Semantic Search for the Enterprise

International Workshop on Search Computing, Brussels, September 2012,
Dr. Christoph Goller, Director Research, IntraFind Software AG
Emergence of Lucene / Solr / Elasticsearch

- Lucene / Solr / Elasticsearch
  - Built in late 90’s by Doug Cutting…. Apache release 2001
  - Wide acceptance by 2005
- 4,000+ sites – Apple, Cisco, EMC, HP, IBM (OmniFind, Watson), MySpace, CNET, Netflix, Salesforce, Twitter, Ebay, Scout24, Gov, Wikipedia…,
- Rapid Innovation, Extensible Architecture, complete control (open source)
- CORE TECHNOLOGY AS GOOD OR BETTER THAN ANY OTHER … AND OPEN SOURCE
- Fulltext Search has become a commodity
- Good Enterprise Search Solutions have to offer more than just full-text search
IntraFind Software AG

- Founding of the company: October 2000
- More than 700 customers mainly in Germany, Austria, and Switzerland
- Partner Network (> 30 VAR & embedding partners)
- Employees: 30
- Lucene Committers: B. Messer, C. Goller

- **Product Company:** iFinder, Topic Finder, Knowledge Map, Tagging Service, …
- Products are a combination of Open Source Components and in-house Development
- Support (up to 7x24), Services, Training, Stable API
- **Text Analytics:** Automatic Generation of Semantics
  - statistical and linguistic methods
  - **stable, scalable, easily adaptable to new domains, easy to use for partners**
  - Linguistic Analyzers for most European Languages (+ Chinese & Japanese)
  - Named Entity Recognition
  - Text Classification
  - Clustering
Semantic Search for the Enterprise
Semantic Web / Semantic Applications

Semantic Web proposed by Tim Berners-Lee, the founder of the WWW

**Idea:** Computers should be able to

- evaluate information according to its meanings
- connect information
- reason with it (inference), generate new information

**Problem:**

„Metacrab“ Critics by Cory Doctorow or Peter Norvig:

- „People are lazy“
- „People are stupid“
- „People cheat and lie“

Where will meanings (reliable annotations and meta information) come from?

**Solution:** Text Analytics can produce meta information (semi-) automatically.
Morphological Analyzer vs. Stemming

Morphological Analyzer:

- **Lemmatizer**: maps words to their base forms

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>going -&gt; go (Verb)</td>
<td>lief -&gt; laufen (Verb)</td>
</tr>
<tr>
<td>bought -&gt; buy (Verb)</td>
<td>rannte -&gt; rennen (Verb)</td>
</tr>
<tr>
<td>bags -&gt; bag (Noun)</td>
<td>Bücher -&gt; Buch (Noun)</td>
</tr>
<tr>
<td>bacteria -&gt; bacterium (Noun)</td>
<td>Taschen -&gt; Tasche (Noun)</td>
</tr>
</tbody>
</table>

- **Decomposer**: decomposes words into their compounds
  
  Kinderbuch (children’s book) -> Kind (Noun) | Buch (Noun)
  Versicherungsvertrag (contract of insurance) -> Versicherung (Noun) | Vertrag (Noun)
  Holztisch (wooden table), Glastisch (table made of glass)

- **Stemmer**: usually simple algorithm

  going -> go
  king -> k
  Messer -> mess
Bad Precision with Algorithmic Stemmer

Apache Lucene/Solr - Who We Are
Lucene/Solr is maintained by a team of volunteer developers. Core Committers Bill Au (billa@...) Michael Busch (buschmi@...) Doron Cohen (doronc@...) Doug Cutting (cutting@...) Shai Erera (shaie@...) Erick Erickson (erick@...) Olis Gospodnetic (olis@...) Martijn van Groningen (mvg@...) Erik Hatcher (ehatcher@...) Chris Hostetter (hossman@...) Jan Hoydahl (janhoy@...) Grant Ingersoll (gingers@...) Mike McCandless (mikemccand@...) Ryan McKinley (ryan@...) Chris Male (chrism@...) Bernhard Messer (bmesser@...) Mark Miller (markmiller@...) Robert Muir (rmuir@...) Stanislaw Osinski (stanislaw@...) Noble Paul (noble@...) Steven Rowe (srowe@...) Uwe Schindler (uschindler@...) Shalin Shrikiar (shalin@...) Yonik Seeley (yonik@...) Koji Sekiguchi (koji@...) Dawid Weiss (dweiss@...) Andi Vajda (vajda@...) Simon Willnauer (simonw@...) Emeritus Committers Josh Bloch Peter Carlson (carlson)
http://lucene.apache.org/java/docs/whoweare.html

Information creation proliferation? Forbes looks to Solr and Lucid Imagination
2011-01-12 11:42
Information: we all love to make more of it, but it sure piles up. Forbes Magazine Online blogger Quentin Hardy takes on making sense of it in a nice post about Solr/Lucene and Lucid Imagination. Some soundbites: Our civilization may pride itself on the amount of information we create – more than every conversation, ever, in a couple of years; enough to jack The Library of Congress to the Moon on data-packed CD-Roms, take your superlative – but we’ve also made a holy mess of it. It is mostly, as they say, “unstructured,” meaning as random as your last 25 e-mails, your tweets, and all those spreadsheets and documents piling up at work. That’s why Solr/Lucene and Lucid Imagination are names you need to know in tech. ... That’s because of the phenomenal
http://www.lucidimagination.com/blog/?p=2874

Information creation proliferation? Forbes looks to Solr and Lucid Imagination
Information: we all love to make more of it, but it sure piles up. Forbes Magazine Online blogger Quentin Hardy takes on making sense of it in a nice post about Solr/Lucene and Lucid Imagination. Some soundbites: Our civilization may pride itself on the amount of information we create – more than every conversation, ever, in a couple of years; enough to jack The Library of Congress to the Moon on data-packed CD-Roms, take your superlative – but we’ve also made a holy mess of it. It is mostly, as they say, “unstructured,” meaning as random as your last 25 e-mails, your tweets, and all those spreadsheets and documents piling up at work. That’s why Solr/Lucene and Lucid Imagination are names you need to know in tech. ... That’s because of the phenomenal
http://www.lucidimagination.com/lucene-solr-blog/information-creation-prolife...

Estimating Memory and Storage for Lucene/Solr
2011-09-14 05:27
if it for what you are actually seeing in your system. It is a DRAFT. It is likely missing a few things, but I am putting it up here and in Subversion as a means to gather feedback. I reserve the right to have messed up the calculations. I feel the values might be a little bit
http://www.lucidimagination.com/blog/?p=4000
Suchergebnisse für "buchen"

STUDENTENANFRAG
Willkommen in der großen Maschine Universität...

Handy-Fahrschein
Von der Deutschen Bahn verfolgt...

SACHBUCH
Liebe und solche Sachen...

Spitzengastronomie
Frankrechts Köche setzen Jean-François Piège auf Platz 1...

8044 ERGEBNISSE
Sortieren nach: RELEVANZ | DATUM

Zufluchsort Deutschland
Das gelobte Land...

Suchen

03.10.2011, DIE ZEIT

03.10.2011, DIE ZEIT

27.09.2011, DIE ZEIT

27.09.2011, DIE ZEIT

Semantic Search for the Enterprise
High Recall and High Precision with Morphological Analyzer

Semantic Search for the Enterprise
Advantages of Morphological Analysis

- Combines high Recall with high Precision for Search Applications
- Improves subsequent statistical methods
- Better suited as descriptions for semantic search / faceting / clustering than artificial stems, categories used for phrase detection
- Reliable Lookup in Lexicon Resources
  - Thesaurus / Ontologies
  - Cross-lingual search
- Available Languages:
  - German, English, Spanish, French, Italian, Dutch, Russian, Polish, Serbo-Croatian, Greek, (Chinese, Japanese, Arabian, Pasthu)
Named Entity Recognition (NER)

Automated extraction of information from unstructured data

- People names
- Company names
- Brands from product lists
- Technical key figures from technical data (raw materials, product types, order IDs, process numbers, eClass categories)
- Names of streets and locations
- Currency and accounting values
- Dates
- Phone numbers, email addresses, hyperlinks
- Search query: Which people / companies strongly correlate with each other?
Frage: Wo liegen Werke von Audi?

NL-Search

Audi – Wikipedia

Die Fahrzeuge der Marke Audi werden außer in den beiden deutschen Werken Ingolstadt und Neckarsulm in Győr (Ungarn), Bratislava (Slowakei), Changchun (Volksrepublik China), Brüssel...

Google

Audi und BMW: Neue Werke in Bayern - Börse - FOCUS Online
Audi-Chef Franz-Josef Paefgen stößt mit seinen Werken in Ingolstadt, Neckarsulm und im ungarischen Győr an die Kapazitätsgrenzen. Wie FOCUS- MONEY erfuhr, ...

INTRA FIND

... höre“ ins Lateinische „Audi“... Im Juli 1910 verließ das erste Fahrzeug mit dem Namen Audi das Zwickauer Werk. 1915 wurde die „Audi Werke AG“ gegründet. Nachdem Audi 1928 ...
Frage: Wer hat Microsoft gegründet?

Semantic Search: Comparison with standard Search Engines

NL-Search

Microsoft

sein Betriebssystem Windows und seine Büro-Software Office. Das Unternehmen wurde 1975 von Bill Gates und Paul Allen gegründet. Der Name Microsoft steht für Microcomputer-Software, ursprünglich...

Google

Microsoft – Wikipedia

Das Unternehmen wurde 1975 von Bill Gates und Paul Allen gegründet. ..... Juli 2004 hat Microsoft bekanntgegeben, dass es nach der nun erfolgten Beilegung ...

INTRA FIND

Microsoft Dynamics NAV ist eine Standardsoftware für ERP-Systeme. 2002 übernahm Microsoft den dänischen Hersteller und integrierte es in seinen Geschäftsbereich Microsoft Business Solutions. Seitdem wird Microsoft Dynamics ...
Google US: Semantic Search

Search
About 35,200,000 results (0.14 seconds)

Ulm
Albert Einstein, Place of birth

Albert Einstein - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Albert_Einstein
Albert Einstein was born in Ulm, in the Kingdom of Württemberg in the German Empire on 14 March 1879. His father was Hermann Einstein, a salesman and ... Brain - Religious views of Albert ... - General relativity - Einstein family

Albert Einstein - Biography
www.nobelprize.org/nobel_prizes/physics
Biography. Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where he later on began ...

10 Strange Facts About Einstein - Neatorama
26 Mar 2007 – When Albert’s mother, Pauline Einstein gave birth to him, she thought that Einstein’s head was so big and misshapen that he was deformed!

When and where was Albert Einstein born and when did he die
wiki.answers.com › Famous People › Scientists › Albert Einstein
Einstein’s Birth and Death. Born: 14 March 1879. Birthplace: Ulm, Germany; Died: 18 April 1955 (Abdominal Aortic Aneurysm); Place of death: Princeton, NJ, ...

Ulm
Ulm is a city in the federal German state of Baden-Württemberg, situated on the River Danube. The city, whose population is estimated at 120,000, forms an urban district of its own and is the administrative seat of the Alb-Donau district. Wikipedia

Founded: 1181
Area: 118.7 km²
Local time: 1:16pm Tuesday (CEST)
Enterprise Search as Central Enterprise Information System

Answer queries such as:
- phone Number [Person]
- department [Person]
- experts in [technology, projects, ….]
- director of [project, ..]
- all invoices of [customer, supplier, …] for the last three months
- raw materials, needed for …
RDF Stores & Enterprise Search

- RDF Stores:
  - Originally designed as Meta-Data model: machine-readable information
  - Triples: subject-predicate-object
  - General Data Model for Knowledge Representation

- Use RDF Stores to store structured data
  - Enterprise Databases
  - Extracted from documents

- Link structured and unstructured data

- RDF stores are designed for machine-usability not human usability:
  - combine Fulltext Search an RDF Stores into a single (Enterprise) Content Repository
  - provide a simple unified intuitive user interface for accessing all (Enterprise) knowledge, structured and unstructured
iFinder: Semantic Search and Navigation

Unique Search Box

Result from Structured Data

Result from Unstructured Data

Facets (correlated terms)

Facets (named entities)

Facets (other meta-data)
Questions?

Dr. Christoph Goller
Director Research

Phone: +49 89 3090446-0
Fax: +49 89 3090446-29
Email: christoph.goller@intrafind.de
Web: www.intrafind.de

IntraFindSoftware AG
Landsberger Straße 368
80687 München
Germany