

Information Retrieval – begleitendes Tutorium II

Thomas Schmidt

Rückblick

→ Lösungen zur Übungsaufgabe online im Grips

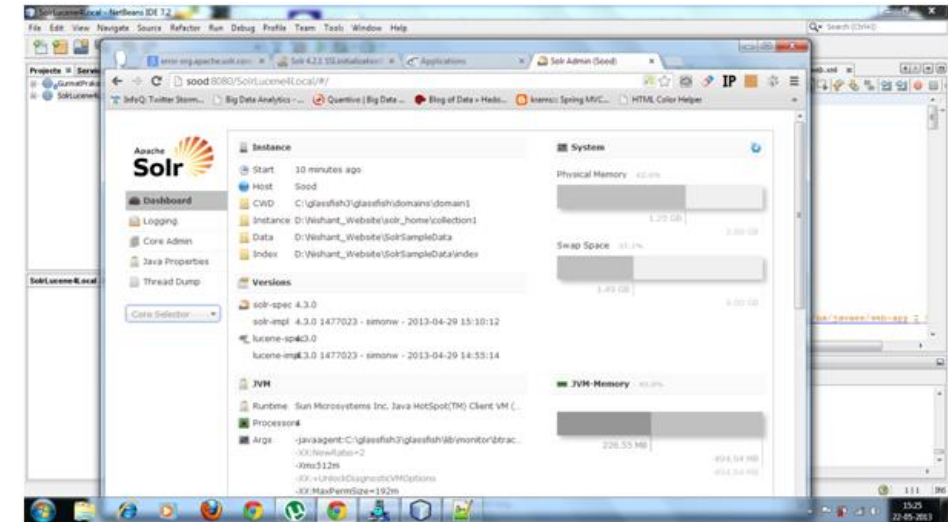


Heute: Solr



Allgemeines

- Standalone Enterprise Search Server
- Indexierung und Query mit XML, JSON oder Binary über HTTP
- flexibel und anpassbar
- HTML-Admin-UI
- nützliche integrierte Technologien → DataImportHandler, SolrItas, Ajax Solr, usw.
- viele nützliche Features: Facetten, Autocomplete, Highlighting, mächtige Query-Syntax
- basiert auf Lucene



Installation

- Solr-4.8.0.zip von runterladen (Link im Grips oder googeln) und entpacken
- über Ausführen – cmd: Konsole öffnen
- Navigation zum solr-4.8.0 – Ordner über cd-Befehl
- zu example-Ordner navigieren, also cd example
- „java -jar start.jar“ eingeben
- Solr nun erreichbar unter: <http://localhost:8983/solr>

Quick tour of Solr



Solr - schema.xml

- Zu finden unter: `solr-4.8.0/example/solr/collection1/conf/schema.xml`
- beliebig anpassbar
- grundlegende Konfiguration
- Konfiguration der Datenfelder
- Definition von Datenfeldtypen

Datenfeldtypen

- besitzen einen Namen
- besitzen eine Java-Klasse, die sie implementiert
- werden später referenziert bei der eigentlichen Felddefinition

```
<fieldType name="string" class="solr.StrField" sortMissingLast="true" />  
  
<!-- boolean type: "true" or "false" -->  
<fieldType name="boolean" class="solr.BoolField" sortMissingLast="true"/>
```

```
<fieldType name="managed_en" class="solr.TextField" positionIncrementGap="100">  
  <analyzer>  
    <tokenizer class="solr.StandardTokenizerFactory"/>  
    <filter class="solr.ManagedStopFilterFactory" managed="english" />  
    <filter class="solr.ManagedSynonymFilterFactory" managed="english" />  
  </analyzer>  
</fieldType>
```


Attribute von Datenfeldtypen – eine Auswahl

Attribut	Beschreibung	Wert
indexed	„true“: Wert des Feldes kann in Querys für das Retrieval von Dokumenten genutzt werden	true or false
stored	„true“: Wert des Feldes kann über Querys abgerufen werden	true or false
sortMissingFirst sortMissingLast	Kontrolle über die Platzierung von Dokumenten falls kein Feld vorhanden	true or false
multiValued	„true“: Einzelnes Dokument kann mehrere Werte für diesen Datenfeldtypen besitzen	true or false
positionIncrementGap	„100“: Verhindert falsches Phrasematching über mehrere Felder hinweg	integer
omitNorms	„true“: Norms werden nicht gespeichert	true or false
omitTermFreqAndPositions	„true“: Term Frequenz und Position im Text werden nicht gespeichert (default für non-Text)	true or false
autoGeneratePhraseQueries	„true“: Solr erstellt automatisch PhraseQuerys für benachbarte Terme (ansonsten “ “)	true or false

Feldtypen nach Anwendungsfall

Use Case	indexed	stored	multiValued	omitNorms	termVectors	termPositions
search within field	true					
retrieve contents		true				
use as unique key	true		false			
sort on field	true		false	true [1]		
use field boosts [5]				false		
document boosts affect searches within field				false		
highlighting	true [4]	true			[2]	true [3]
faceting [5]	true					
add multiple values, maintaining order			true			
field length affects doc score				false		
MoreLikeThis [5]					true [6]	

Textanalyse

Verarbeitet Textstream, erstellt Tokenstream
Chaining möglich

```
<fieldType name="managed_en" class="solr.TextField" positionIncrementGap="100">  
  <analyzer>  
    <tokenizer class="solr.StandardTokenizerFactory"/>  
    <filter class="solr.ManagedStopFilterFactory" managed="english" />  
    <filter class="solr.ManagedSynonymFilterFactory" managed="english" />  
  </analyzer>  
</fieldType>
```

Verarbeitet Tokenstream, erstellt neuen
Tokenstream
Chaining möglich

```
<filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
<filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>
<filter class="solr.LowerCaseFilterFactory"/>
```

```
<fieldtype name="phonetic" stored="false" indexed="true" class="solr.TextField" >
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.DoubleMetaphoneFilterFactory" inject="false"/>
  </analyzer>
</fieldtype>
```

```
<filter class="solr.PatternReplaceFilterFactory"
  pattern="([a-z])" replacement="" replace="all"
/>
```

Felder

```
<field name="sku" type="text_en_splitting_tight" indexed="true" stored="true" omitNorms="true"/>
<field name="name" type="text_general" indexed="true" stored="true"/>
<field name="manu" type="text_general" indexed="true" stored="true" omitNorms="true"/>
<field name="cat" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="features" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="includes" type="text_general" indexed="true" stored="true" termVectors="true" termPositions="true"
  termOffsets="true" />

<field name="weight" type="float" indexed="true" stored="true"/>
<field name="price" type="float" indexed="true" stored="true"/>
<field name="popularity" type="int" indexed="true" stored="true" />
<field name="inStock" type="boolean" indexed="true" stored="true" />
```

Copyfields und dynamicfields

Copyfield:

- Inhalt von einem Feld wird in den Index eines anderen kopiert
- gleiches Feld in verschiedenen Indizes oder mehrere Felder in eines
- Sortieren, Facetten

Dynamicfield:

- dienen dazu Felder zu indexieren, die nicht explizit im Schema enthalten sind
- Name mit Wildcard (*)

Schritte für das Schema-Design

1. Welche Art von Suchmöglichkeiten werden unterstützt
2. Welche Art von Entitäten sollen von der Suche zurückgeliefert werden
3. Denormalisierung in Beziehung stehender Daten
 - Redundantes abspeichern der Daten für schnellen Zugriff
 - 1:1 Relation Bsp: Namensfeld für Künstler, Album etc.
 - 1:n Relation Feld mit mehreren Werten: multiValued="true"



Laden der Beispieldaten

- cd example/exampledocs
- java -jar post.jar *.xml
- Iteration mittels SimplePostTool (HTTP – POST)

```
<add><doc>
  <field name="id">3007WFP</field>
  <field name="name">Dell Widescreen UltraSharp 3007WFP</field>
  <field name="manu">Dell, Inc.</field>
  <!-- Join -->
  <field name="manu_id_s">dell</field>
  <field name="cat">electronics and computer1</field>
  <field name="features">30" TFT active matrix LCD, 2560 x 1600,
  | .25mm dot pitch, 700:1 contrast</field>
  <field name="includes">USB cable</field>
  <field name="weight">401.6</field>
  <field name="price">2199</field>
  <field name="popularity">6</field>
  <field name="inStock">true</field>
  <!-- Buffalo store -->
  <field name="store">43.17614, -90.57341</field>
</doc></add>
```


Suche in Solr

- Vielfältige Query-Anfragen
- Beispiel eines Response-Headers:

```
<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">21</int>
  <lst name="params">
    <str name="facet">true</str>
    <str name="indent">true</str>
    <str name="facet.query">dell</str>
    <str name="q">dell</str>
    <str name="_">1399149152571</str>
    <str name="wt">xml</str>
  </lst>
</lst>
```

Request-Handler (qt)

/select

— common —

q

dell

fq

sort

start, rows

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

xml ▼

☒ indent

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☐ facet

☐ spatial

☐ spellcheck

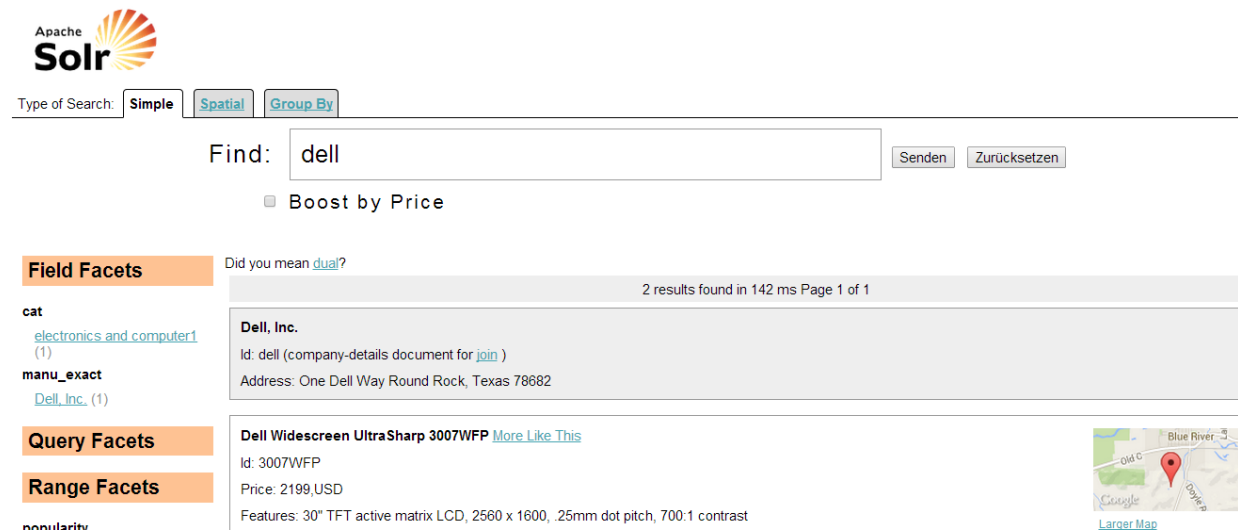
Resultat

```
<result name="response" numFound="1" start="0">
  <doc>
    <str name="id">3007WFP</str>
    <str name="name">Dell Widescreen UltraSharp 3007WFP</str>
    <str name="manu">Dell, Inc.</str>
    <str name="manu_id_s">dell</str>
    <arr name="cat">
      <str>electronics and computer1</str>
    </arr>
    <arr name="features">
      <str>30" TFT active matrix LCD, 2560 x 1600, .25mm dot pitch, 700:1 contrast</str>
    </arr>
    <str name="includes">USB cable</str>
    <float name="weight">401.6</float>
    <float name="price">2199.0</float>
    <str name="price_c">2199,USD</str>
    <int name="popularity">6</int>
    <bool name="inStock">true</bool>
    <str name="store">43.17614,-90.57341</str>
    <long name="_version_">1467113110701604864</long></doc>
</result>
```

Ein einfaches Search-UI - Solritas

<http://localhost:8983/solr/browse>

- zahlreiche Features (Autocomplete, More-Like-This, Facettierte Suche, Rechtschreibprüfung usw.)
- UI anpassbar (→ siehe kommende Sitzungen)



The screenshot displays the Apache Solr browse interface. At the top, the Apache Solr logo is visible. Below it, the 'Type of Search' section includes buttons for 'Simple', 'Spatial', and 'Group By'. The search bar contains the text 'dell', with 'Find:' to its left and 'Senden' and 'Zurücksetzen' buttons to its right. Below the search bar, there is a checkbox for 'Boost by Price'. The main content area shows search results. On the left, there are 'Field Facets' for 'cat' (electronics and computer1 (1)) and 'manu_exact' (Dell, Inc. (1)). Below these are 'Query Facets' and 'Range Facets'. The main results area shows two results. The first result is for 'Dell, Inc.' with details: 'Id: dell (company-details document for join)' and 'Address: One Dell Way Round Rock, Texas 78682'. The second result is for 'Dell Widescreen UltraSharp 3007WFP' with details: 'Id: 3007WFP', 'Price: 2199,USD', and 'Features: 30" TFT active matrix LCD, 2560 x 1600, .25mm dot pitch, 700:1 contrast'. A 'More Like This' link is provided for the second result. A small map is visible in the bottom right corner of the results area.

Ausblick

In 2 Wochen:

- Solr: config.xml
- DataImportHandler und Indexierung
- Spezielle Features wie Autocomplete, Facetten, Did you mean, Spellchecker etc.

Nächste Woche: Methoden für die
Anforderungsanalyse → Präsentation des UI

