A Brief Introduction to

@蒋锴_USTC

2012-11-24

Outline

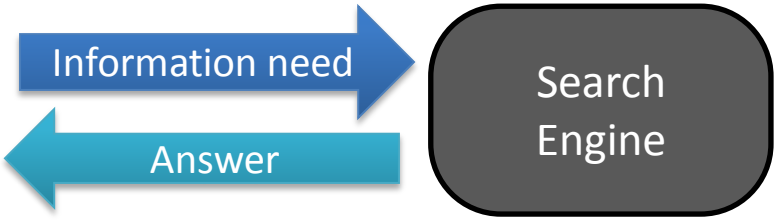
- Backgrounds
  - Search and search server/service
- Overview
  - Concepts and features
- Solr in Action
  - Deploy solr
  - Index and query a sample movie database
- p.s.
  - Beyond basics, falltrap and practice
  - reference

Backgrounds

Search, search, search

- Search is everywhere.

Various modality



Doesn't always require explicit input

Search server, why?

- Share across multiple application and reduce overhead.
- Transparent to applications' development and deployment.
- Benefit performance. Caching, pre-warming, load balance, HA, etc.

RDBMS is not designed for searching

- `like '%'` can be slow, even with index
- Especially when update operations are performed frequently

Options

- Google Search Appliance

  - Google's secret recipe
  - Software & hardware in one box
  - expensive

- Solr

Apache Solr

  - Mature, active community, lots of packages and docs
  - Free, lucene based

- Sphinx

Sphinx

  - Fast, written in c++
  - Simple, but lack of fancy features, eg. Facet, word/pdf parsing support, etc
  - GPL

- Elastic search

elasticsearch.

  - Born with "cloud" in mind
  - Free, lucene based

Why Solr?

- Free
- Powerful
- Mature

- After this talk, you will have your own answer.

Overview

What is Solr

- Solr is an open source search engine managed as part of the Apache Lucene project. Lucene is a core Java library for building search applications. Solr is based on Lucene and was originally developed at CNET for use on websites such as CNET Reviews.

- Who use Solr?

  http://wiki.apache.org/solr/PublicServers

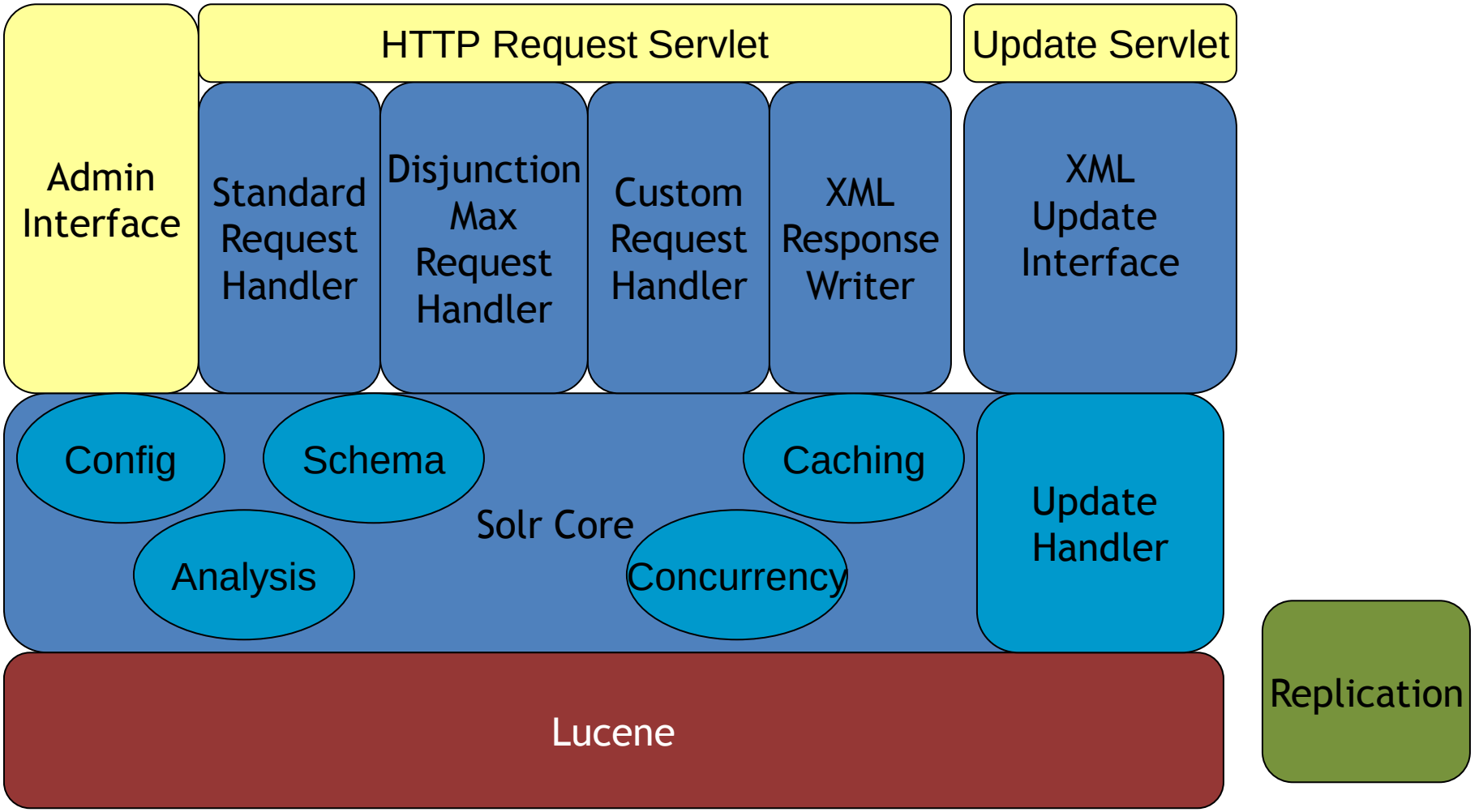
  - Whitehouse
  - Instagram
  - FCC
  - AOL
  - AT&T
  - Netflix
  - CNET
  - SourceForge
  - Digg
  - JSTOR

  - And more…

Features

- A full text search server based on Lucene
- XML/HTTP Interfaces
- Loosely coupled document model
- Flexible query syntax
- Import kinds of data in kinds of ways
- Faceted Browsing
- Highlighting
- Spell check
- “More like this”
- Web Administration Interface
- Extensive Caching
- HA, Index Replication

Architecture



ref: Yonik Seeley yonik@apache.org

XML/HTTP Interfaces

- **XML/HTTP**

  easily adapt to different platforms, devices and programming languages

```
Eg.
curl -s http://localhost:8983/solr/update --data-binary
'XMLContent' -H 'Content-type:text/xml; charset=utf-8'
```

- Add doc

```
<add>
    <doc>
      <field name="id">230012</field>
      <field name="title">功夫</field>
      <field name="genre">喜剧</field>
      <field name="actors">周星驰，黄圣依
       </field>
    </doc>
</add>
```

- Delete
  - Delete by Id

```
<delete><id>05591</id></delete>
```

  - Delete by Query

```
<delete>
   <query>actors:bruces</query>
</delete>
```

XML/HTTP Interfaces

- <commit/> makes changes visible
  - closes IndexWriter
  - removes duplicates
  - opens new IndexSearcher
    - newSearcher/firstSearcher events
    - cache warming
    - “register” the new IndexSearcher
- <optimize/> same as commit, merges all index segments.
- <rollback/> to last commit point
- Search results is XML too. Or, json, other formats through xslt

Solr Document Model: preliminary

Document → Term → Inverted index

IR is finding material of an unstructured nature, that satisfies an information need from within large collections.

↓ Tokenization

IR/is/finding/material/of/an/unstructured/nature/that/satisfies/an/information/need/from/within/large/collections

↓ remove stop words

IR/finding/material/unstructured/nature/satisfies/information/need/large/collections

↓ normalization

information/retrieval/finding/material/unstructured/nature/satisfies/information/need/large/collections

↓ stemming

information/retrieval/find/material/structure/nature/satisfy/information/need/large/collection <- Term

for more: http://www-nlp.stanford.edu/IR-book/

Solr Document Model: preliminary

- Inverted Index

DocumentId	Term
001	election win politic
002	nba win game
003	play game xbox

Term	DocumentId
election	001
game	002,003
nba	002
play	003
politic	001
win	001,002
xbox	003

Solr Document Model

Document

bal_id	inner_id	title	alias	director	actors	genre	rating	website_url	image
95276	13358	残梦		杨紫婷	杨紫婷,孙文婷,霍政谚,陈	同性	0	http://movie.douban.con	s6907514.jpg
95277	13359	灵异女孩月子 201		小川通仁	相武紗季,塚本高史,石垣(	恐怖	6.2	http://movie.douban.con	s6917025.jpg
95278	13360	Larva		Joo-Gong Meang		动画,短片	9	http://movie.douban.con	s6916993.jpg

Document is consisted by fields.

Schema.xml

```
<fields>
   <field name="title" type="text_cn" indexed="true" stored="true"/>
   <field name="alias" type="text_cn" indexed="true" stored="true"/>
   <field name="director" type="text_cn" indexed="true" stored="true"/>
   <field name="actors" type="text_cn" indexed="true" stored="true"/>
   <field name="genre" type="text_cn" indexed="true" stored="true"/>
   <field name="rating" type="float" indexed="false" stored="true"/>
   <field name="play_url" type="string" indexed="false" stored="true"/>
   <field name="website_url" type="string" indexed="false" stored="true"/>
   <field name="image" type="string" indexed="false" stored="true"/>
   <field name="source" type="string" indexed="false" stored="true"/>
  <field name="id" type="string" indexed="true" stored="true" required="true" />
  <field name="inner_id" type="string" indexed="true" stored="true"/>
  </fields>
```

NOTE:
Solr instance / Core
One core, one *"table"*. A core can only index one kind of document.

Field is defined by FieldType,
and fieldType determines how the field will be processed(tokenize, stem, filter, etc.)

```
  <fieldType name="text_cn" class="solr.TextField" positionIncrementGap="100">
       <analyzer>
           <tokenizer class="com.chenlb.mmseg4j.solr.MMSegTokenizerFactory" mode="complex"
dicPath="C:\xampp\tomcat\webapps\videosearch\mmseg4j\data"/>
           <filter class="solr.LowerCaseFilterFactory"/>
       </analyzer>
   </fieldType>
```

Powerful query syntax

- Query arguments for HTTP GET/POST to /select
  - “q” the query
  - “start” (0) offset
  - “rows” (10) number of docs
  - “fl” (*) fields to return
  - “qt” (standard) query type, maps to query handler
  - “df” (schema) default field to search
  - “wt” writer type (response format)
  - Include (+), exclude (-)
  - Field-specific searching: <fieldname>:<fieldvalue>
  - Wildcard searching: “*” or “?”
  - Range searching: Timestamp:[2006-01-01 TO *]
  - Proximity searching: “~”: “video ipod”~3 (up to 3 words apart)
  - Fuzzy searches: Belkin~0.8 (will find words close spellings)

Multiple ways to import data

- Import records from a database using the Data Import Handler (DIH).
- Load a CSV file (comma separated values), including those exported by Excel or MySQL.
- POST JSON/XML documents
- Index binary documents such as Word and PDF with Solr Cell (ExtractingRequestHandler).
- Use SolrJ for Java or other Solr clients to programatically create documents to send to Solr

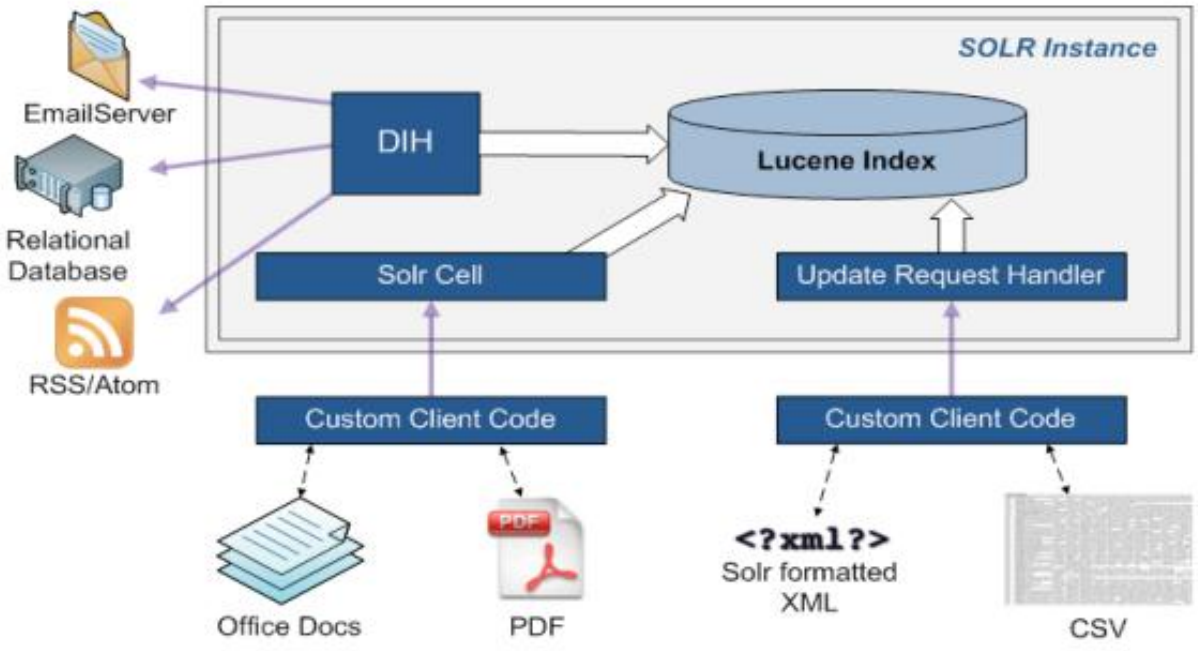
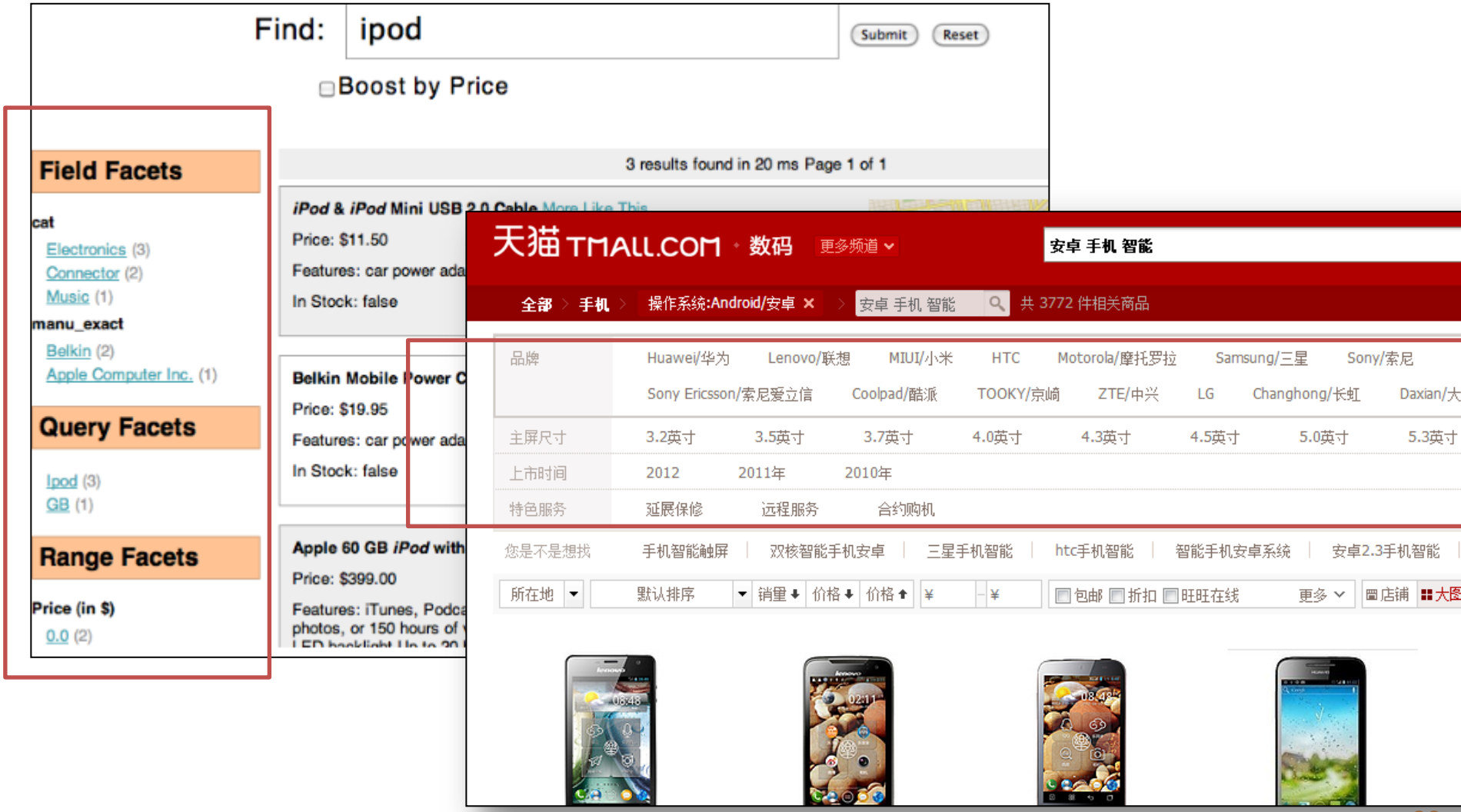


Illustration: SolrRecipes by Lucid Imagination

Faceted Browsing

widely used in E-commerce

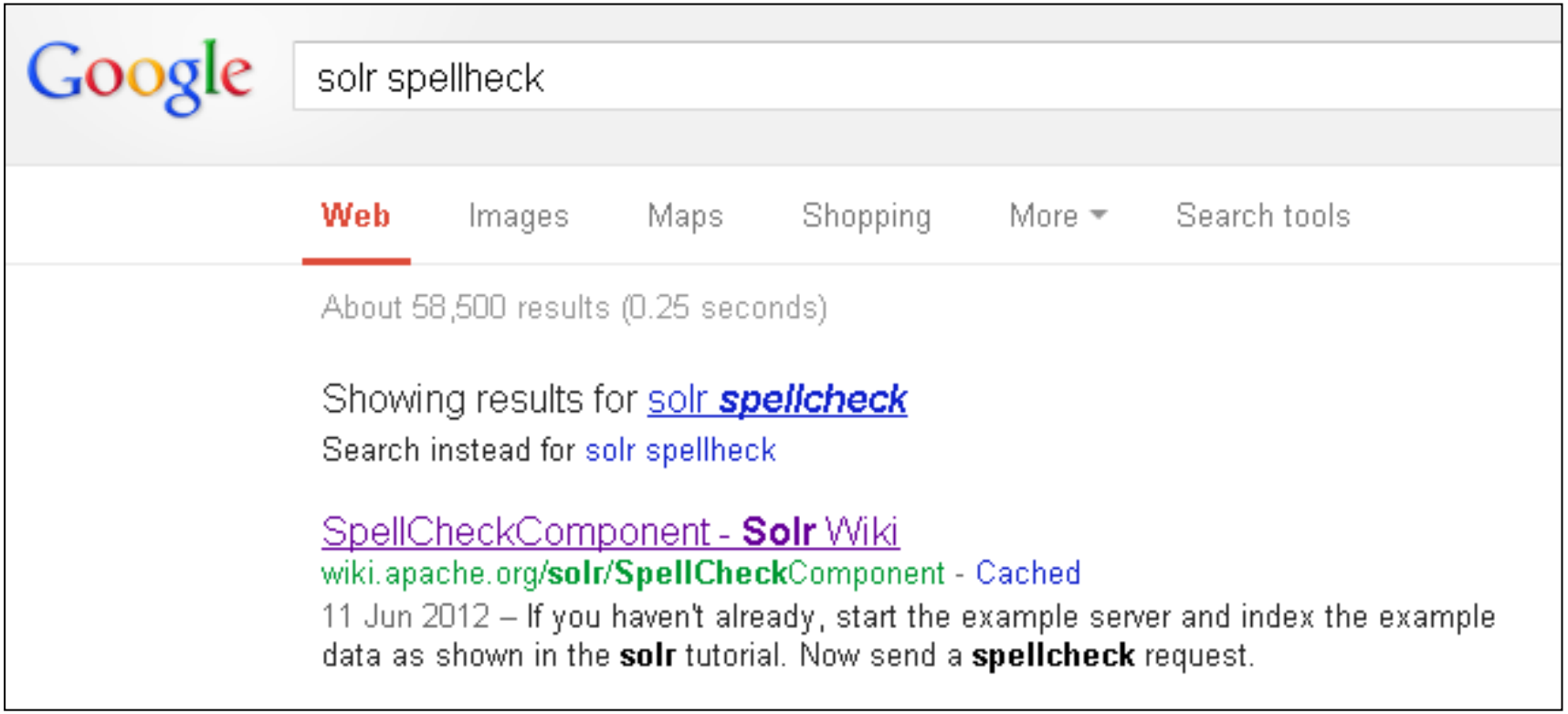


Highlighting

- Only text that has been both indexed and stored may be highlighted.
- Eg.
  - http://localhost:8081/videosearch/movieindex/select/?q=title:爱情&hl=true&hl.fl=title

For more highlighting parameters, check lucidworks.lucidimagination.com/display/solr/Highlighting

Spell check



- howto: http://searchhub.org/2010/08/31/getting-started-spell-checking-with-apache-lucene-and-solr/

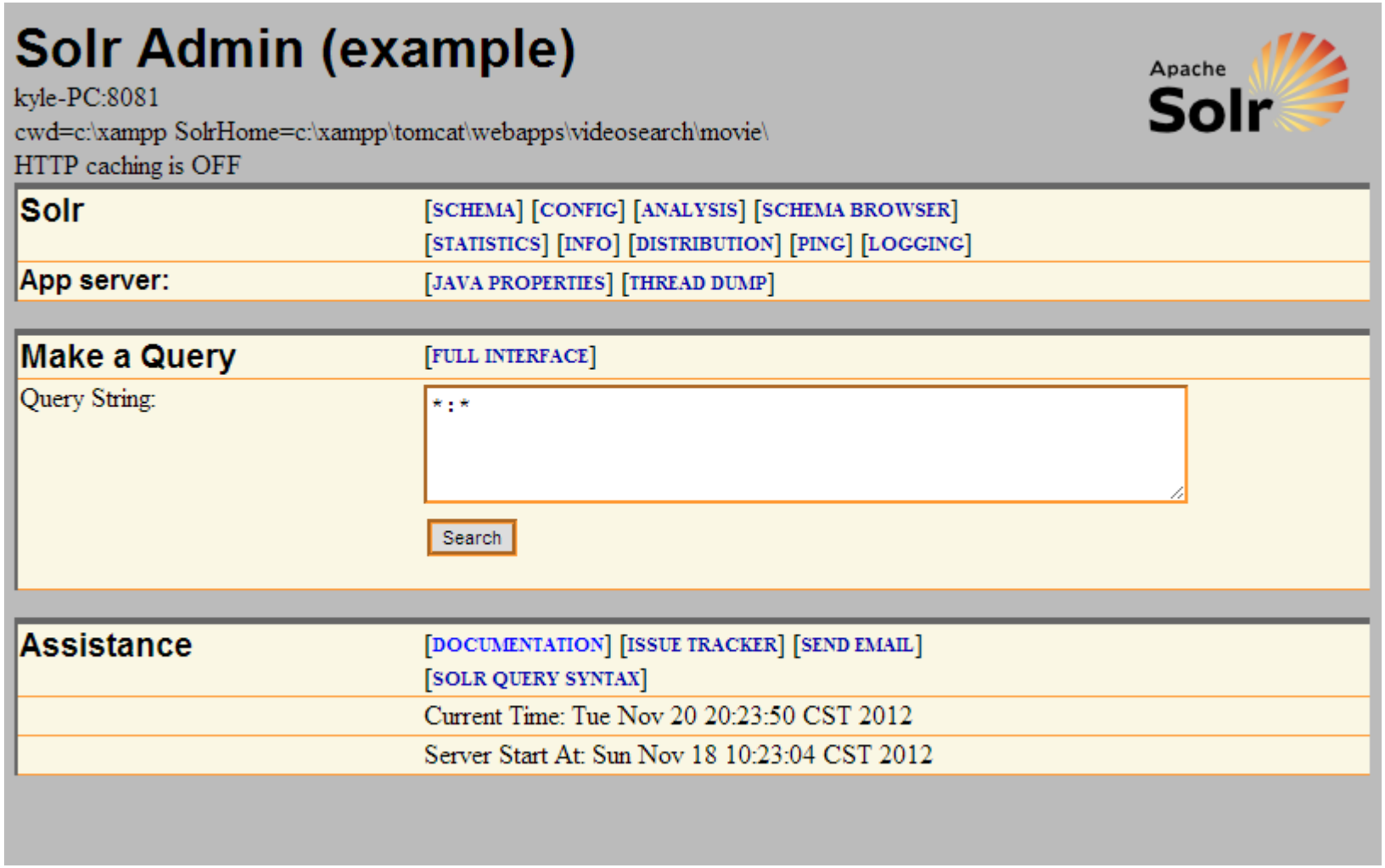
MoreLikeThis

- Basically, you will find it calculates similarity of documents using vector space model, if you know about NLP or IR
- schema.xml

```
<field name="cat" ... termVectors="true" />
```

- simple tutorial http://blog.brattland.no/node/18

Web Admin Interface



Web Admin Interface

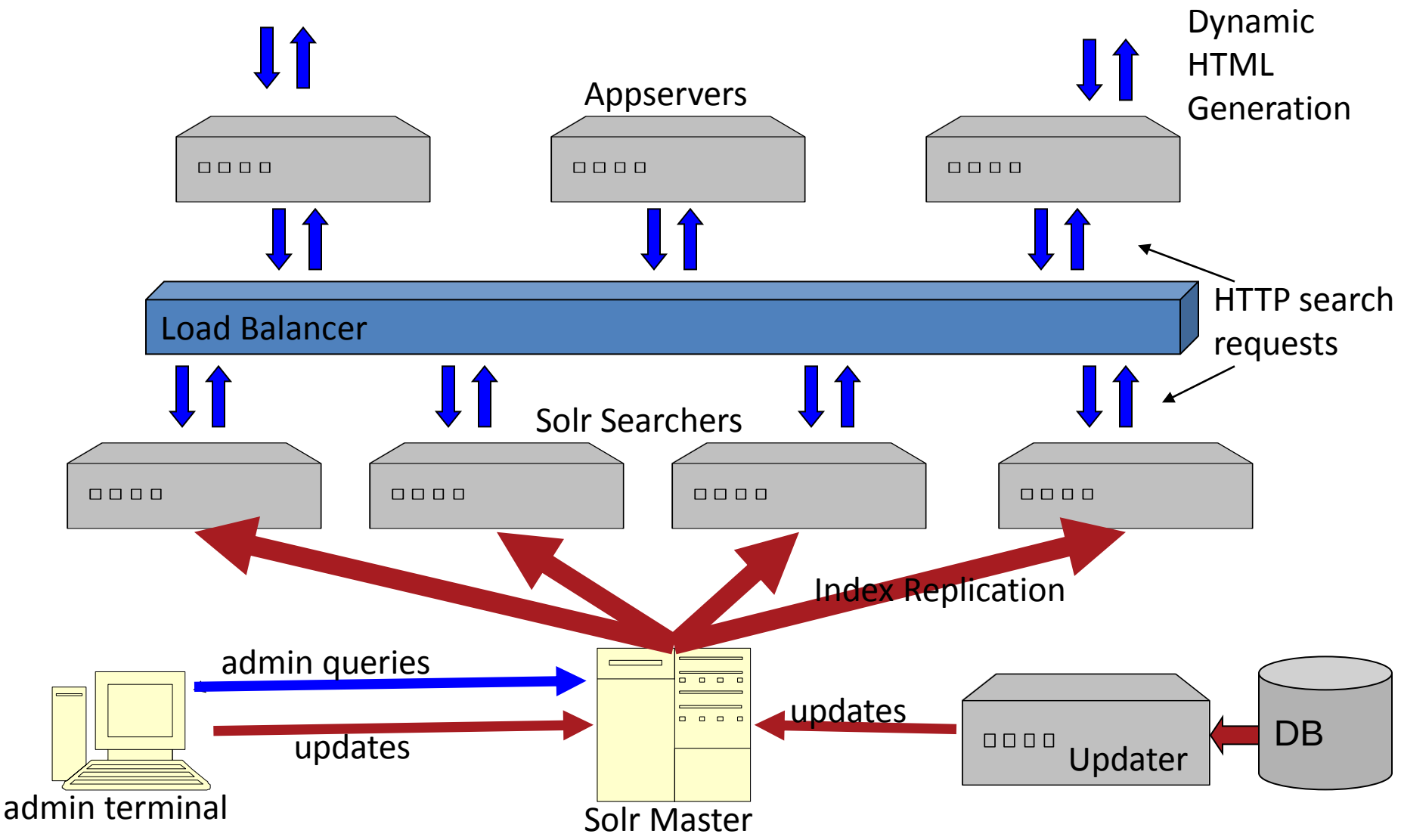
- Show Config, Schema, Distribution info
- Query Interface
- Statistics
  - Caches: lookups, hits, hitratio, inserts, evictions, size
  - RequestHandlers: requests, errors
  - UpdateHandler: adds, deletes, commits, optimizes
  - IndexReader, open-time, index-version, numDocs, maxDocs,
- Analysis Debugger
  - Shows tokens after each Analyzer stage
  - Shows token matches for query vs index

Caching

- Extensive caching
  - filterCache
  - fieldValueCache
  - queryResultCache
  - documentCache
  - User/Generic Caches
- Cache Warming and Autowarming

read more: http://wiki.apache.org/solr/SolrCaching

High Availability



Good News

- You can build yourself a search server with all these amazing features without writing any "*code*."
- All you need to do is to configure some xml files.

Why Solr?

- Shouldn't be a question
- Let there be light. Let there be solar

Solr in Action

Deploy

- Put *.war* into tomcat *webapps* directory
- Configure webapp: **`web.xml`**

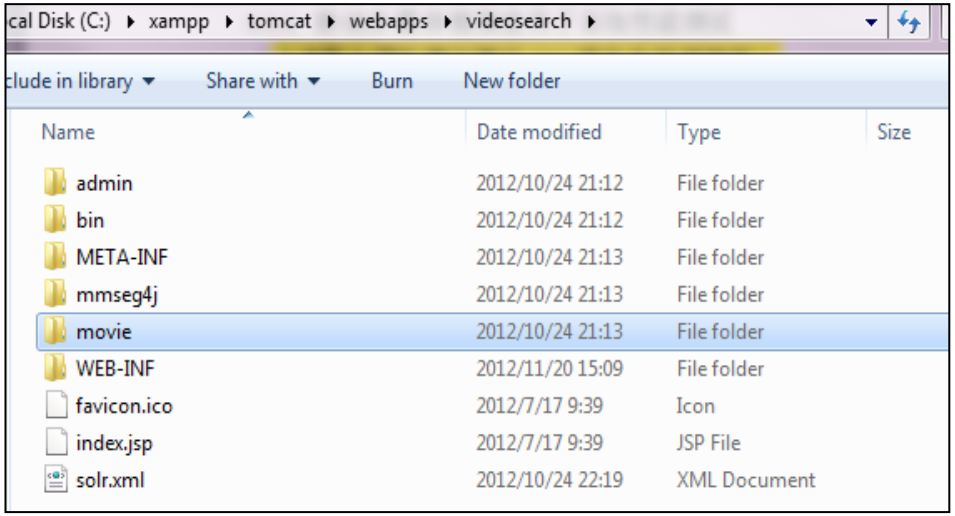
```
<env-entry>
   <env-entry-name>solr/home</env-entry-name>
   <env-entry-value>c:\xampp\tomcat\webapps\videosearch</env-entry-value>
   <env-entry-type>java.lang.String</env-entry-type>
</env-entry>
```

Deploy

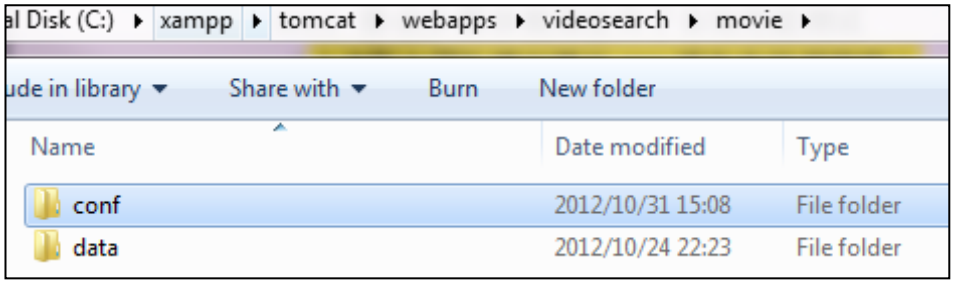
- Configure solr cores: **`solr.xml`**

```
<solr persistent="false">
  <cores adminPath="/admin/cores">
    <core name="movieindex" instanceDir="movie"/>
  </cores>
</solr>
```

Can add more core other than "movieindex"



Inside **Solr Home**



To be more general, **instanceDir** is used to represent "movie" hereafter

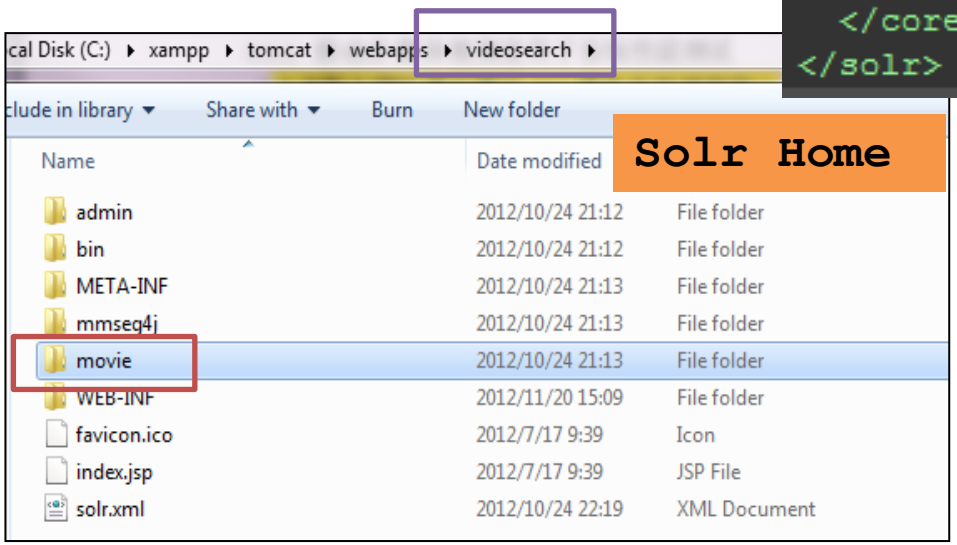
Inside **InstanceDir**

Deploy

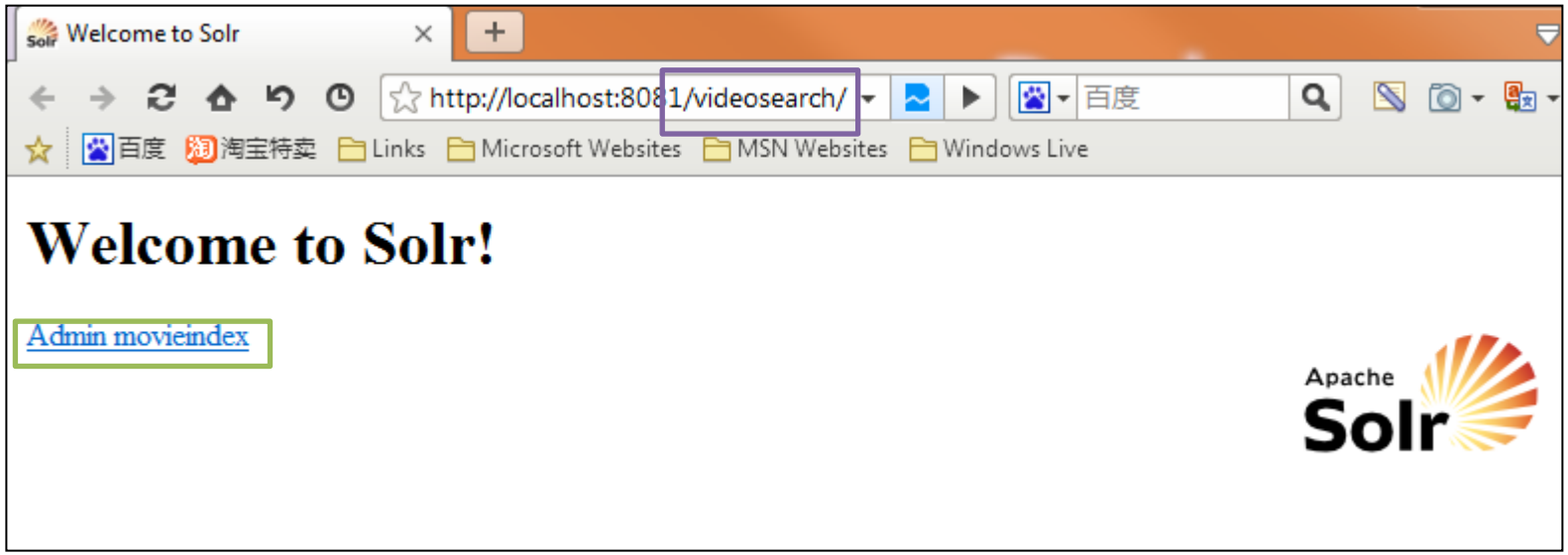
- Visit webapp:

```
<solr persistent="false">
  <cores adminPath="/admin/cores">
    <core name="movieindex" instanceDir="movie"/>
  </cores>
</solr>
```

solr.xml



Solr Home



Import data through DIH

- Request handler
  - Configure **`solrconfig.xml`** in `instancedir/conf`
  - Add a requestHandler

```
<requestHandler name="/dataimport" class="org.apache.solr.handler.dataimport.DataImportHandler">
 <lst name="defaults">
   <str name="config">data-config.xml</str>
 </lst>
 </requestHandler>
```

Import data through DIH

- Sample data: A movie database

global_id	inner_id	title	alias	director	actors	genre	rating	website_url	image	play_url	source
95276	13358	残梦		杨紫婷	杨紫婷,孙文婷,霍政谚,陈	同性	0	http://movie.douban.con	s6907514.jpg		douban
95277	13359	灵异女孩月子 201		小川通仁	相武紗季,塚本高史,石垣(	恐怖	6.2	http://movie.douban.con	s6917025.jpg		douban
95278	13360	Larva		Joo-Gong Meang		动画,短片	9	http://movie.douban.con	s6916993.jpg		douban
95279	13361	女性癌者 The Ny		拉斯·冯·提尔			0	http://movie.douban.con	s/movie-default-medi		douban
95280	13362	女朋友 * 男朋友		杨雅喆	桂纶镁,张孝全	爱情	0	http://movie.douban.con	s/movie-default-medi		douban
95281	13363	UN-GO剧场版 UI	UN-GO剧场	水岛精二	胜地凉,丰崎爱生,山本希	动画	0	http://movie.douban.con	s6918552.jpg		douban
95282	13364	死党 ผีกู เพื่อนกัน	Friends Never		马里奥·毛瑞尔,Natcha Ja		0	http://movie.douban.con	s6920606.jpg		douban
95283	13365	超市试吃				喜剧,短片	4.4	http://movie.douban.con	s6927343.jpg		douban
95284	13366	气球冷藏				喜剧,短片	6.2	http://movie.douban.con	s6927345.jpg		douban
95285	13367	剪刀手				喜剧,短片	6.5	http://movie.douban.con	s6927389.jpg		douban
95286	13368	民的1911		鲁艺		剧情,历史,动画	0	http://movie.douban.con	s6927432.jpg		douban
98302	tt0004465	The Perils of Paul		Louis J. Gasnier,Dor	Pearl White,Crane Wilbu	Action	0	www.imdb.com/title/tt00			imdb
98303	tt0006206	Les vampires		Louis Feuillade	Musidora,Édoua	Action,Adventure,(	0	www.imdb.com/title/tt00	action/tt0006206.jpg		imdb
98304	tt0006333	20,000 Leagues U		Stuart Paton	Allen Holubar,Curtis Ben	Action,Adventure	0	www.imdb.com/title/tt00	action/tt0006333.jpg	http://player.yo	imdb
98305	tt0009682	Tarzan of the Ape		Scott Sidney	Elmo Lincoln,Enid Marke	Action,Adventure	0	www.imdb.com/title/tt00	action/tt0009682.jpg		imdb
98306	tt0011036	Bullet Proof		Lynn Reynolds	Harry Carey,William Ryn	Action,Western	0	www.imdb.com/title/tt00			imdb
98307	tt0012752	The Three Muske		Fred Niblo	Douglas Fairbanks,L&ea	Action,Adventure,F	0	www.imdb.com/title/tt00	action/tt0012752.jpg		imdb
98308	tt0014945	Girl Shy		Fred C. Newmeyer,S	Harold Lloyd,Jobyna Ral	Action,Comedy,Ro	0	www.imdb.com/title/tt00	action/tt0014945.jpg		imdb
98309	tt0016634	Beau Geste		Herbert Brenon	Ronald Colman,Neil Har	Action,Adventure,I	0	www.imdb.com/title/tt00			imdb
98310	tt0016641	Ben-Hur: A Tale (		Fred Niblo, &8	Ramon Novarro,Francis J	Action,Adventure,I	0	www.imdb.com/title/tt00	action/tt0016641.jpg		imdb
98311	tt0016654	The Black Pirate		Albert Parker	Douglas Fairbanks,Billie I	Adventure,Action	0	www.imdb.com/title/tt00	action/tt0016654.jpg		imdb
98312	tt0017925	The General		Clyde Bruckman,Bu	Buster Keaton,Marion M	Comedy,Romance,	0	www.imdb.com/title/tt00	action/tt0017925.jpg	http://player.yo	imdb
98313	tt0018578	Wings		William A. Wellman	Clara Bow,Richard Arlen	Drama,Romance,W	0	www.imdb.com/title/tt00	action/tt0018578.jpg		imdb
98314	tt0019412	Speedy		Ted Wilde	Harold Lloyd,Ann Christy	Action,Comedy,Fa	0	www.imdb.com/title/tt00	action/tt0019412.jpg		imdb
98315	tt0019421	Steamboat Bill, Jr		Charles Reisner,&nb	Buster Keaton,Tom McG	Action,Comedy,Dr	0	www.imdb.com/title/tt00	action/tt0019421.jpg	http://player.yo	imdb
98316	tt0020570	Die weiße H		Arnold Fanck,Georg	Gustav Diessl,Leni Riefen	Action,Adventure,I	0	www.imdb.com/title/tt00	action/tt0020570.jpg		imdb
98317	tt0020629	All Quiet on the V		Lewis Milestone	Lew Ayres,Louis Wolheir	Action,Drama,Hist	0	www.imdb.com/title/tt00	action/tt0020629.jpg		imdb
98318	tt0020815	The Dawn Patrol		Howard Hawks	Richard Barthelmess,Dou	War,Drama,Action	0	www.imdb.com/title/tt00			imdb
98319	tt0021140	Men Without Wo		John Ford	Kenneth MacKenna,Fran	Action,Drama	0	www.imdb.com/title/tt00			imdb
98320	tt0022111	The Maltese Falc		Roy Del Ruth	Bebe Daniels,Ricardo Co	Action,Drama,Rom	0	www.imdb.com/title/tt00			imdb

Define document

- data-config.xml
  - Import data from mysql

```
1  <dataConfig>
2      <dataSource type="JdbcDataSource"
3      name="ds1"
4      driver="com.mysql.jdbc.Driver"
5      url="jdbc:mysql://localhost:3306/video_info"
6      user="root" password="abcd"
7      batchSize="-1"/>
8      <document>
9          <entity name="movie" dataSource="ds1" query="select global_id,
   inner_id,title,alias,director,actors,genre,rating,play_url,website_url,
   image,source from total_movie">
10             <field column="global_id" name="id"/>
11         </entity>
12     </document>
13 </dataConfig>
14
```

- There is more: delta-import

  read: http://wiki.apache.org/solr/DataImportHandler#Using_delta-import_command

Define field types

- Configure **schema.xml** in **instanceDir/conf**

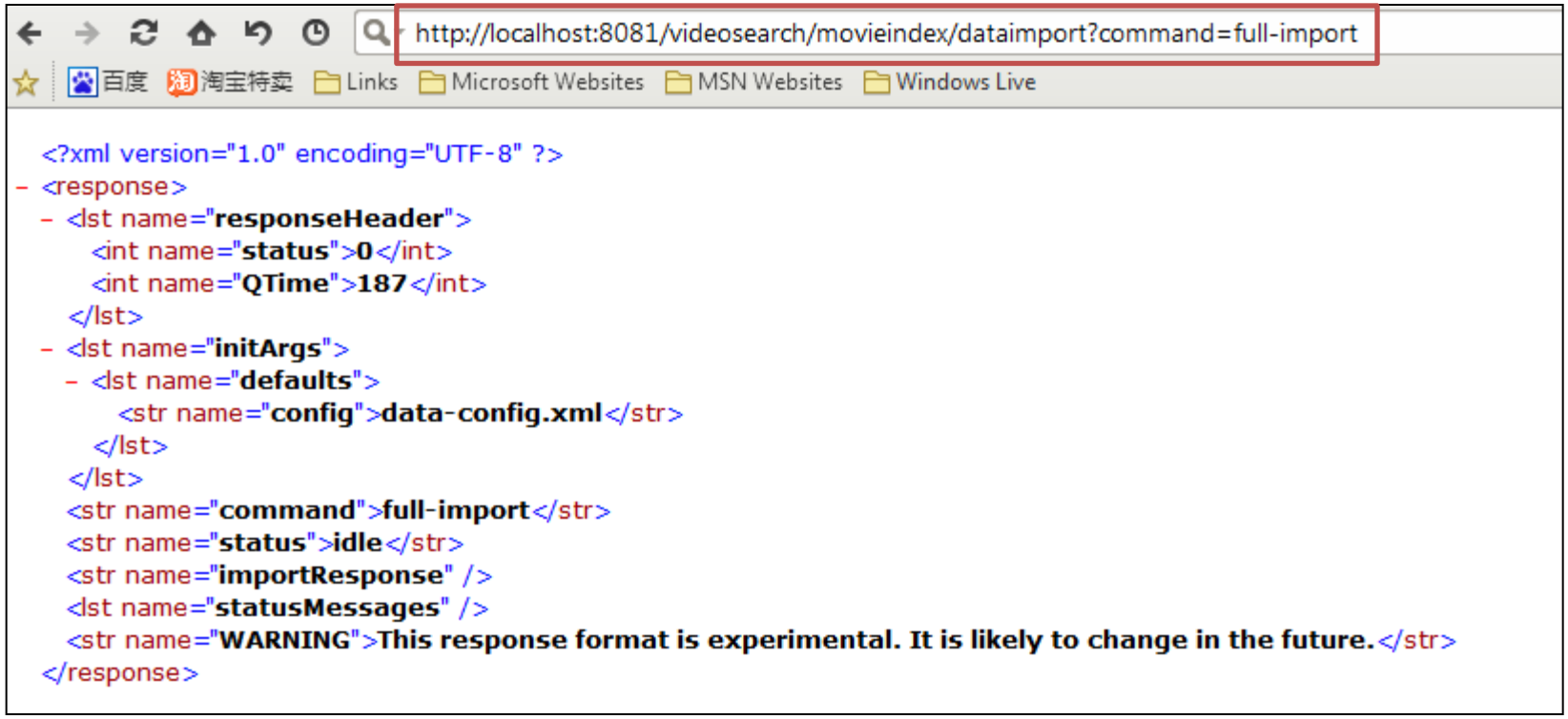
```
  <fieldType name="text_cn" class="solr.TextField" positionIncrementGap="100">
       <analyzer>
           <tokenizer class="com.chenlb.mmseg4j.solr.MMSegTokenizerFactory" mode="complex"
dicPath="C:\xampp\tomcat\webapps\videosearch\mmseg4j\data"/>
           <filter class="solr.LowerCaseFilterFactory"/>
       </analyzer>
   </fieldType>
 </types>


 <fields>
   <field name="title" type="text_cn" indexed="true" stored="true"/>
   <field name="alias" type="text_cn" indexed="true" stored="true"/>
   <field name="director" type="text_cn" indexed="true" stored="true"/>
   <field name="actors" type="text_cn" indexed="true" stored="true"/>
   <field name="genre" type="text_cn" indexed="true" stored="true"/>
   <field name="rating" type="float" indexed="false" stored="true"/>
   <field name="play_url" type="string" indexed="false" stored="true"/>
   <field name="website_url" type="string" indexed="false" stored="true"/>
   <field name="image" type="string" indexed="false" stored="true"/>
   <field name="source" type="string" indexed="false" stored="true"/>
  <field name="id" type="string" indexed="true" stored="true" required="true" />
  <field name="inner_id" type="string" indexed="true" stored="true"/>
  </fields>
```

Add custom fieldType

Declair fieldType of the document's fields, and whether a field should be indexed or stored

Trigger import process



Trigger import process

Processing…

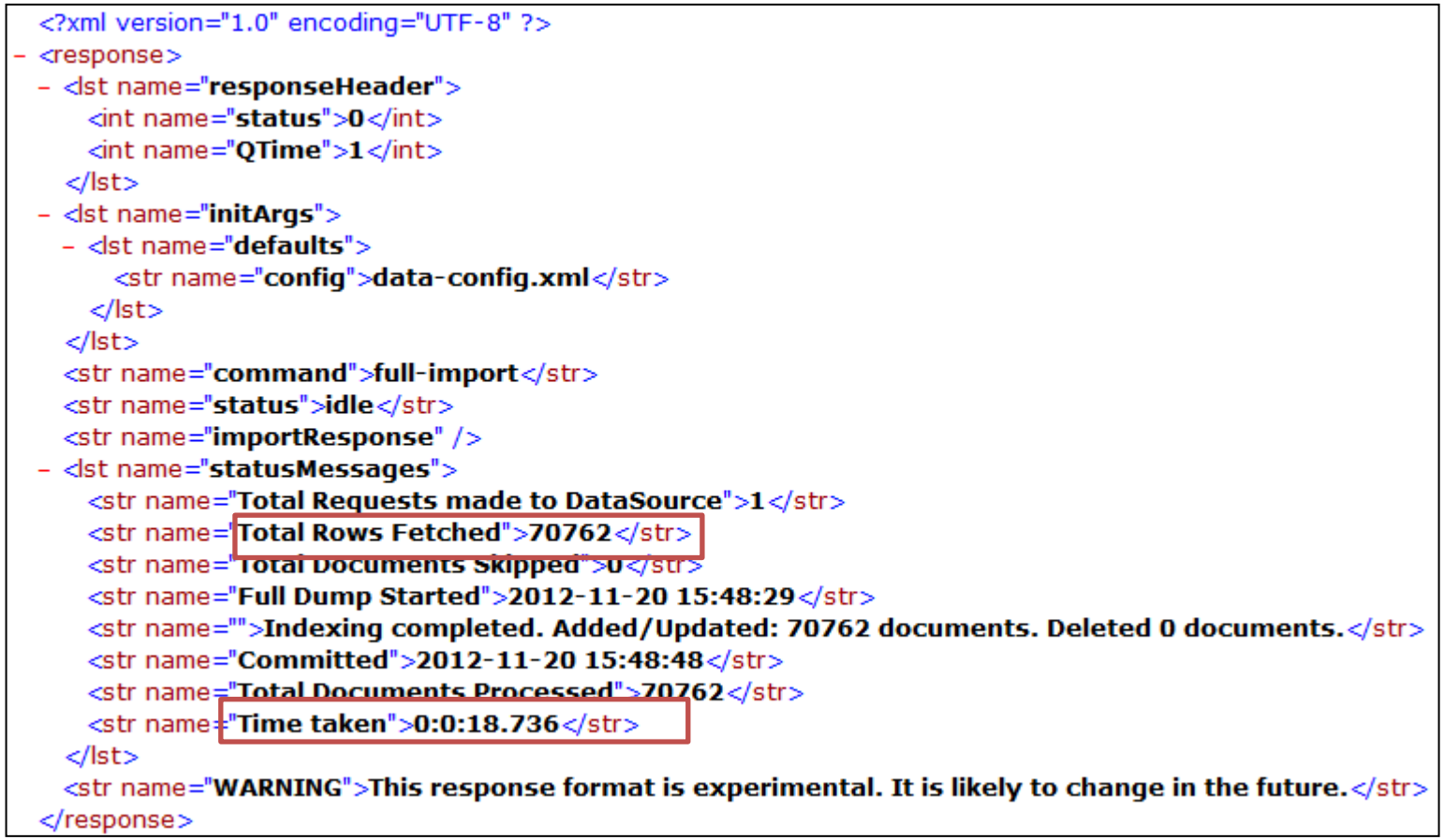
http://localhost:8081/videosearch/movieindex/dataimport

```
<?xml version="1.0" encoding="UTF-8" ?>
- <response>
- <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">1</int>
  </lst>
- <lst name="initArgs">
  - <lst name="defaults">
      <str name="config">data-config.xml</str>
    </lst>
  </lst>
  <str name="status">busy</str>
  <str name="importResponse">A command is still running...</str>
- <lst name="statusMessages">
    <str name="Time Elapsed">0:0:2.771</str>
    <str name="Total Requests made to DataSource">1</str>
    <str name="Total Rows Fetched">20334</str>
    <str name="Total Documents Processed">20333</str>
    <str name="Total Documents Skipped">0</str>
    <str name="Full Dump Started">2012-11-20 15:49:45</str>
  </lst>
  <str name="WARNING">This response format is experimental. It is likely to change in the future.</str>
</response>
```

Trigger import process

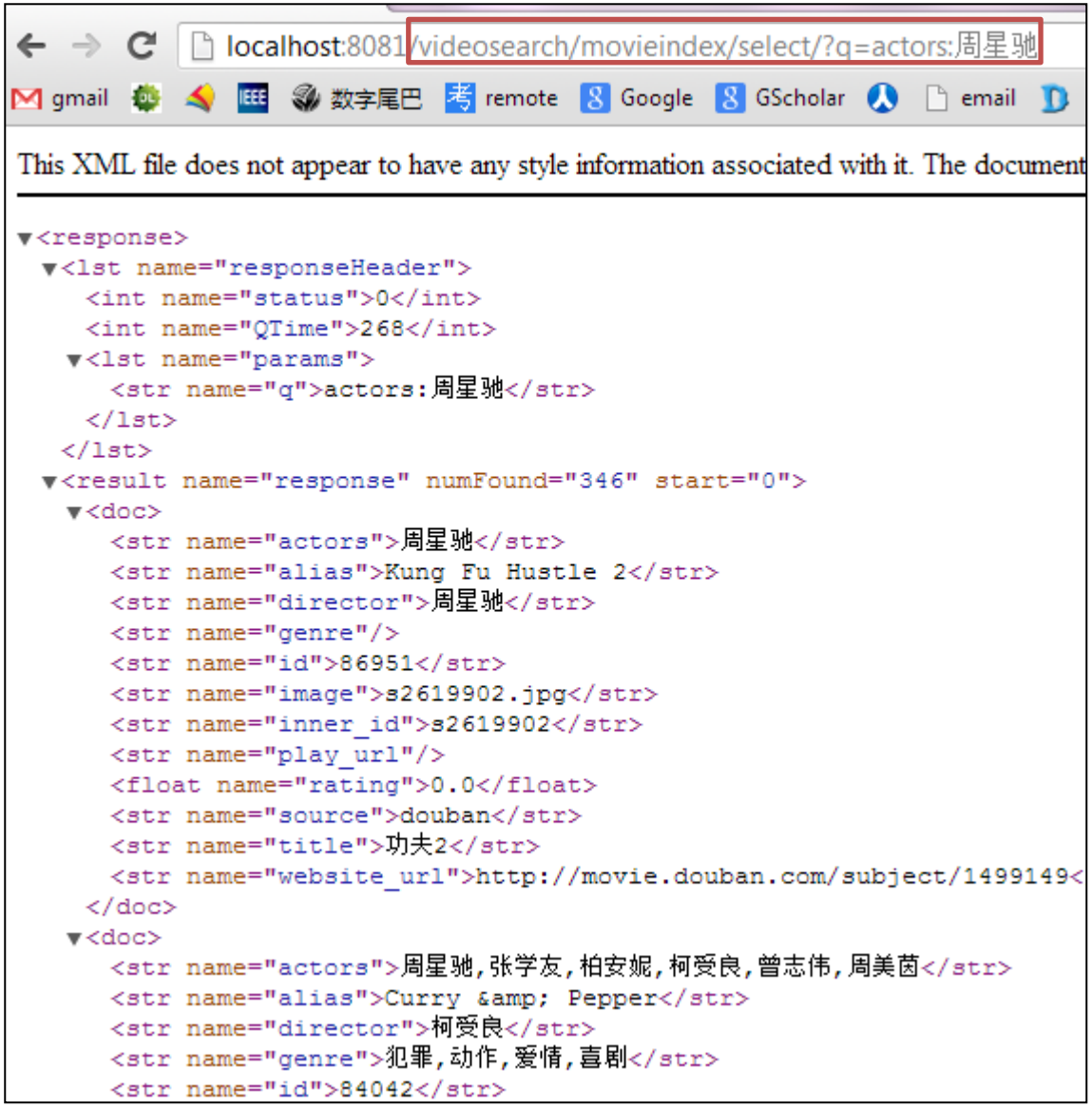
Done!!!

http://localhost:8081/videosearch/movieindex/dataimport

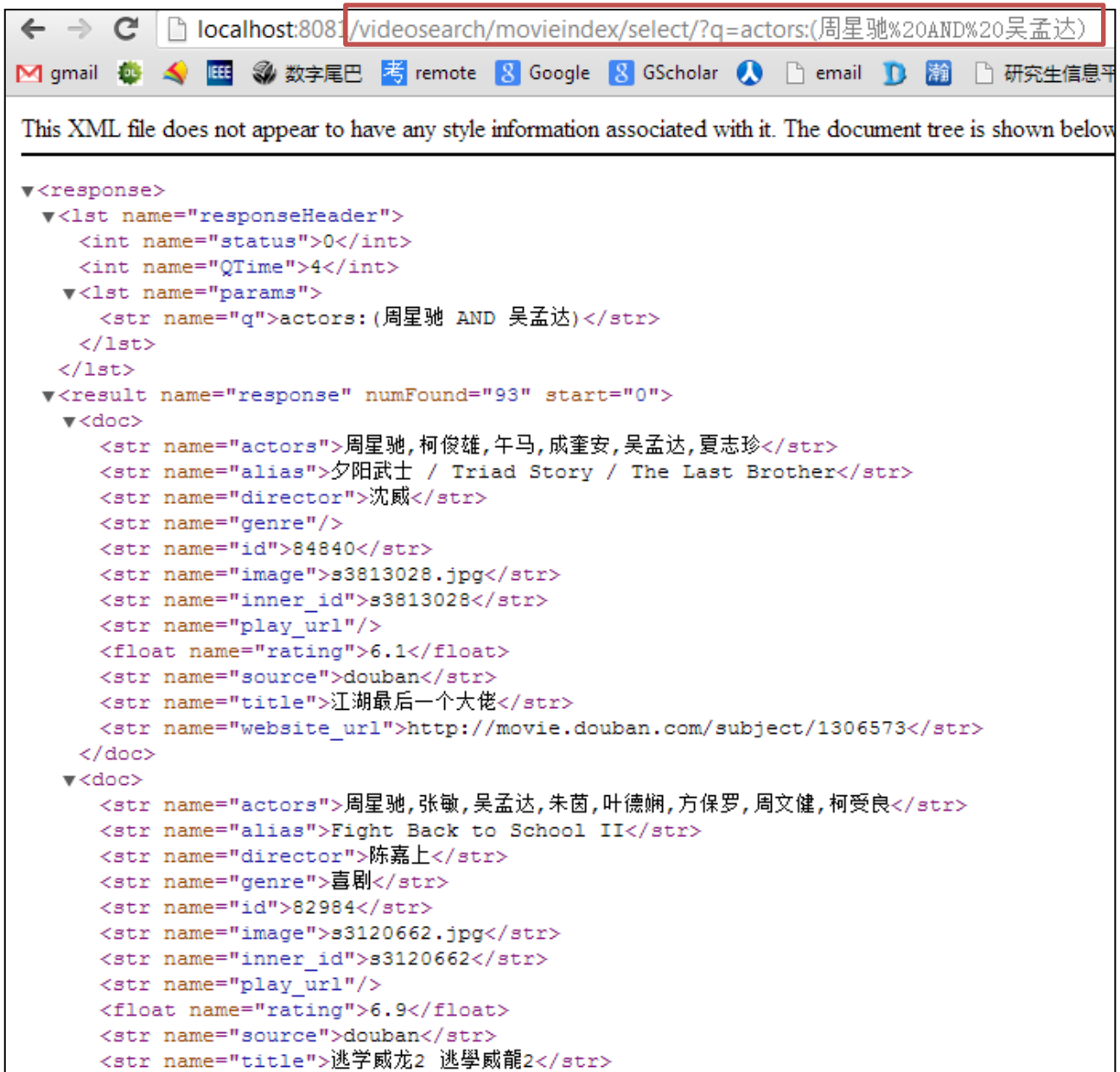


```
<?xml version="1.0" encoding="UTF-8" ?>
- <response>
  - <lst name="responseHeader">
      <int name="status">0</int>
      <int name="QTime">1</int>
    </lst>
  - <lst name="initArgs">
    - <lst name="defaults">
        <str name="config">data-config.xml</str>
      </lst>
    </lst>
    <str name="command">full-import</str>
    <str name="status">idle</str>
    <str name="importResponse" />
  - <lst name="statusMessages">
      <str name="Total Requests made to DataSource">1</str>
      <str name="Total Rows Fetched">70762</str>
      <str name="Total Documents Skipped">0</str>
      <str name="Full Dump Started">2012-11-20 15:48:29</str>
      <str name="">Indexing completed. Added/Updated: 70762 documents. Deleted 0 documents.</str>
      <str name="Committed">2012-11-20 15:48:48</str>
      <str name="Total Documents Processed">70762</str>
      <str name="Time taken">0:0:18.736</str>
    </lst>
    <str name="WARNING">This response format is experimental. It is likely to change in the future.</str>
  </response>
```

Search

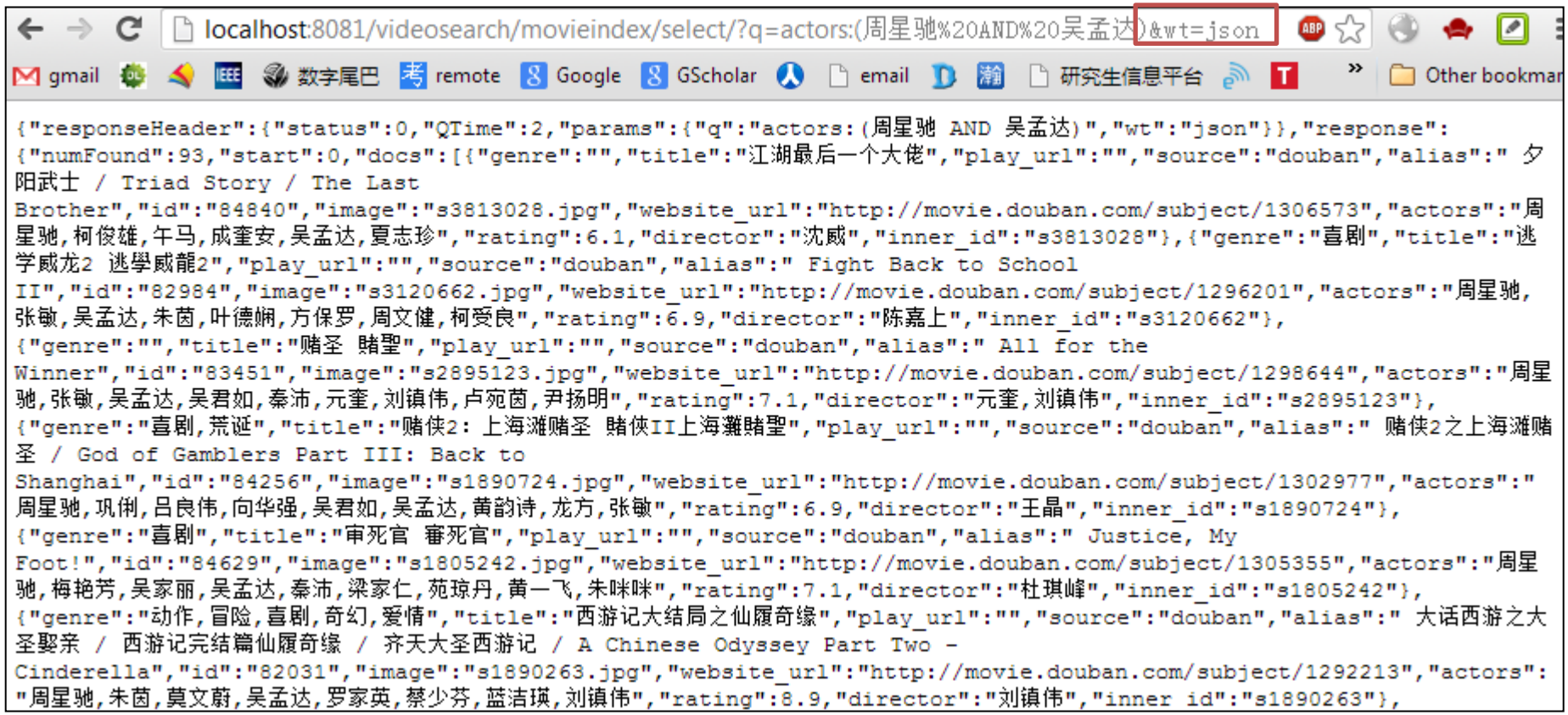

localhost:8081/videosearch/movieindex/select/?q=actors:周星驰
gmail
数字尾巴
remote
Google
GScholar
email
This XML file does not appear to have any style information associated with it. The document
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">268</int>
<lst name="params">
<str name="q">actors:周星驰</str>
</lst>
</lst>
<result name="response" numFound="346" start="0">
<doc>
<str name="actors">周星驰</str>
<str name="alias">Kung Fu Hustle 2</str>
<str name="director">周星驰</str>
<str name="genre"/>
<str name="id">86951</str>
<str name="image">s2619902.jpg</str>
<str name="inner_id">s2619902</str>
<str name="play_url"/>
<float name="rating">0.0</float>
<str name="source">douban</str>
<str name="title">功夫2</str>
<str name="website_url">http://movie.douban.com/subject/1499149<
</doc>
<doc>
<str name="actors">周星驰,张学友,柏安妮,柯受良,曾志伟,周美茵</str>
<str name="alias">Curry & Pepper</str>
<str name="director">柯受良</str>
<str name="genre">犯罪,动作,爱情,喜剧</str>
<str name="id">84042</str>

Search


localhost:8081/videosearch/movieindex/select/?q=actors:(周星驰%20AND%20吴孟达)
gmail
数字尾巴
remote
Google
GScholar
email
研究生信息平
This XML file does not appear to have any style information associated with it. The document tree is shown below
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">4</int>
<lst name="params">
<str name="q">actors:(周星驰 AND 吴孟达)</str>
</lst>
</lst>
<result name="response" numFound="93" start="0">
<doc>
<str name="actors">周星驰,柯俊雄,午马,成奎安,吴孟达,夏志珍</str>
<str name="alias">夕阳武士 / Triad Story / The Last Brother</str>
<str name="director">沈威</str>
<str name="genre"/>
<str name="id">84840</str>
<str name="image">s3813028.jpg</str>
<str name="inner_id">s3813028</str>
<str name="play_url"/>
<float name="rating">6.1</float>
<str name="source">douban</str>
<str name="title">江湖最后一个大佬</str>
<str name="website_url">http://movie.douban.com/subject/1306573</str>
</doc>
<doc>
<str name="actors">周星驰,张敏,吴孟达,朱茵,叶德娴,方保罗,周文健,柯受良</str>
<str name="alias">Fight Back to School II</str>
<str name="director">陈嘉上</str>
<str name="genre">喜剧</str>
<str name="id">82984</str>
<str name="image">s3120662.jpg</str>
<str name="inner_id">s3120662</str>
<str name="play_url"/>
<float name="rating">6.9</float>
<str name="source">douban</str>
<str name="title">逃学威龙2 逃學威龍2</str>

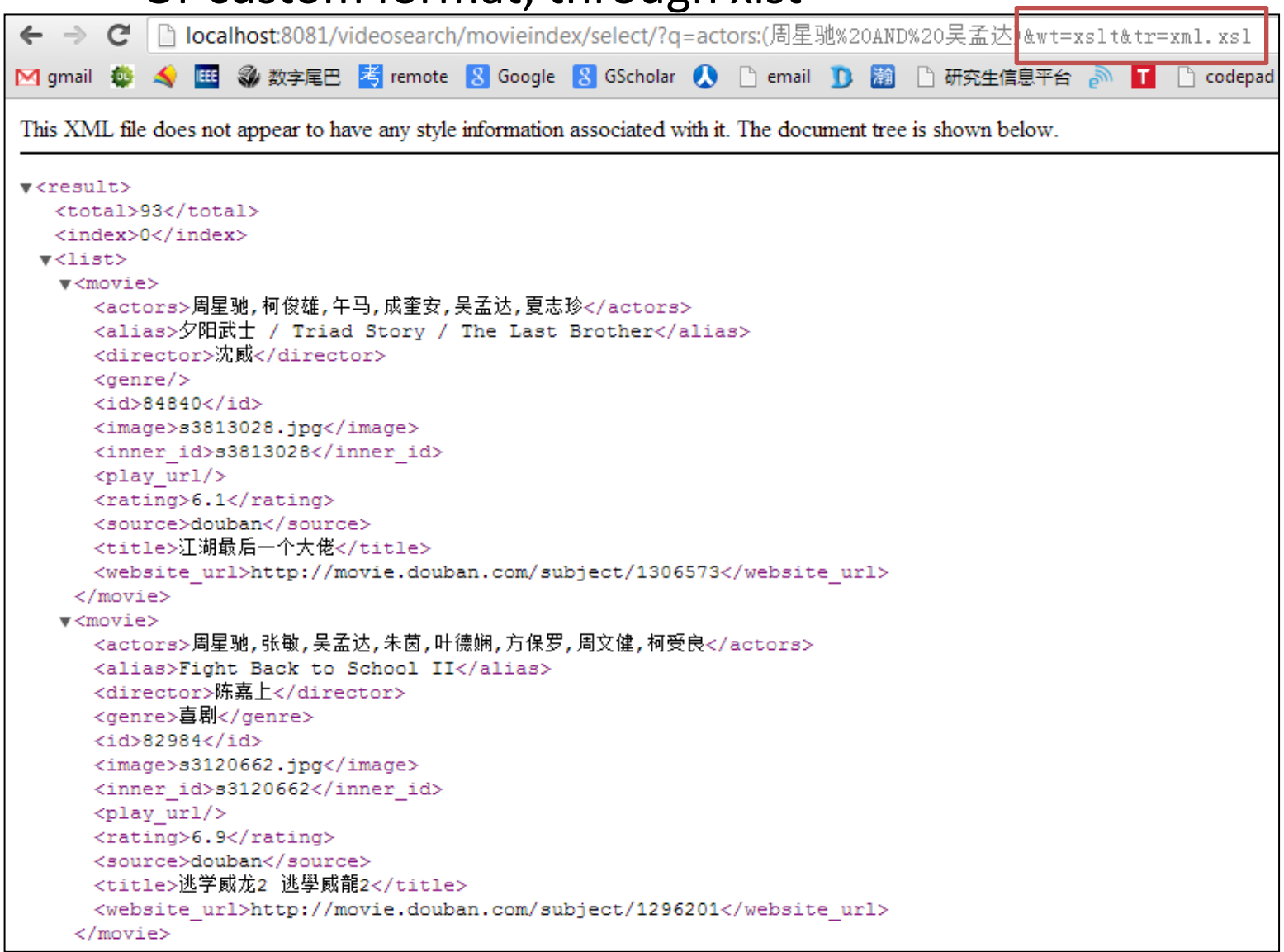
Control output format

- Xml is default
- json
  - Friendly to ajax app, and python, etc



Control output format

- Or custom format, through xlst



Control output format

- instanceDir/conf/xslt/xml.xsl

```
<?xml version='1.0' encoding='UTF-8'?>
<xsl:stylesheet version='1.0'  xmlns:xsl='http://www.w3.org/1999/XSL/Transform'>
<xsl:output method="xml" encoding="UTF-8"/>

    <xsl:template match="/">
        <result>
            <total>
                <xsl:value-of select="response/result/@numFound"/>
            </total>
            <index>
                <xsl:value-of select="response/result/@start"/>
            </index>
            <list>
                <xsl:apply-templates select="response/result"/>
            </list>
        </result>
    </xsl:template>

    <xsl:template match="result">
        <xsl:for-each select="doc">
            <movie>
                <xsl:for-each select="*">
                    <xsl:variable name="nodename" select="@name"/>
                    <xsl:element name="{$nodename}">
                        <xsl:value-of select="."/>
                    </xsl:element>
                </xsl:for-each>
            </movie>
        </xsl:for-each>
    </xsl:template>
</xsl:stylesheet>
```

Requires knowledge of xpath and xslt

p.s.

Encoding

- change your tomcat configuration
- tomcat\conf\server.xml

```
<Connector URIEncoding="UTF-8" port="8081" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

multiValued

book

bookid	intro
b001	xxx
b002	xx
b003	xxxxx

book_tag_relation

bookid	tagid
b001	t011
b001	t012
b001	t211

- nested entity in DIH's data-config

```
<dataConfig>
  <dataSource type="JdbcDataSource" … />
  <document>
    <entity name="book" dataSource="ds1" query="select bookid, intro from book">
            <entity name="book_tagids" dataSource="ds1" query="select tagid from book_tag_relation where
              bookid='${book.bookid}'">
            </entity>
    </entity>
  </document>
</dataConfig>
```

- schema.xml

```
<field name="tagid" type="string" indexed="true" stored="true" multiValued="true"/>
```

Configure buffered reading in DIH

- mysql
  - `<dataSource type="JdbcDataSource" name="ds1" driver="com.mysql.jdbc.Driver" url="jdbc:mysql://localhost:3306/video_info" user="root" password="abcd1234" batchSize="-1"/>`
- sqlserver
  - `<dataSource type="JdbcDataSource" name="ds1" driver="com.microsoft.sqlserver.jdbc.SQLServerDriver" url="jdbc:sqlserver://localhost:1433;databaseName=video_info;responseBuffering=adaptive;" user="sa" password="abcd1234"/>`

Search query syntax

- Query arguments for HTTP GET/POST to /select
  - “q” the query
  - “start” (0) offset
  - “rows” (10) number of docs
  - “fl” (*) fields to return
  - “qt” (standard) query type, maps to query handler
  - “df” (schema) default field to search
  - “wt” writer type (response format)
  - Include (+), exclude (-)
  - Field-specific searching: <fieldname>:<fieldvalue>
  - Wildcard searching: “*” or “?”
  - Range searching: Timestamp:[2006-01-01 TO *]
  - Proximity searching: “~”: “video ipod”~3 (up to 3 words apart)
  - Fuzzy searches: Belkin~0.8 (will find words close spellings)

DisMax

- DisMax is an abbreviation Disjunction Max. Disjunction: search query over multiple fields, Max: max score of fields is the final score, not sum.

- User enter a simple query foo, searching with query:”title:foo OR body:foo” is weak. DisMax is more flexible in ranking.

```
<requestHandler name="dismax"class="solr.SearchHandler" default="true">
  <lst name="defaults">
    <str name="defType">dismax</str>
    <str name="echoParams">explicit</str>
    <float name="tie">0.01</float>
    <str name="qf">text^0.5 category^1.5 title^2 body^1permalink^10.0
author^1.8 tag^1.3</str>
    <str name="pf">text^0.2 title^4 author^1.8 body^1</str>
    <str name="mm">3<60%</str>
  </lst>
</requestHandler>
```

tie: Tie breaker. 0: pure dismax, 1: dis sum

qf: query field.

pf:phrase field. boost docs when query is adjacent

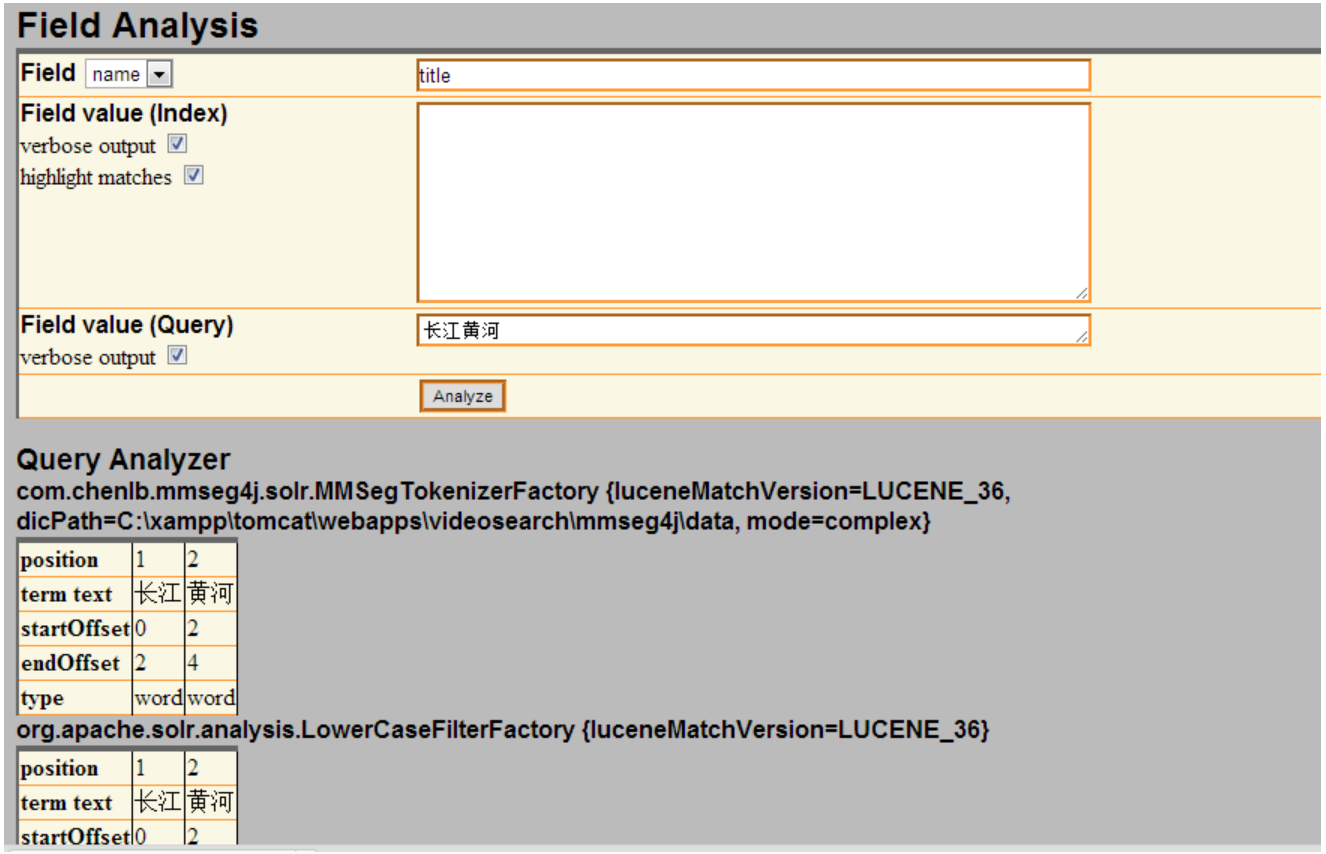
mm: minimum match

Check http://wiki.apache.org/solr/ExtendedDisMax

And http://searchhub.org/2010/05/23/whats-a-dismax/ for more explanation

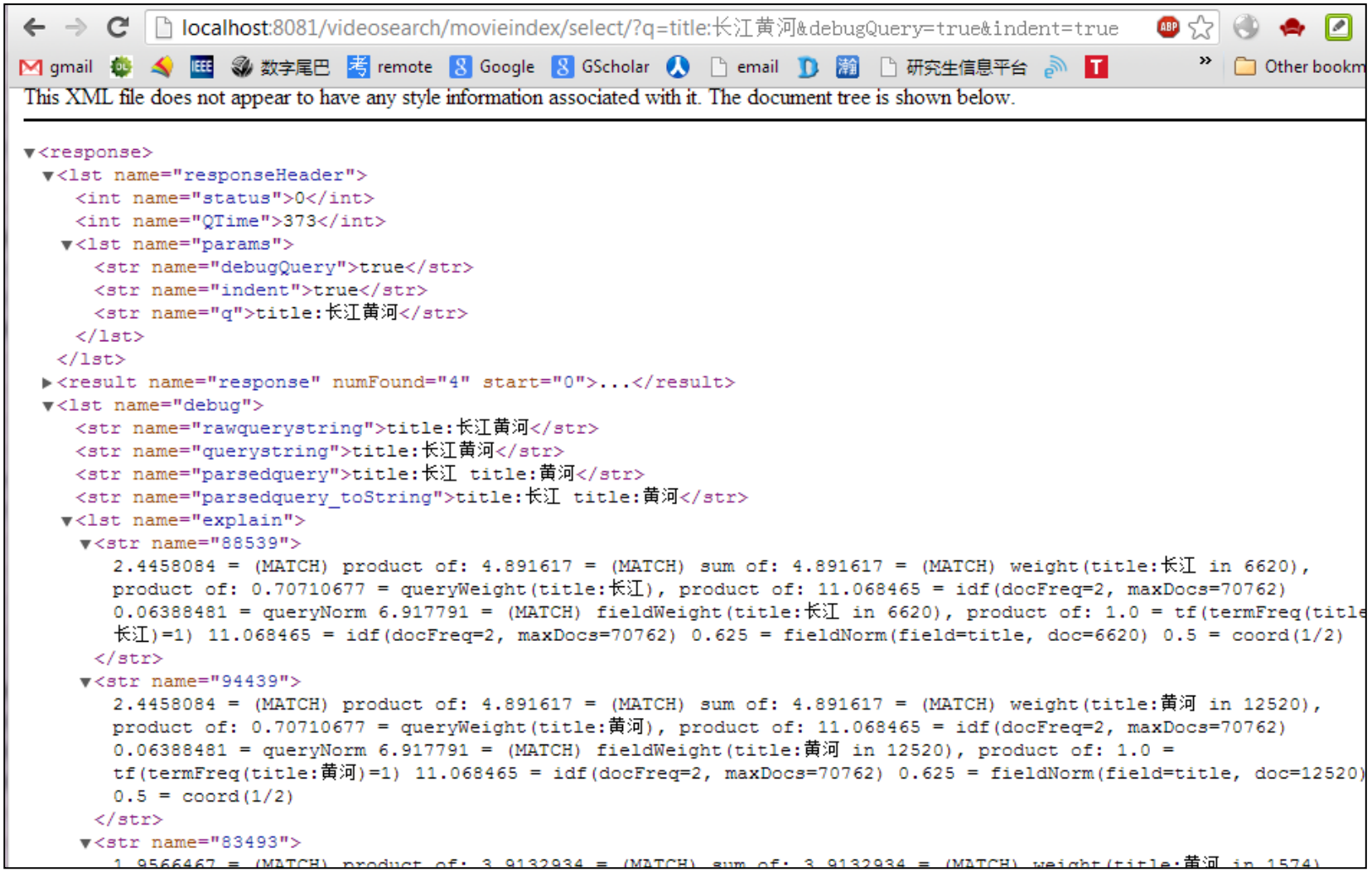
Debug Query

- When the search fails, you can debug you query to see how the query is processed behind the scene.
- http://url-to-your-solor-core/admin/analysis.jsp

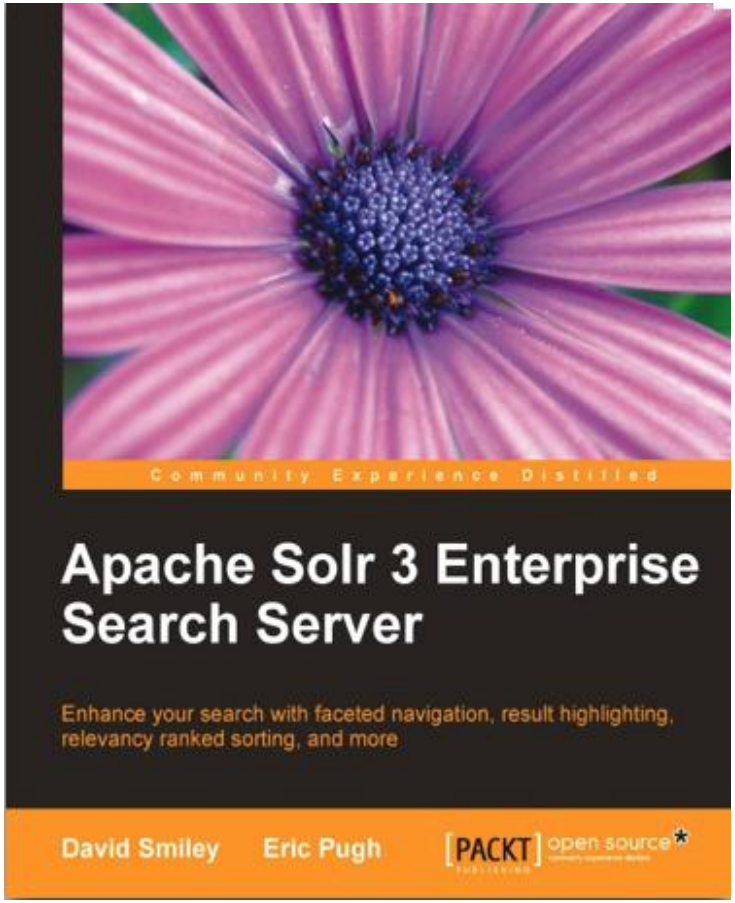


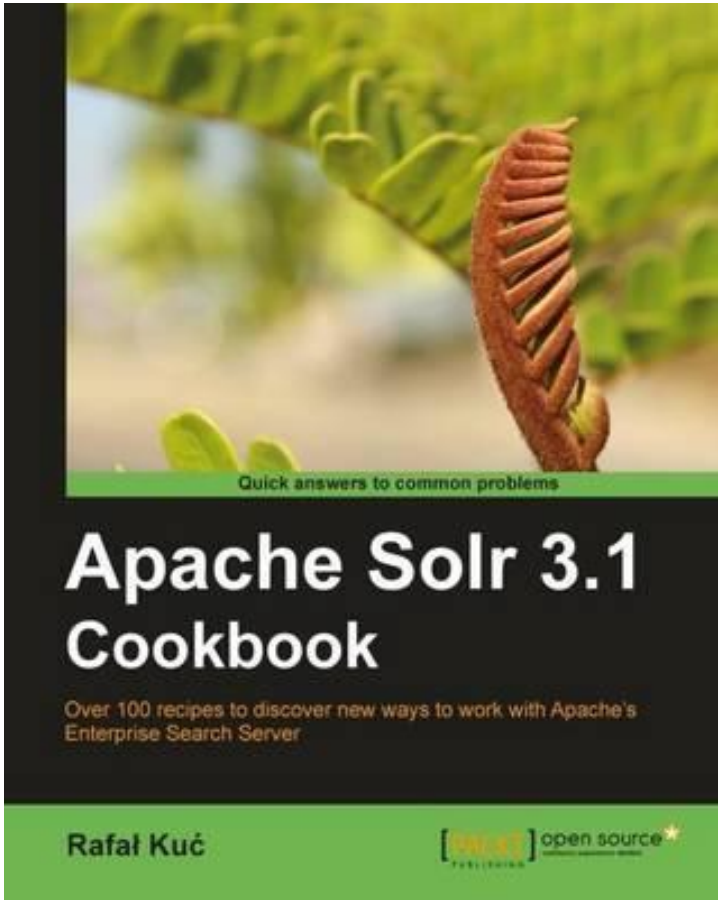
Debug Query

- or add "debugQuery=true" to your query



Books


Community Experience Distilled
Apache Solr 3 Enterprise Search Server
Enhance your search with faceted navigation, result highlighting, relevancy ranked sorting, and more
David Smiley
Eric Pugh
[PACKT] open source


Quick answers to common problems
Apache Solr 3.1 Cookbook
Over 100 recipes to discover new ways to work with Apache's Enterprise Search Server
Rafał Kuć
open source

Crawler

- Apache Nutch
  - Apache Nutch is an open source web-search software project. Nutch is a project of the Apache Software Foundation and is part of the larger Apache community of developers and users.

- Heritrix
  - Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project.

Document content extraction

- The Apache Tika™ toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries.

- Supported formats:
  - HyperText Markup Language
  - XML and derived formats
  - Microsoft Office document formats
  - OpenDocument Format
  - Portable Document Format
  - Electronic Publication Format
  - Rich Text Format
  - Compression and packaging formats
  - Text formats
  - Audio formats
  - Image formats
  - Video formats
  - Java class files and archives
  - The mbox format

End.