Search and browse super powers
an introduction to Solr

Ben Gaucherin - 11/05/2011
A quick tour of the VM

- The VirtualBox VM is a TurnkeyLinux Core (Ubuntu) setup with both a LAMP stack and a Solr instance running on Tomcat.
- The user root’s password is turnkey123.
- Tomcat is installed in /usr/share/tomcat6/.
- The Tomcat admin user is admin and the password is turnkey123.
- The Solr home directory which includes both a conf and data directory is /etc/solr/.
- The php files are in /var/www/solr_tut.
Part I – Key concepts for search
The basic model

1. Feed “stuff” into the engine
2. Search for “stuff”
Inverted index

• A “forward index” establishes a relationship between a document and the terms within it
  
  Doc 1 → {“The” “little” “brown” “fox”}
  Doc 2 → {“Bounce” “little” “brown” “ball”}

• An “inverted index” establishes a relationship between terms and the documents where they can be found

  “The” → Doc 1
  “Bounce” → Doc 2
  “little” → Doc 1, Doc 2
  “brown” → Doc 1, Doc 2
  “ball” → Doc 2
  “fox” → Doc 1
Obama names Elena Kagan solicitor general | Harvard...
President-elect Barack Obama has nominated Harvard Law School (HLS) Dean Elena Kagan as solicitor general. If confirmed...
news.harvard.edu/gazette/story/2009/01/obama-names-elena-kagan-solicitor-general/

Howell Jackson named as prospective acting dean of Harvard...
... Harvard Law School (HLS), subject to the US Senate’s confirmation of Dean Elena Kagan’s nomination to serve as US Solicitor General, President...

Harvard Gazette: Archibald Cox dies at 92
ProfessorEmeritus Archibald Cox, former Watergate special prosecutor and solicitor general, died Saturday (May 29) in his home in...
news.harvard.edu/gazette/2004/06.03/15-cox.html

Guantanamo Bay: Memo to W. Finally we agree on something...
... On your behalf, the Solicitor General of the United States argued that all of Guantánamo, including every square inch on which your base sits, is...
www.people.fas.harvard.edu/~jdoming/images/id_guantanamo.pdf

HPAC PR | Harvard Gazette
... Obama names Elena Kagan solicitor general. January 5, 2009. President-elect Barack Obama has nominated Harvard Law...
news.harvard.edu/gazette/section/hpac-pr/page/9/?gz_page_type=archive

Colleagues recall Kagan's years at Harvard | Harvard Gazette
... 2009. In 2009, she was confirmed as US solicitor general. The ... Obama names Elena Kagan solicitor general. President-elect ...
news.harvard.edu/gazette/story/2010/08/senate-approves-kagan-for-high-court/
Faceted search

• Items in the index have facets
  – For example working with books we could have the following facets: author(s), isbn, subject(s), publication date, publisher, etc.

• When you perform a search, facets provide additional information about the results

• Just about anything can be a facet

• With faceted search: no single taxonomy, no assumption of global relevance, etc.
<table>
<thead>
<tr>
<th>Course Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstraction and Design in Computation (CSCI E-250 (23371))</strong></td>
</tr>
<tr>
<td>Harvard Extension School</td>
</tr>
<tr>
<td>Spring 2012</td>
</tr>
<tr>
<td>John Gregory Morrisett Ph.D., Allen B. Cutting Professor of Computer Science, Harvard University</td>
</tr>
<tr>
<td>Online only</td>
</tr>
<tr>
<td><strong>Accumulating Evidence: How to Conduct a Quantitative Research Synthesis That Informs Educational Policy and Practice (Part I) (S032A1)</strong></td>
</tr>
<tr>
<td>Harvard Graduate School of Education</td>
</tr>
<tr>
<td>Spring 2012</td>
</tr>
<tr>
<td>James S. Kim</td>
</tr>
<tr>
<td>1:00 p.m. - 4:00 p.m.</td>
</tr>
<tr>
<td><strong>Adv Comp Bio &amp; Bioinformatics (BIOS13-01)</strong></td>
</tr>
<tr>
<td>Harvard School of Public Health</td>
</tr>
<tr>
<td>Fall 2011</td>
</tr>
<tr>
<td>Guocheng Yuan and Winston Hide</td>
</tr>
<tr>
<td>3:30 p.m. - 5:20 p.m.</td>
</tr>
<tr>
<td><strong>Advanced Epidemiologic Methods (EPI207-01)</strong></td>
</tr>
<tr>
<td>Harvard School of Public Health</td>
</tr>
<tr>
<td>Fall 2011</td>
</tr>
<tr>
<td>James M. Robins and Miguel Hernan</td>
</tr>
<tr>
<td>Monday 3:30 p.m. - 5:20 p.m.; Wednesday 3:30 p.m. - 5:20 p.m.; Thursday 1:30 p.m. - 3:20 p.m.</td>
</tr>
</tbody>
</table>
Part II – The Solr application
What is Solr?

- An Apache Software Foundation project
- Extends the capabilities of the Apache Lucene search engine
- Solr is a Java web application distributed as a WAR file – it runs in a servlet container (e.g. Tomcat)
HTTP Server
Servlet Container – e.g. Tomcat
HTTP Request/Response
Solr
Solr Index
Solr Engine
Solr application tour
Key applications elements

• Using `tomcat_dir/conf/Catalina/localhost/solr.xml` we tell Tomcat where `solr/home` is

• `solr/home` contains a `conf` directory that contains a number of configuration files:
  - `solrconfig.xml`
  - `schema.xml`
  - `spellings.txt`, `stopwords.txt`, `synonyms.txt`, etc.

• And the index will be located here:
  ```
  <dataDir>
    ${solr.data.dir:/Applications/tomcat/conf/solr_conf/data}
  </dataDir>
  ```
Part III – Indexing our data
The schema - types

```xml
<?xml version="1.0" encoding="utf-8" ?>
<schema name="ShelfLife Demo" version="1.0">
  <types>
    <fieldType name="text" class="solr.TextField" positionIncrementGap="100">
      <analyzer type="index">
        ...
      </analyzer>
      <analyzer type="query">
        ...
      </analyzer>
    </fieldType>
  </types>
</schema>
```
The schema - fields

<fields>
  <field name="isbn" type="text" indexed="true" stored="true" required="true" />

  <field name="title" type="string" indexed="true" stored="true" required="true" />
  <field name="title_det" type="text" index="true" stored="true" required="true" />
  <copyField source="title" dest="title_det" />

  ...

</fields>

<uniqueKey>isbn</uniqueKey>

<defaultSearchField>title</defaultSearchField>

<solrQueryParser defaultOperator="OR" />
</schema>
The data

• Posted to /update

```xml
<?xml version="1.0" encoding="utf-8" ?>
<add>
  <doc>
    <field name='title'>The little prince</field>
    <field name='subject'>space</field>
    <field name='subject'>story</field>
    ...
  </doc>
  <doc>
    <field name='subject'>space</field>
    <field name='subject'>story</field>
    ...
  </doc>
  <commit/>
  <optimize/>
</add>
```
solr_data_post.php
Part IV – Solr queries 101
You can talk to Solr directly

- The */select* request handler
- \texttt{q}: the actual query
- \texttt{wt}: to specify the format for the results
- \texttt{start} and \texttt{rows}: to help with pagination through large numbers of results

http://192.168.56.101:8080/solr/

\texttt{select}?\texttt{q}=\ast:\ast\texttt{&wt}=\texttt{xml}\texttt{&start}=5\texttt{&rows}=10
Adding facets

- We just turn facets on and specify the facet fields and additional configuration as needed

http://192.168.56.101:8080/solr/select?q=*:*&wt=xml&facet=true&facet.field=author
And many more

- Other query parameters (e.g. `hl` for highlighting)

- Other request handlers (see `solrconfig.xml` for a full list)
  - `/admin`
  - `/admin/ping`
  - `/debug/dump`
  - `...`
solr_search.php
Part V – Faceted search user interface
Some examples…

Subjects:
Some examples…

Cross Registration

- Any Cross-Reg Status
  - Eligible for cross-registration (319)
  - Not eligible for cross-registration (122)

Calendar Period

- Any Calendar Period
  - Jun to Aug 2011 (Summer Term) (21)
  - Sep to Dec 2011 (Fall Term) (206)
  - Jan 2012 (Winter Session) (8)
  - Jan to May 2012 (Spring Term) (197)

Level

- Any Credit Level
  - Graduate (354)
  - Undergraduate (118)

School

- All Schools
  - Faculty of Arts and Sciences (252)
  - Harvard Business School - MBA Program (1)
  - Harvard Extension School (77)
  - Harvard Graduate School of Design (15)
  - Harvard Graduate School of Education (12)
  - Harvard Kennedy School (7)
  - Harvard Law School (13)
  - Harvard Medical School (11)
  - Harvard School of Public Health (33)
  - Harvard Summer School (20)

Days

- Any Day
  - Monday (85)
  - Tuesday (114)
  - Wednesday (96)
  - Thursday (105)
  - Friday (48)
  - Saturday (2)

Start Time

- Any Time
  - Morning (8am-noon) (87)
  - Afternoon (Noon-5pm) (105)
  - Evening (after 5pm) (64)

Offered

- Any Offered Status
  - Not offered (1)
  - Offered (440)
Some examples...

Harvard University 2011-2012 Course Catalog

Cross Registration
- Any Cross-Reg Status
  - Eligible for cross-registration (78)
  - Not eligible for cross-registration (4)
- Calendar Period
- Any Calendar Period
  - Sep to Dec 2011 (Fall Term)
- Level
- Any Credit Level
  - Graduate (74)
  - Undergraduate (12)
- School

NARROWED BY
- Sep to Dec 2011 (Fall Term)
- Engineering and Applied Sciences
- Faculty of Arts and Sciences
- Fall 2011-2012

82 Courses Found

Course Information
- Applied Mathematics 205 - Advanced Scientific Computing; Numerical Methods
  - (1370)
  - Faculty of Arts and Sciences
  - Fall 2011-2012
  - David J. Knezevic

Sort by: Course Name | School | Department | Day | Time | Relevance
Schedule | Relevance

SMTWTFS
10:00 a.m. - 11:00 a.m.
Some examples...
Part VI – Beyond the basics
Beyond the basics

- DisMax – meant to simplify complex queries
- Multi-core – multiple schemas in one instance
- Shards – Distributed search, BIG data split into partitions
- Solr a new kind of database?
A quick recap

- Install Solr
  - Install a servlet container (e.g. Tomcat)
  - Install the Solr WAR file
  - Set solr/home and solr.data.dir
- Define your schema and create/find your data
- Load data into Solr using /update
- Create a user interface that generates Solr queries

- The art of implementing Solr
  - Get it to work with simple data and settings first
  - Then tweak, adjust, refine, iterate
The End

bengaucherin.wordpress.com