

A Brief Introduction to

Ultra-fast Lucene-based Search Server



Apache Solr offers Lucene's capabilities in an easy to use, fast search server with additional features like faceting, scalability and much more

@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

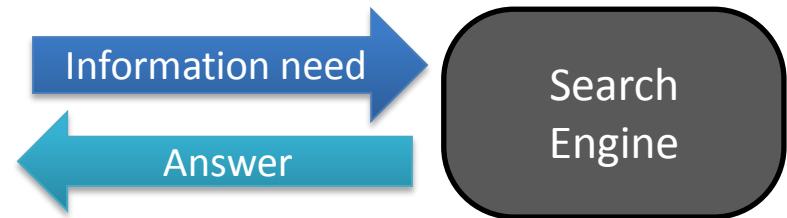
Backgrounds

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

The screenshot shows a search interface for querying freight car parameters. The user has entered '6' in the 'Self-weight (t)' field. A search error message is displayed below the form, stating: 'System operation error: org.springframework.dao.DataIntegrityViolationException: could not execute query; SQL [select * from TB_INFO_CLCS where flag = 'Y' and czdm = 'G' and ziz like '%6%' order by cxdm]; nested exception is org.hibernate.exception.DataException: could not execute query'. This error occurs because the database does not have an index on the 'ziz' column, making the search operation inefficient.

中国铁路客户服务中心
www.12306.cn是中国铁路客户服务中心唯一网站。截
2012年9月27日 星期四 首页 | 客运服务 | 货运服务 | 行包服务 | 车站引导 | 铁路常识 | 站车风采 | 客户信
| 首页 > 货运服务 > 罐车参数查询

搜索条件

车种车型 自重(t) 载重(t) 搜索

系统运行期错误:

查看详细

```
org.springframework.dao.DataIntegrityViolationException: could not execute query; SQL [select * from TB_INFO_CLCS where flag = 'Y' and czdm = 'G' and ziz like '%6%' order by cxdm]; nested exception is org.hibernate.exception.DataException: could not execute query
```

Resources Network Sources Timeline Profiles Audits Console

```
<pre>
    org.springframework.dao.DataIntegrityViolationException: could not execute query; SQL [select * from TB_INFO_CLCS where
flag = 'Y' and czdm = 'G' and ziz like '%6%' order by cxdm]; nested exception is org.hibernate.exception.DataException: could not execute query
    at org.springframework.orm.hibernate3.SessionFactoryUtils.convertHibernateAccessException(SessionFactoryUtils.java:642)
    at org.springframework.orm.hibernate3.HibernateAccessor.convertHibernateAccessException(HibernateAccessor.java:412)
    at org.springframework.orm.hibernate3.HibernateTemplate.doExecute(HibernateTemplate.java:411)

```

Options

- Google Search Appliance



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

- Sphinx



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

- Solr



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

- Elastic search



- 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?

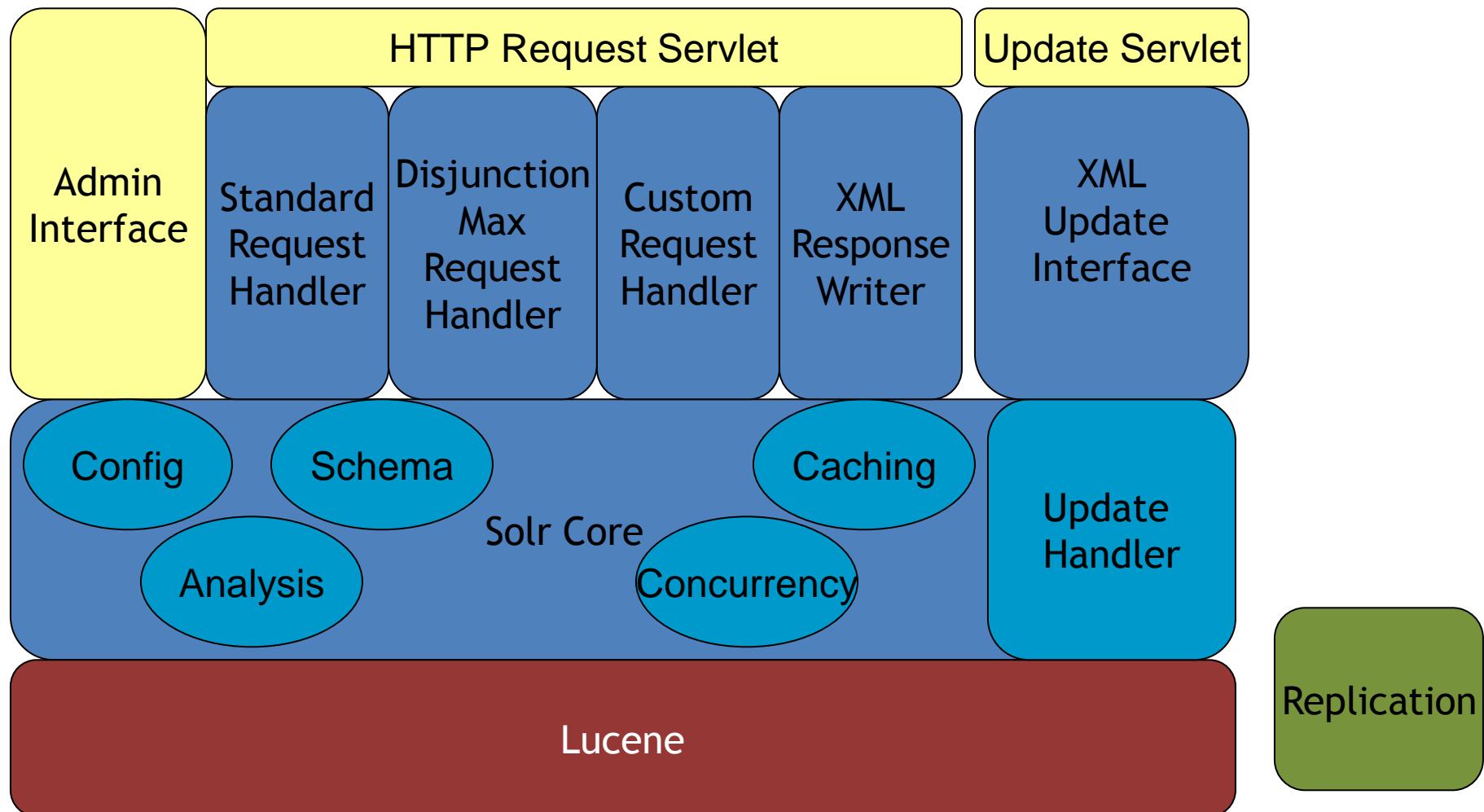
<http://wiki.apache.org/solr/PublicServers>

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



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.com/s6907514.jpg	
95277	13359	灵异女孩月子201		小川通仁	相武纱季,塙本高史,石垣	恐怖	6.2	http://movie.douban.com/s6917025.jpg	
95278	13360	Larva		Joo-Gong Meang		动画,短片	9	http://movie.douban.com/s6916993.jpg	

Document is consisted by fields.

```
<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>
```

Schema.xml

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 programmatically create documents to send to Solr

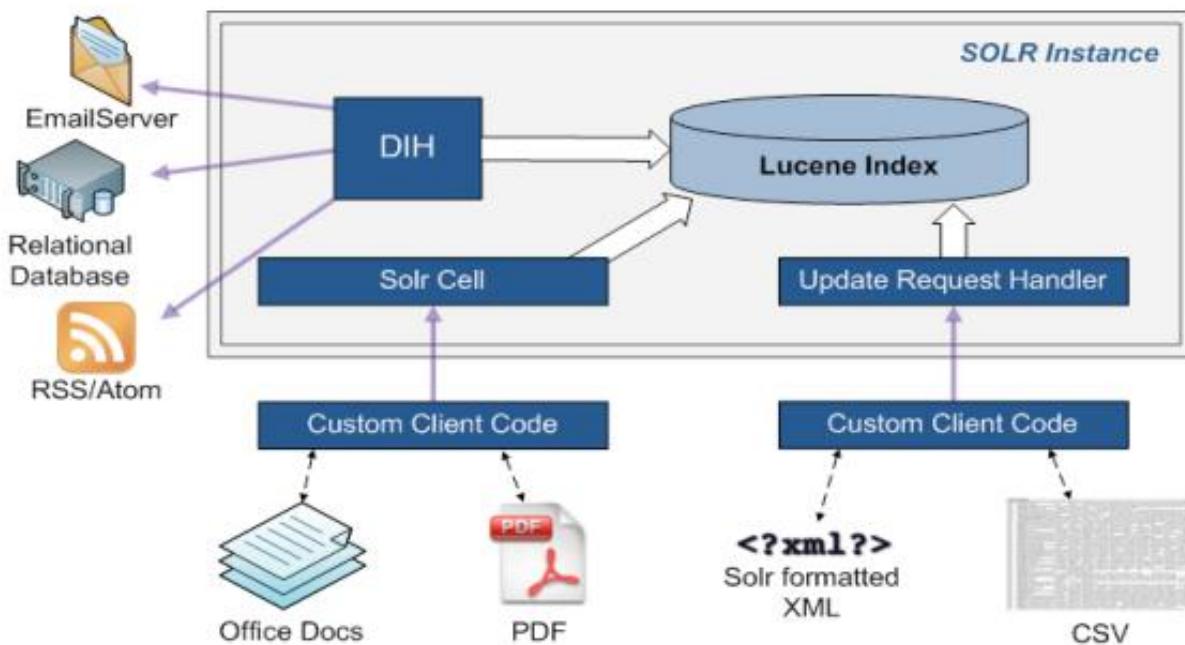


Illustration: SolrRecipes
by Lucid Imagination

Faceted Browsing

widely used in E-commerce

Find:

Boost by Price

Field Facets

cat	Electronics (3)
	Connector (2)
	Music (1)
manu_exact	Belkin (2)
	Apple Computer Inc. (1)

Query Facets

ipod (3)
GB (1)

Range Facets

Price (in \$)	0.0 (2)
---------------	-------------------------

3 results found in 20 ms Page 1 of 1

iPod & iPod Mini USB 2.0 Cable [More Like This](#)

Price: \$11.50
Features: car power adap
In Stock: false

天猫 TMALL.COM · 数码 [更多频道](#)

安卓 手机 智能

全部 > 手机 > 操作系统:Android/安卓 x > 安卓 手机 智能

共 3772 件相关商品

品牌	Huawei/华为	Lenovo/联想	MIUI/小米	HTC	Motorola/摩托罗拉	Samsung/三星	Sony/索尼
	Sony Ericsson/索尼爱立信	Coolpad/酷派	TOOKY/京崎	ZTE/中兴	LG	Changhong/长虹	Daxian/大

Belkin Mobile Power C

Price: \$19.95
Features: car power adap
In Stock: false

Apple 60 GB iPod with

Price: \$399.00
Features: iTunes, Podcasts, photos, or 150 hours of video playback. Up to 10 hours of battery life.

您是不是想找 手机智能触屏 | 双核智能手机安卓 | 三星手机智能 | htc手机智能 | 智能手机安卓系统 | 安卓2.3手机智能

所在地 默认排序 销量 价格 价格 包邮 折扣 旺旺在线 更多 店铺 大图



Highlighting

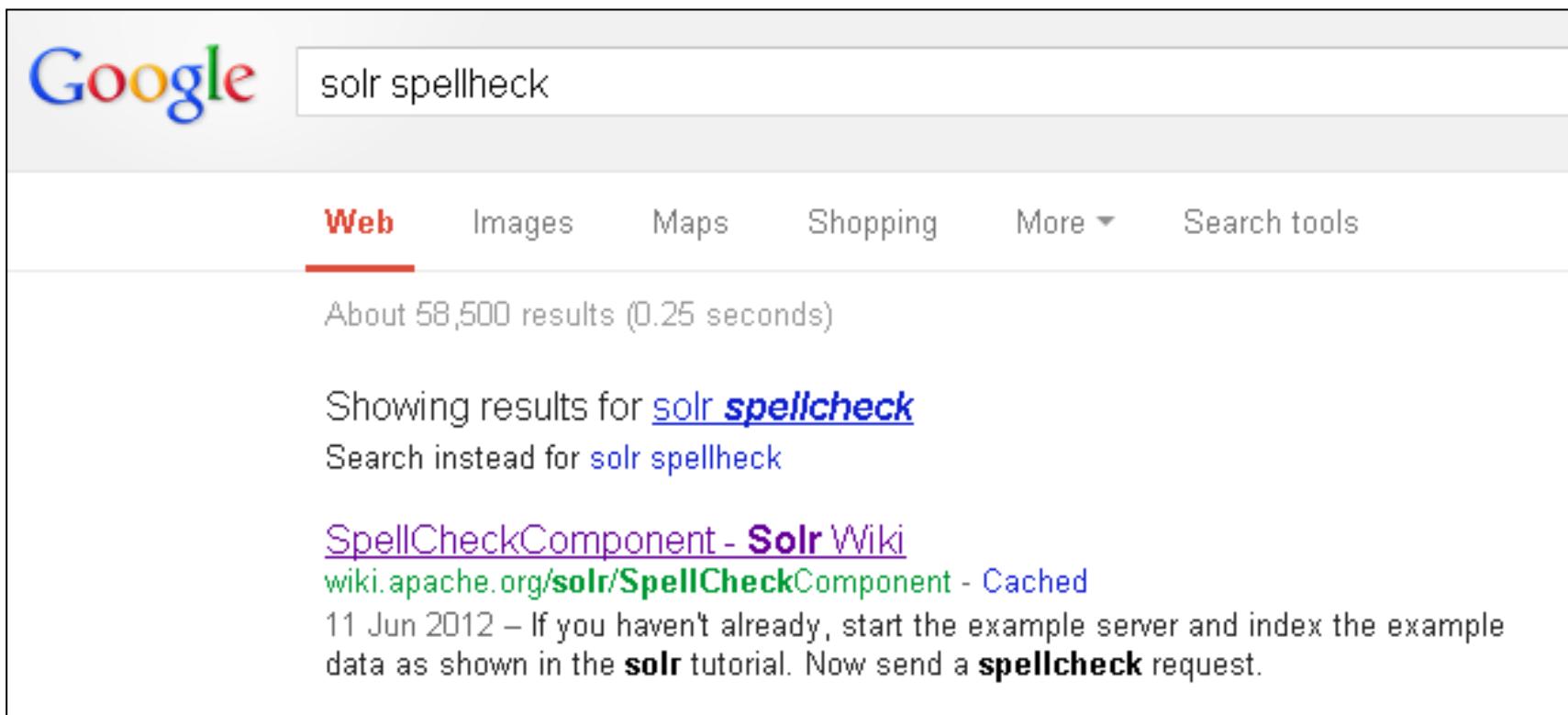
- Only text that has been both indexed and stored may be highlighted.
- Eg.
 - <http://localhost:8081/videosearch/movieindex/select/?q=title:爱情&hl=true&hl.fl=title>

```
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">1326</int>
  <lst name="params">
    <str name="q">title:爱情</str>
    <str name="hl.fl">title</str>
    <str name="hl">true</str>
  </lst>
</lst>
<result name="response" numFound="91" start="0">
  <doc>
    <str name="actors">林青霞</str>
    <str name="alias">Ai qing chang pao</str>
    <str name="director">陈耀圻 (Yao-chi Chen)</str>
    <str name="genre"/>
    <str name="id">84777</str>
    <str name="image">s2713078.jpg</str>
    <str name="inner_id">s2713078</str>
    <str name="play_url"/>
    <float name="rating">0.0</float>
    <str name="source">douban</str>
    <str name="title">爱情长跑</str>
    <str name="website_url">http://movie.douban.com/subject/1306315</str>
  </doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
  <doc>...</doc>
</result>
```

```
<lst name="highlighting">
  <lst name="84777">
    <arr name="title">
      <str><em>爱情</em>长跑</str>
    </arr>
  </lst>
  <lst name="90047">...</lst>
  <lst name="93477">...</lst>
  <lst name="88150">...</lst>
  <lst name="91940">...</lst>
  <lst name="82739">...</lst>
  <lst name="84522">...</lst>
  <lst name="84742">...</lst>
  <lst name="84753">...</lst>
  <lst name="85425">...</lst>
</response>
```

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

Spell check



A screenshot of a Google search results page. The search query "solr spellcheck" is entered in the search bar. The "Web" tab is selected, showing approximately 58,500 results found in 0.25 seconds. The top result is a link to the "SpellCheckComponent - Solr Wiki" page, which is a cached version from June 11, 2012. The page content describes how to start the example server, index data, and send a spellcheck request.

Google solr spellcheck

Web Images Maps Shopping More ▾ Search tools

About 58,500 results (0.25 seconds)

Showing results for [solr **spellcheck**](#)
Search instead for [solr spellcheck](#)

[SpellCheckComponent - Solr Wiki](#)
[wiki.apache.org/solr/SpellCheckComponent](#) - Cached
11 Jun 2012 – If you haven't already, start the example server and index the example data as shown in the **solr** tutorial. Now send a **spellcheck** request.

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

MoreLikeThis

中国经验

[高德思]

对跨国公司以及那些试图更多了解跨国公司如何在中国市场成功运作的人而言，尤其如此。以中国为主题的商业类图书越来越多，《中国 CEO—20 位商界领袖的经验之谈》（以下简称“中国 CEO”）是其中极具价值的一本。对跨国公司以及那些试图更多了解跨国公司如何在中国市场成功运作的人而言，尤其如此。……

2006年09月30日 ·  发现类似文章

苹果 vs. 谷歌：Android侵权案内幕

[Philip Elmer-DeWitt]

相关标签: 苹果 谷歌 Android

史蒂夫·乔布斯声称谷歌“剽窃”了苹果的创新。上周，ITC表示认可。当iPhone收到一条含有电话号码、电子邮件地址、网页链接或街道地址的消息时，这些相关信息会自动高亮、添加上下划线并能转换成可点击的链接。例如如果你点击电话号码，iPhone会询问你是否需要拨打；点

[主网页链接](#) iPhone

- 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

Solr Admin (example)

kyle-PC:8081

cwd=c:\xampp SolrHome=c:\xampp\tomcat\webapps\videosearch\movie\

HTTP caching is OFF



Solr

[SCHEMA] [CONFIG] [ANALYSIS] [SCHEMA BROWSER]
[STATISTICS] [INFO] [DISTRIBUTION] [PING] [LOGGING]

App server:

[JAVA PROPERTIES] [THREAD DUMP]

Make a Query

[FULL INTERFACE]

Query String:

* : *

Search

Assistance

[DOCUMENTATION] [ISSUE TRACKER] [SEND EMAIL]
[SOLR QUERY SYNTAX]

Current Time: Tue Nov 20 20:23:50 CST 2012

Server Start At: Sun Nov 18 10:23:04 CST 2012

Web Admin Interface

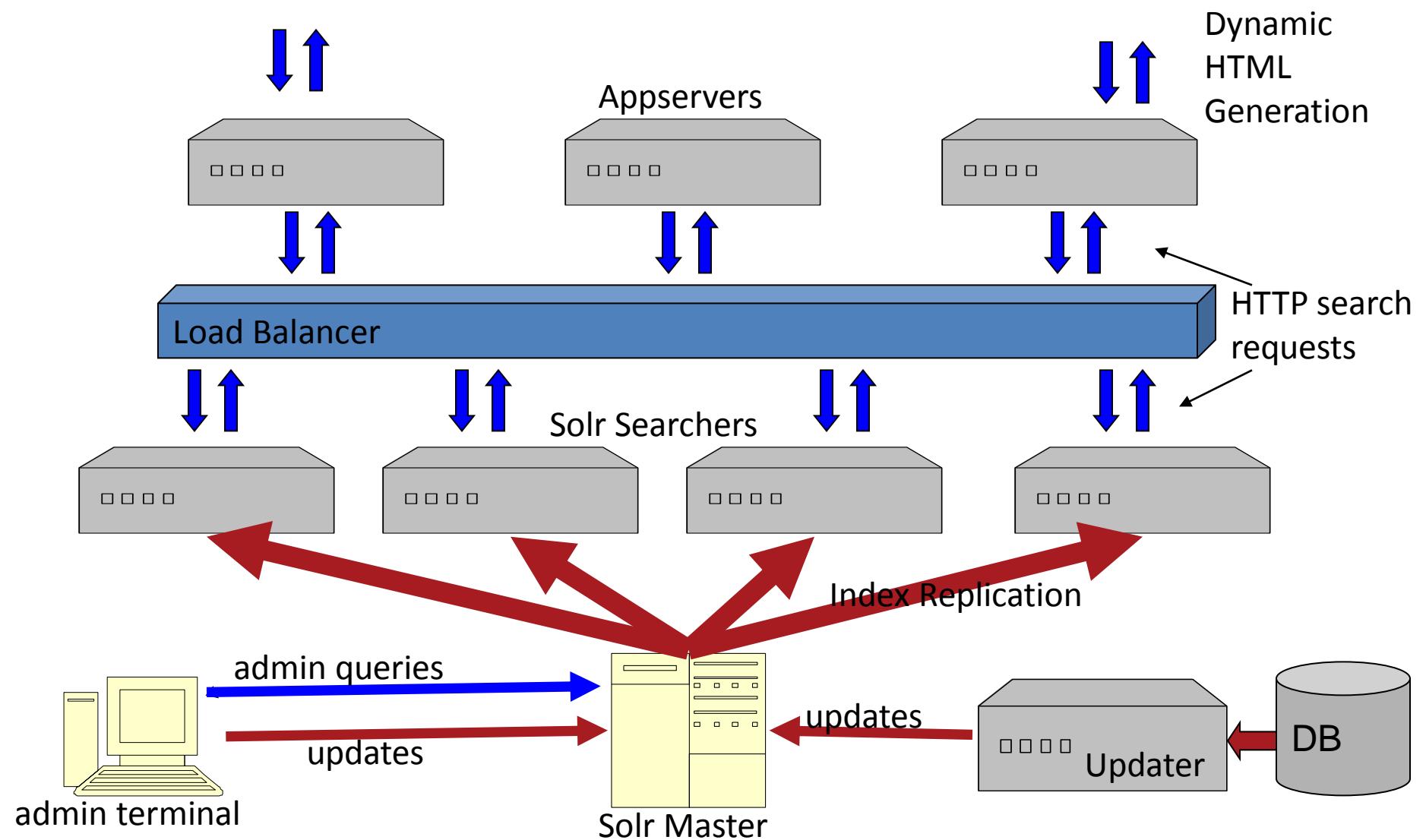
- Show Config, Schema, Distribution info
- Query Interface
- Statistics
 - Caches: lookups, hits, hitratio, inserts, evictions, size
 - RequestHandlers: requests, errors
 - UpdateHandler: adds, deletes, commits, optimizes
 - IndexReader, open-time, index-version, numDocs, maxDocs,
- Analysis Debugger
 - 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"

Name	Date modified	Type	Size
admin	2012/10/24 21:12	File folder	
bin	2012/10/24 21:12	File folder	
META-INF	2012/10/24 21:13	File folder	
mmseg4j	2012/10/24 21:13	File folder	
movie	2012/10/24 21:13	File folder	
WEB-INF	2012/11/20 15:09	File folder	
favicon.ico	2012/7/17 9:39	Icon	
index.jsp	2012/7/17 9:39	JSP File	
solr.xml	2012/10/24 22:19	XML Document	

Inside Solr Home

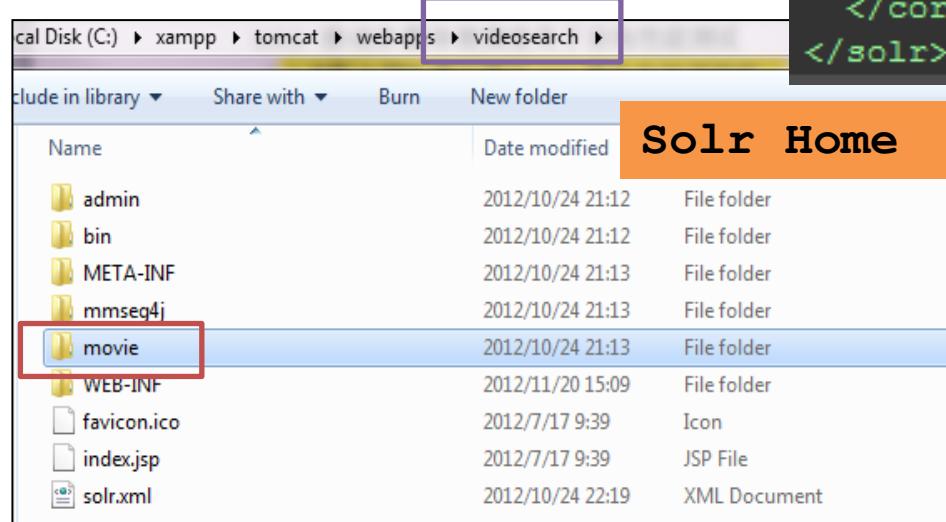
Name	Date modified	Type
conf	2012/10/31 15:08	File folder
data	2012/10/24 22:23	File folder

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

A screenshot of a web browser displaying the 'Solr Home' page. The URL bar shows 'http://localhost:8081/videosearch/'. The page content includes 'Welcome to Solr!' and a link 'Admin movieindex' (highlighted with a green box). In the bottom right corner, there is an Apache Solr logo.

Welcome to Solr!

Admin movieindex

Apache Solr

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.com/s/6907514.jpg			douban
95277	13359	灵异女孩月子 201		小川通仁	相武紗季,塙本高史,石垣リサ	恐怖	6.2	http://movie.douban.com/s/6917025.jpg			douban
95278	13360	Larva		Joo-Gong Meang		动画,短片	9	http://movie.douban.com/s/6916993.jpg			douban
95279	13361	女性瘾者 The Ny		拉斯·冯·提尔			0	http://movie.douban.com/s/movie-default-media			douban
95280	13362	女朋友 *男朋友		杨雅喆	桂纶镁,张孝全	爱情	0	http://movie.douban.com/s/movie-default-media			douban
95281	13363	UN-GO剧场版 UN-GO剧场	UN-GO剧场	水岛精二	胜地凉,丰崎爱生,山本希望	动画	0	http://movie.douban.com/s/6918552.jpg			douban
95282	13364	死党 ดีที่สุดเพื่อนกัน	Friends Never		马里奥·毛瑞尔,Natcha Jai		0	http://movie.douban.com/s/6920606.jpg			douban
95283	13365	超市试吃				喜剧,短片	4.4	http://movie.douban.com/s/6927343.jpg			douban
95284	13366	气球冷藏				喜剧,短片	6.2	http://movie.douban.com/s/6927345.jpg			douban
95285	13367	剪刀手				喜剧,短片	6.5	http://movie.douban.com/s/6927389.jpg			douban
95286	13368	民的1911		鲁艺		剧情,历史,动画	0	http://movie.douban.com/s/6927432.jpg			douban
98302	tt0004465	The Perils of Paul		Louis J. Gasnier,Dorothy Gish	Pearl White,Crane Wilbur	Action	0	www.imdb.com/title/tt0004465			imdb
98303	tt0006206	Les vampires		Louis Feuillade	Musidora,Édoua	Action,Adventure,Horror	0	www.imdb.com/title/tt0006206.action/tt0006206.jpg			imdb
98304	tt0006333	20,000 Leagues Under the Sea		Stuart Paton	Allen Holubar,Curtis Ben	Action,Adventure	0	www.imdb.com/title/tt0006333.action/tt0006333.jpg	http://player.youku.com/embed/	imdb	
98305	tt0009682	Tarzan of the Apes		Scott Sidney	Elmo Lincoln,Enid Marke	Action,Adventure	0	www.imdb.com/title/tt0009682.action/tt0009682.jpg			imdb
98306	tt0011036	Bullet Proof		Lynn Reynolds	Harry Carey,William Ryn	Action,Western	0	www.imdb.com/title/tt0011036			imdb
98307	tt0012752	The Three Musketeers		Fred Niblo	Douglas Fairbanks,Lewis Stone	Action,Adventure,Fantasy	0	www.imdb.com/title/tt0012752.action/tt0012752.jpg			imdb
98308	tt0014945	Girl Shy		Fred C. Newmeyer	Harold Lloyd,Jobyna Ralston	Action,Comedy,Romantic	0	www.imdb.com/title/tt0014945.action/tt0014945.jpg			imdb
98309	tt0016634	Beau Geste		Herbert Brenon	Ronald Colman,Neil Hamilton	Action,Adventure,I	0	www.imdb.com/title/tt0016634			imdb
98310	tt0016641	Ben-Hur: A Tale of the Christ		Fred Niblo,Charles Roscoe Gibly	Ramon Novarro,Francis X.	Action,Adventure,I	0	www.imdb.com/title/tt0016641.action/tt0016641.jpg			imdb
98311	tt0016654	The Black Pirate		Albert Parker	Douglas Fairbanks,Billie Dove	Adventure,Action	0	www.imdb.com/title/tt0016654.action/tt0016654.jpg			imdb
98312	tt0017925	The General		Clyde Bruckman,Buster Keaton	Buster Keaton,Marion Morehouse	Comedy,Romance	0	www.imdb.com/title/tt0017925.action/tt0017925.jpg	http://player.youku.com/embed/	imdb	
98313	tt0018578	Wings		William A. Wellman	Clara Bow,Richard Arlen	Drama,Romance,War	0	www.imdb.com/title/tt0018578.action/tt0018578.jpg			imdb
98314	tt0019412	Speedy		Ted Wilde	Harold Lloyd,Ann Christie	Action,Comedy,Fantasy	0	www.imdb.com/title/tt0019412.action/tt0019412.jpg			imdb
98315	tt0019421	Steamboat Bill, Jr.		Charles Reisner,John Wengraf	Buster Keaton,Tom McGowan	Action,Comedy,Dr	0	www.imdb.com/title/tt0019421.action/tt0019421.jpg	http://player.youku.com/embed/	imdb	
98316	tt0020570	Die weiße Faust		Arnold Fanck,Georg P. Schüller	Gustav Diessl,Leni Riefenstahl	Action,Adventure,I	0	www.imdb.com/title/tt0020570.action/tt0020570.jpg			imdb
98317	tt0020629	All Quiet on the Western Front		Lewis Milestone	Lew Ayres,Louis Wolheim	Action,Drama,Historic	0	www.imdb.com/title/tt0020629.action/tt0020629.jpg			imdb
98318	tt0020815	The Dawn Patrol		Howard Hawks	Richard Barthelmess,Douglas Fairbanks	War,Drama,Action	0	www.imdb.com/title/tt0020815.action/tt0020815.jpg			imdb
98319	tt0021140	Men Without Women		John Ford	Kenneth MacKenna,Frank Morgan	Action,Drama	0	www.imdb.com/title/tt0021140.action/tt0021140.jpg			imdb
98320	tt0022111	The Maltese Falcon		Roy Del Ruth	Bebe Daniels,Ricardo Cortez	Action,Drama,Romantic	0	www.imdb.com/title/tt0022111.action/tt0022111.jpg			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,
10             inner_id,title,alias,director,actors,genre,rating,play_url,website_url,
11             image,source from total_movie">
12             <field column="global_id" name="id"/>
13         </entity>
14     </document>
15 </dataConfig>
16
```

- 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**

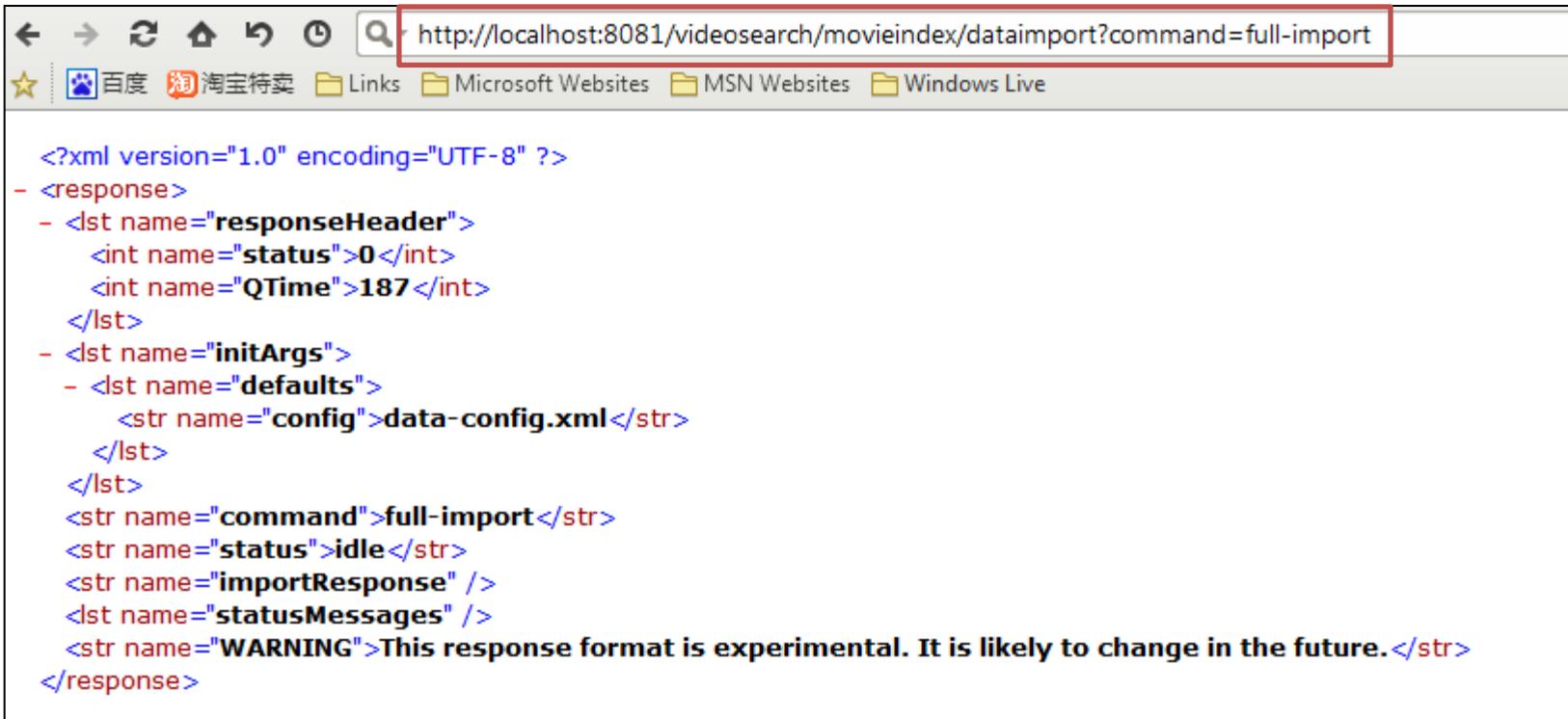
```
<fieldType name="text_cn" class="solr.TextField" positionIncrementGap="100">
    <analyzer>
        <tokenizer class="com.chenlb.mmseg4j.solr.MMSegTokenizerFactory" mode="complex"
dicPath="C:\xampp\tomcat\webapps\videosearch\mmseg4j\data"/>
        <filter class="solr.LowerCaseFilterFactory"/>
 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



The screenshot shows a web browser window with the URL `http://localhost:8081/videosearch/movieindex/dataimport?command=full-import` highlighted in a red box. The page content is an XML document with the following structure:

```
<?xml version="1.0" encoding="UTF-8" ?>
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">187</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="WARNING">This response format is experimental. It is likely to change in the future.</str>
</response>
```

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

This XML file does not appear to have any style information associated with it. The document structure is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<response>
    <lst name="responseHeader">
        <int name="status">0</int>
        <int name="QTime">268</int>
    </lst>
    <lst name="params">
        <str name="q">actors:周星驰</str>
    </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</str>
        </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">840422</str>
        </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">840422</str>
        </doc>
    </result>
</response>
```

Search

Control output format

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

localhost:8081/videosearch/movieindex/select/?q=actors:(周星驰%20AND%20吴孟达)&wt=json

```
{"responseHeader":{"status":0,"QTime":2,"params":{"q":"actors:(周星驰 AND 吴孟达)","wt":"json"}}, "response": {"numFound":93,"start":0,"docs": [{"genre":"","title":"江湖最后一个大佬","play_url":"","source":"douban","alias":" 夕阳武士 / Triad Story / The Last Brother","id":"84840","image":"s3813028.jpg","website_url":"http://movie.douban.com/subject/1306573","actors":"周星驰,柯俊雄,午马,成奎安,吴孟达,夏志珍","rating":6.1,"director":"沈威","inner_id":"s3813028"}, {"genre":"喜剧","title":"逃学威龙2 逃學威龍2","play_url":"","source":"douban","alias":" Fight Back to School II","id":"82984","image":"s3120662.jpg","website_url":"http://movie.douban.com/subject/1296201","actors":"周星驰,张敏,吴孟达,朱茵,叶德娴,方保罗,周文健,柯受良","rating":6.9,"director":"陈嘉上","inner_id":"s3120662"}, {"genre":"","title":"赌圣 賭聖","play_url":"","source":"douban","alias":" All for the Winner","id":"83451","image":"s2895123.jpg","website_url":"http://movie.douban.com/subject/1298644","actors":"周星驰,张敏,吴孟达,吴君如,秦沛,元奎,刘镇伟,卢宛茵,尹扬明","rating":7.1,"director":"元奎,刘镇伟","inner_id":"s2895123"}, {"genre":"喜剧,荒诞","title":"赌侠2：上海滩赌圣 賭俠II上海灘賭聖","play_url":"","source":"douban","alias":" 赌俠2之上海滩赌圣 / God of Gamblers Part III: Back to Shanghai","id":"84256","image":"s1890724.jpg","website_url":"http://movie.douban.com/subject/1302977","actors":"周星驰,巩俐,吕良伟,向华强,吴君如,吴孟达,黄韵诗,龙方,张敏","rating":6.9,"director":"王晶","inner_id":"s1890724"}, {"genre":"喜剧","title":"审死官 審死官","play_url":"","source":"douban","alias":" Justice, My Foot!","id":"84629","image":"s1805242.jpg","website_url":"http://movie.douban.com/subject/1305355","actors":"周星驰,梅艳芳,吴家丽,吴孟达,秦沛,梁家仁,苑琼丹,黄一飞,朱咪咪","rating":7.1,"director":"杜琪峰","inner_id":"s1805242"}, {"genre":"动作,冒险,喜剧,奇幻,爱情","title":"西游记大结局之仙履奇缘","play_url":"","source":"douban","alias":" 大话西游之大圣娶亲 / 西游记完結篇仙履奇緣 / 齐天大圣西游记 / A Chinese Odyssey Part Two - Cinderella","id":"82031","image":"s1890263.jpg","website_url":"http://movie.douban.com/subject/1292213","actors":"周星驰,朱茵,莫文蔚,吴孟达,罗家英,蔡少芬,蓝洁瑛,刘镇伟","rating":8.9,"director":"刘镇伟","inner_id":"s1890263"}]}
```

Control output format

- Or custom format, through xslt

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<result>
  <total>93</total>
  <index>0</index>
  <list>
    <movie>
      <actors>周星驰, 柯俊雄, 午马, 成奎安, 吴孟达, 夏志珍</actors>
      <alias>夕阳武士 / Triad Story / The Last Brother</alias>
      <director>沈威</director>
      <genre/>
      <id>84840</id>
      <image>s3813028.jpg</image>
      <inner_id>s3813028</inner_id>
      <play_url/>
      <rating>6.1</rating>
      <source>douban</source>
      <title>江湖最后一个大佬</title>
      <website_url>http://movie.douban.com/subject/1306573</website_url>
    </movie>
    <movie>
      <actors>周星驰, 张敏, 吴孟达, 朱茵, 叶德娴, 方保罗, 周文健, 柯受良</actors>
      <alias>Fight Back to School II</alias>
      <director>陈嘉上</director>
      <genre>喜剧</genre>
      <id>82984</id>
      <image>s3120662.jpg</image>
      <inner_id>s3120662</inner_id>
      <play_url/>
      <rating>6.9</rating>
      <source>douban</source>
      <title>逃学威龙2 逃學威龍2</title>
      <website_url>http://movie.douban.com/subject/1296201</website_url>
    </movie>
  </list>
</result>
```

Control output format

- instanceDir/conf/xslt/xml.xsl

```
1 <?xml version='1.0' encoding='UTF-8'?>
2 <xsl:stylesheet version='1.0' xmlns:xsl='http://www.w3.org/1999/XSL/Transform'>
3 <xsl:output method="xml" encoding="UTF-8"/>
4
5     <xsl:template match="/">
6         <result>
7             <total>
8                 <xsl:value-of select="response/result/@numFound"/>
9             </total>
10            <index>
11                <xsl:value-of select="response/result/@start"/>
12            </index>
13            <list>
14                <xsl:apply-templates select="response/result"/>
15            </list>
16        </result>
17    </xsl:template>
18
19    <xsl:template match="result">
20        <xsl:for-each select="doc">
21            <movie>
22                <xsl:for-each select="*">
23                    <xsl:variable name="nodename" select="@name"/>
24                    <xsl:element name="${nodename}">
25                        <xsl:value-of select=".//${nodename}"/>
26                    </xsl:element>
27                </xsl:for-each>
28            </movie>
29        </xsl:for-each>
30    </xsl:template>
31 </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> tie: Tie breaker. 0: pure dismax, 1: dis sum
    <str name="qf">text^0.5 category^1.5 title^2 body^1 permalink^10.0
      author^1.8 tag^1.3</str> qf: query field.
    <str name="pf">text^0.2 title^4 author^1.8 body^1</str> pf:phrase field. boost docs
      when query is adjacent
    <str name="mm">3<60%</str> mm: minimum match
  </lst>
</requestHandler>
```

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 your query to see how the query is processed behind the scene.
- <http://url-to-your-solr-core/admin/analysis.jsp>

Field Analysis

Field	<input type="button" value="name"/>	title
Field value (Index)		
verbose output	<input checked="" type="checkbox"/>	
highlight matches	<input checked="" type="checkbox"/>	
Field value (Query)		
verbose output	<input checked="" type="checkbox"/>	
长江黄河		
<input type="button" value="Analyze"/>		

Query Analyzer

```
com.chenlb.mmseg4j.solr.MMSegTokenizerFactory {luceneMatchVersion=LUCENE_36,
dicPath=C:\xampp\tomcat\webapps\videosearch\mmseg4j\data, mode=complex}
```

position	1	2
term text	长江	黄河
startOffset	0	2
endOffset	2	4
type	word	word

```
org.apache.solr.analysis.LowerCaseFilterFactory {luceneMatchVersion=LUCENE_36}
```

position	1	2
term text	长江	黄河
startOffset	0	2

Debug Query

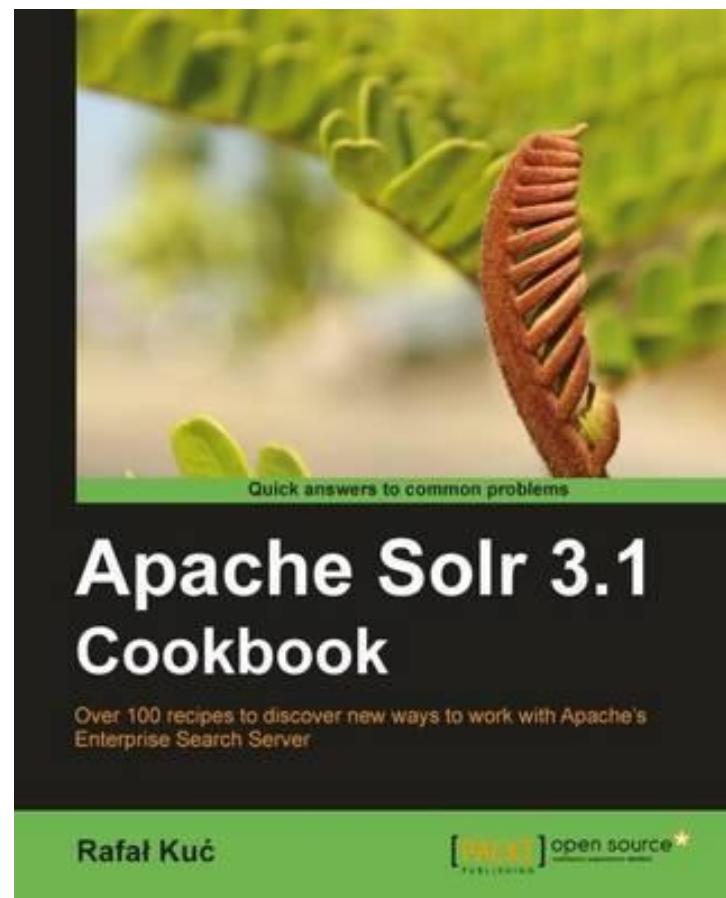
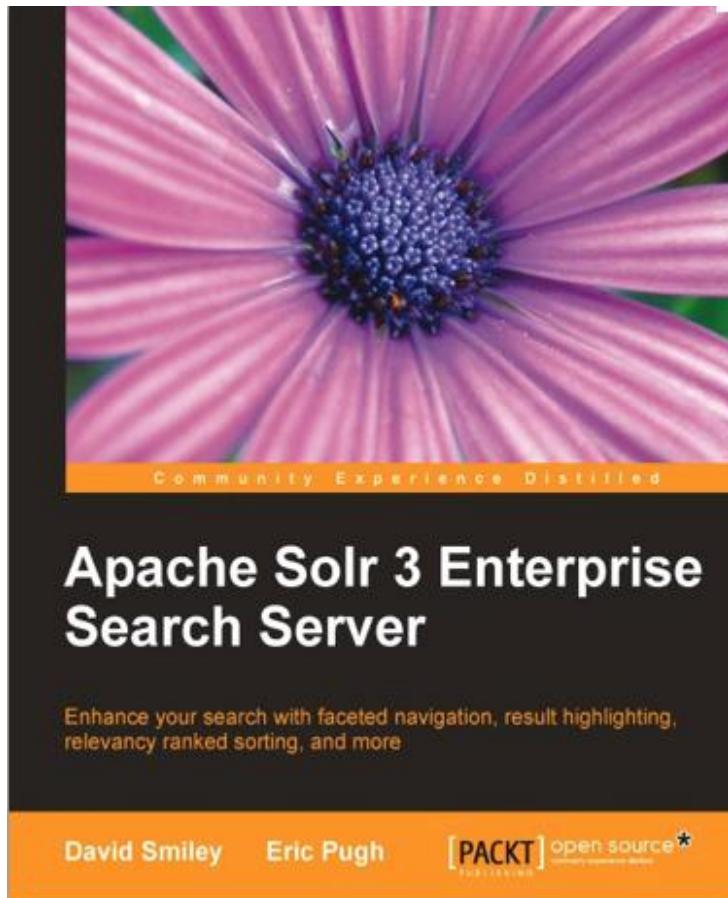
- or add “debugQuery=true” to your query



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">373</int>
  <lst name="params">
    <str name="debugQuery">true</str>
    <str name="indent">true</str>
    <str name="q">title:长江黄河</str>
  </lst>
</lst>
<result name="response" numFound="4" start="0">...</result>
<lst name="debug">
  <str name="rawquerystring">title:长江黄河</str>
  <str name="querystring">title:长江黄河</str>
  <str name="parsedquery">title:长江 title:黄河</str>
  <str name="parsedquery_toString">title:长江 title:黄河</str>
  <lst name="explain">
    <str name="88539">
      2.4458084 = (MATCH) product of: 4.891617 = (MATCH) sum of: 4.891617 = (MATCH) weight(title:长江 in 6620),
      product of: 0.70710677 = queryWeight(title:长江), product of: 11.068465 = idf(docFreq=2, maxDocs=70762)
      0.06388481 = queryNorm 6.917791 = (MATCH) fieldWeight(title:长江 in 6620), product of: 1.0 =
      tf(termFreq(title:长江)=1) 11.068465 = idf(docFreq=2, maxDocs=70762) 0.625 = fieldNorm(field=title, doc=6620) 0.5 = coord(1/2)
    </str>
    <str name="94439">
      2.4458084 = (MATCH) product of: 4.891617 = (MATCH) sum of: 4.891617 = (MATCH) weight(title:黄河 in 12520),
      product of: 0.70710677 = queryWeight(title:黄河), product of: 11.068465 = idf(docFreq=2, maxDocs=70762)
      0.06388481 = queryNorm 6.917791 = (MATCH) fieldWeight(title:黄河 in 12520), product of: 1.0 =
      tf(termFreq(title:黄河)=1) 11.068465 = idf(docFreq=2, maxDocs=70762) 0.625 = fieldNorm(field=title, doc=12520) 0.5 = coord(1/2)
    </str>
    <str name="83493">
      1.0566467 = (MATCH) product of: 3.0132934 = (MATCH) sum of: 3.0132934 = (MATCH) weight(title:黄河 in 1574)
    </str>
  </lst>
</lst>
```

Books



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.