



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 / elesticsearch
 - 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

IntraFind Software AG

INTRAFIND



IntraFind Software AG

INTRAFIND



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

INTRAFIND

Morphological Analyzer:

• Lemmatizer: maps words to their base forms

EnglishGermangoing -> go (Verb)lief -> laufen (Verb)bought -> buy (Verb)rannte -> rennen (Verb)bags -> bag (Noun)Bücher -> Buch (Noun)bacteria -> bacterium (Noun)Taschen -> Tasche (Noun)



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

INTRAFIND

[WEB] 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@...) Otis Gospodnetic (otis@...) Martijn van Groningen (mvg@...) Erik Hatcher (ehatcher@...) Chris Hostetter (hossman@...) Jan Høydahl (janhoy@...) Grant Ingersoll (gsingers@...) Mike McCandless (mikemccand@...) Ryan McKinley (ryan@...) Chris Male (chrism@...) Bernhard Messer (bmesser@...) Mark Miller (markrmiller@...) Robert Muir (rmuir@...) Stanislaw Osinski (stanislaw@...) Noble Paul (noble@...) Steven Rowe (sarowe@...) Uwe Schindler (uschindler@...) Shalin Shekhar Mangar (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

[LUCID] 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 emails, 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

[LUCID] 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 emails, 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...

[LUCID] Estimating Memory and Storage for Lucene/Solr

2011-09-14 05:27

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

High Recall and High Precision with Morphological Analyzer

INTRAFIND

Suchergebnisse für "buchen"

IHRE SUCHE	buchen					Suchen
DATUM FILTERN	ALLE INHALTE	HEUTE	24 STUNDER	N 7 TAGE	30 TAGE	
	VOM			BIS		
CEMALLES DATUM	14		I COLOR			

ZUFLUCHTSORT DEUTSCHLAND Das gelobte Land

... Familie schon einmal einen Kurzurlaub in Deutschland gebucht, auf dem Oktoberfest waren sie. Die Lebensfreude der Deutschen habe seinen Kindern gefallen, sagt er. Der Rest

habe sie eher kaltgelassen, vor allem das Essen. Er lächelt. »Aber vielleicht habe [weiter...] 03.10.2011. DIE ZEIT



8044 ERGEBNISSE

Sortieren nach: RELEVANZ DATUM

STUDENTENANDRANG

Willkommen in der großen Maschine Universität

Filter: ALLE INHALTE NUR REZENSIONEN

... rückt, der lieber dort als in Saarbrücken studiert usw. Darum müssen die Unis die Studiengänge "über**buchen**". Zum Semesterstart merken sie dann oft, dass sie sich geirrt

haben: Das "Annahmeverhalten" war besser als geschätzt. Dann haben mehr

Studierende [weiter...] 18.10.2011, ZEIT ONLINE



SACHBUCH Liebe und solche Sachen

..., wird nie mehr in aller Unschuld in diesem angesagten Restaurant für einen Jahrestag der Liebe den teuren Fensterplatz **buchen** oder naiv die Reise zu zweit in den

Süden für eine individuelle Entscheidung halten, also blind sein gegenüber dem Ausleben von [weiter...]



HANDY-FAHRSCHEIN

Von der Deutschen Bahn verfolgt

... Endpunkt jeder Fahrt einzugeben. Das System berechnet den Fahrpreis und **bucht** ihn vom Konto des Nutzers ab. Selbstverständlich traut die Bahn ihren Kunden nicht und will

überprüfen, ob deren Angaben stimmen. Sie sammelt dazu Bewegungsdaten. Das ist [weiter...] 27.09.2011, ZEIT ONLINE



SPITZENGASTRONOMIE

Frankreichs Köche setzen Jean-François Piège auf Platz 1

... ständig aus**gebucht**. Die Köche als Leser des Branchenmagazins Le Chef konnten den ihrer Ansicht nach

besten Kollegen selbst bestimmen. Es gibt keine Liste oder Vorauswahl. Neben Piège kürten sie Alexandre Jean zum "Sommelier des Jahres [weiter...] 27.09.2011, ZEIT ONLINE

High Recall and High Precision with Morphological Analyzer

INTRAFIND

Suchergebnisse für "Buch"

Buch					Suchen
ALLE INHALTE	HEUTE	24 STUNDEN	7 TAGE	30 TAGE	
VOM		E	3I S		100
	Buch ALLE INHALTE VOM	Buch ALLE INHALTE HEUTE VOM	Buch ALLE INHALTE HEUTE 24 STUNDEN	Buch ALLE INHALTE HEUTE 24 STUNDEN 7 TAGE VOM BIS	Buch ALLE INHALTE HEUTE 24 STUNDEN 7 TAGE 30 TAGE VOM BIS

93445 ERGEBNISSE

Sortieren nach: RELEVANZ | DATUM

Filter: ALLE INHALTE NUR REZENSIONEN



POLITISCHE GEFANGENE Magischer Realismus

... Tagen hatte der alte Report des heute 84-jährigen Schriftstellers Platz eins auf Irans Bestsellerliste erklommen, wie die News-Webseite Aftab wissen ließ. Kurz darauf war das

Buch ausverkauft. Was kümmert die Teheraner urplötzlich das [weiter...] 18.10.2011, DIE ZEIT



Das deutsche Kapitel

...Heinrich August Winkler Geschichte des Westens Die Zeit der Weltkriege 1914-1945 Politisches **Buch** C.H. Beck München 2011 1350 38 Der Westen [weiter...]

18.10.2011, DIE ZEIT

Wirtschaftspolitik

... Praxis bedeutet. "Es gibt keine Inflationsgefahren" (DIE ZEIT, Nr. 25/2010) Jürgen Stark, Chefvolkswirt der Europäischen Zentralbank, soircht in einem Interview über die Rettung kriselnder Staaten, die Grenzen des Lehr**buch**wissens und Merkels Sparpaket [weiter...] 18.10.2011, ZEIT ONLINE



THINKTANKS

Politische Vordenker gesucht

... besänftigen konnte. Cross-over: Institut Solidarische Moderne Der Name ist trügerisch. Unter einem Institut stellt man sich etwas anderes vor: ein paar schicke Räume,

Bücher-Regale, Konferenztische und viele [weiter...] 18.10.2011, ZEIT ONLINE



ROMAN "GRUBER GEHT" Ein Mann, ein Krebs

... es dann eben doch: Er, das Testosteronbündel, lässt Trost und Nähe zu. Es sind vor allem die beiläufigen, absurden Momente, die das **Buch** anrührend machen. Wie Gruber,

schon vom Krebs gezeichnet, in seiner bislang unberührten Edelstahlküche [weiter...] 18.10.2011, ZEIT ONLINE



KOLONIALGE SCHICHTE

Schädel im Schrank

... Arzt und Rasseforscher Eugen Fischer hatte sie »bestellt«. Von der Exekution existieren Fotos; auch beschrieb der Kolonialarzt Wilhelm Wendland 1939 in seinem **Buch** Im

Wunderland der Papuas, wie er die Köpfe nach der Erschießung abschnitt [weiter...] 18.10.2011, DIE ZEIT



- 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)



Die <mark>Beiersdorf AG</mark> ist als Dachmarke Hersteller zahlreicher Markenprodukte, darunter Marken wie Nivea, Labello, Hansaplast, Futuro, Eucerin, Florena oder Tesa. Außerdem gehören zur Kosmetiksparte die Marken Juvena, la prairie of Switzerland, 8x4, atrix und die Haarpflegeprodukte von Marlies Möller. Weiterhin stellt Beiersdorf verschiedene Körperpflegeprodukte (Basis PH, Doppel Dusch, Gammon) her.

Der Hauptsitz befindet sich in Hamburg, weitere deutsche Standorte sind Baden-Baden, Berlin, Emmerich am Rhein, Hannover, Heitersheim, Offenburg und Waldheim. Der Standort in Wien wird weiter als Zentrale für Mittel- und Osteuropa ausgebaut.

Nivea ist eine geschützte Marke der Beiersdorf AG. Die 1911 auf den Markt gekommene Hautpflegecreme NIVEA Creme ist das bekannteste Produkt der Beiersdorf AG. Den Namen leitete Oscar Troplowitz vom lateinischen Adjektiv niveus (zu nix, nivis, Schnee) ab, er bedeutet "die Schneeweiße". Zuvor gab es bereits seit 1906 eine ebenfalls weiße Seife.

Zusammensetzung der Hautpflegecreme Grundlage war die Entdeckung von Eucerit, einem aus Schafswollfett gewonnenen Emulgator, dem ersten Wasser-in-Öl-Emulgator. 1911 entwickelte der Besitzer von Beiersdorf, Oscar Troplowitz, eine Hautcreme in enger Zusammenarbeit mit dem Chemiker Isaac Lifschütz und dem Dermatologen Paul Gerson Unna. Im Dezember desselben Jahres kam die erste Hautcreme der Welt mit langanhaltender Wirkung auf den Markt. Die Rezeptur ist seit den Anfangstagen nahezu unverändert geblieben: unter anderem Glyzerin, Panthenol, Zitronensäure, Wasser, Emulgator (Eucerit) und Duftstoffe. ✓
 Cat_ADJECTIVE
 Cat_NOUN
 Cat_VERB
 ✓ City
 ✓ Organization
 ✓ Produkt
 ✓ PersonName
 Location
 OrgCompanyTemp
 Org_Company
 ✓ ProduktMarke
 ✓ Rohstoff
 ✓ date

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?

Semantic Search: Comparison with standard Search Engines



Frage: Wo liegen Werke von Audi?

NL-Search

Audi.

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

Audi - Wikipedia

Google

Die Fahrzeuge der Marke Audi werden außer in den beiden deutschen Werken Ingolstadt und ... 1915 wurde die "Audi Werke AG" gegründet. Nachdem Audi 1928 in ...

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 Györ an die Kapazitätsgrenzen. Wie FOCUS- MONEY erfuhr, ...

INTRAFIND

... höre" ins Lateinische "Audi". Im Juli 1910 verließ das erste Fahrzeug mit dem Namen Audi das Zwickauer <mark>Werk</mark>. 1915 wurde die "Audi Werke</mark> AG" gegründet. Nachdem Audi 1928 ...

Semantic Search: Comparison with standard Search Engines



Frage:

Wer hat Microsoft gegründet?

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 ...

INTRAFIND

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



atei Bearbeiten Ansicht F Google Where was E	Favoriten Extras ? Instein born	Übersetzen 🔹 🥜 🔔 Where 🥞 was 🏝 Einstein 🧟 born Mehr	» Anmelden 🌂
+You Search Ima	ges Maps Play YouTube News Gmail Documents Calendar More≁		
Google	Where was Einstein born	٩	Sign in
Search	About 35,200,000 results (0.14 seconds)		\$
Web	Ulm	UIm	
Images	Albert Einstein, Place of birth	Ulm is a city in the federal German state of Baden-Wür	ttemberg,
Maps	Albert Finstein Wikingdia the free encyclopedia	situated on the River Danube. The city, whose populati at 120,000, forms an urban district of its own and is the	on is estimated administrative
Videos	en.wikipedia.org/wiki/Albert_Einstein	seat of the Alb-Donau district. Wikipedia	
News	Albert E instein 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	Population: 121,648 (2008)	
Shopping	Brain - Religious views of Albert General relativity - Einstein family	Founded: 1181	
More	Albert Einstein - Biography	Area: 118.7 km²	
	www.nobelprize.org > Nobel Prizes > Nobel Prize in Physics	Local time: 1:16pm Tuesday (CEST)	
Show search tools	Biography. Albert Linstein was born at Ulm , in Wurttemberg, Germany, on March 14, 1879. Six weeks later the family moved to Munich, where he later on began		Feedback
	40 Character Frank in Nacharacter		
	www.neatorama.com/2007/03/26/10-strange-facts-about-einstein/		
	26 Mar 2007 – When Albert's mother, Pauline Einstein gave birth to him, she thought that Einstein's head was so his and misshapen that he was deformed		
	and Enstein a near was so by and missingpen that he was reformed.		
	When and where was Albert Einstein born and when did he die		
	Wiki answers.com > > Pamous 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,		

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 ...

iFinder

Index

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

INTRAFIND



Semantic Search for the Enterprise

Questions?

Dr. Christoph Goller Director Research

Phone: +49 89 3090446-0

- Fax: +49 89 3090446-29
- Email: <u>christoph.goller@intrafind.de</u>
- Web: <u>www.intrafind.de</u>

IntraFindSoftware AG Landsberger Straße 368 80687 München Germany



