Ontology Matcher: add an alternative indexing form to a string span

Given an existing taxonomy or ontology, normalized index terms can be produced.
In this video:

People: Allame, Angeliere, Antin, Audrey Pulvar, Australia, Canada, ETA, Esonne, Europe, Gregory Coupet, Geraldine Giraud, Gerard Sylve, Jean Francois Cope, Jean Pierre Prevert, Julian Drey, Kabina, Lherber, Lens, MarieChristine Van, Marseille, Mathieu Canada, Michel Chassang, Moscow, Nicole Boyaut, Oise, PS, FSG, Russia, SaintPetersbourg, Suisse, Sydney, UMP, Yamaha

Organization:  

Location: 

Return to results

6 hours ago.

moscow

Select sources

In this video:

priere de meme que son clocher la messes n'est pas dite.
Toujours a l'etranger en Russie et vingt-six morts cent blesses et dix-huit disparus c'est le dernier bilan de l'attentat contre le train Moscou Saint-Petersbourg avant hier les restes d'un engin explosif ont ete decouverts sous les rails et l'enquete est confuse au FSB Mathieu Canada.
Toute la nuit ils ont travaille pour reparer ce qui avait ete detruit des employes des chemins de fer russes pose de nouveaux rails alors droit ou l'exploision a eu lieu le trafic reprend sur la ligne Moscou.
Semantics in Search

› Suggestions
› Spell correction
› Variant forms
  • Regular expressions
  • Natural language processing
› Synonym expansion, cross lingual search
› Related terms, related queries
› Entities, Classification, facets
› Sentiment Analysis
Search Based Applications
Provide interactive reporting

Infinite number of dimensions

Manage simultaneously both dimensions and single records
Build decisional tools on unstructured information

Reconnect the causes (unstructured) with the effect (structured)
Facets can make search results into Business intelligence aggregating numerous unstructured data sources.
Examples of Search Based Applications
Hybrid Search:
Unify structured and unstructured data sources

A unique search bar
Structured results from a Database
Unstructured data from the Web
Merged Facets
Unify structured and unstructured data sources

Structured results from a Database

Unstructured data from the Web

Semantics from Text
Directories: Urbanizer

Semantic Analytics for Mood-Based Local Search

YPG Database + Web & UGC content + Social Networking
Map-Based Access and Reporting for Water Quality Data

Sourcier

Points d'eau : 20996 résultats(s)

10153X0061/BOYNE
Localisation : Cazouls-D'Hérault
Données : 293 prélèvements - 524 analyses

10154X0058/F
Localisation : Montagnac
Données : 353 prélèvements - 120 analyses

10154X0076/MAGMART
Localisation : Périssan
Données : 760 prélèvements - 396 analyses

10155X0010/P
Localisation : Servian
Données : 452 prélèvements - 124 analyses

10155X0026/AEP
Localisation : Lieuran-Les-Bastides
Données : 657 prélèvements - 206 analyses

Source: Portail ADES. Données exportées du 31/07/2009 - Conditions d'utilisation - EXALEAD - BRGM - Contacts
**Description**

- Agile MDM project
- Part catalogue normalization & enrichment
- Aggregation of content from ERP & Directories
- ERP (SAP) & Directories (online directories)
- Aggregation of:
  - Part number
  - Technical description (size, length)
  - Description
- Semantic analysis:
  - Merging information from SAP and vertical directories on parts
  - Similar product feature bases on technical description
Current and Future Needs: Domain Specific Semantics
Remaining challenges: Intelligent query interpretation

Intelligent analysis of query to extract relevant terms:

- 2 stars: mandatory field
- Swimming pool: Optional with a boost on matching results
Description

- B2C Touristic Portal
- Automatic aggregated content
- Exalead 360 + Business Console modules
- Aggregation of:
  - Online news content
  - Hotel directory
  - Entertainment directory
  - Railway station location
- Semantic analysis:
  - Type “Corrida” or “Beaujolais” and the system auto-discover entertainment or railways stations co-cited in the news.
Search engines can handle the semantics of databases (but not the transactions)

Facets are semantic dimensions

Semantics allows for « business intelligence » type reporting

Search Based Applications use the power of search engines (intuitive, scaling, agility) to extract and merge information from databases and text

More semantic types are needed for text