Apache Solr 3.5 with RankingAlgorithm 1.4.1

By

Nagendra Nagarajayya
transaxtions llc
http://solr-ra.tgels.org

Table of Contents

Apache Solr 3.5 with RankingAlgorithm 1.4.1

By

Nagendra Nagarajayya
http://solr-ra.tgels.org
updated 2012/03/05

1. Introduction

Apache Solr (a lightning fast, open source search platform) can now work with a new search library RankingAlgorithm instead of Lucene. Solr with RankingAlgorithm search seems to be comparable to Google site search (see Perl index comparison results) for certain queries and much better than Lucene.

RankingAlgorithm 1.4.1 enables Solr to rank product searches very accurately and also enables Near Real Time Search.

Three Algorithms are available SIMPLE, SIMPLE1 and COMPLEX. SIMPLE and SIMPLE1 are a very fast algorithms and can return queries in <50ms on a 10m wikipedia index (complete index). It can also scale to 100m docs or maybe more. SIMPLE1 is the fastest algorithm and is well suited for autocomplete type of processing. COMPLEX is a more complex algorithm so is a little slower compared to the SIMPLE, but can also still return queries in < 100ms on a 10m wikipedia index (complete index). COMPLEX is more accurate and should be able to give you the best rankings as compared to SIMPLE.

RankingAlgorithm has been integrated into Solr in such a way that either Lucene or the RankingAlgorithm can be used to do the search. RankingAlgorithm scoring does not break any of the existing functionality. So shard, faceting, highlighting, replication, etc. still work as before.

2. Using Solr with RankingAlgorithm

There is no change in the way you access Solr. All searches work the same as before.

So a Solr search such as:

http://localhost:8983/solr/?q=california gold rush&fl=score

should still work as before. The only difference is that Solr instead of using Lucene library for search, uses the RankingAlogrithm library to search and rank the documents. The returned scores are different from Lucene and reflects the relevancy of a document.

As said above, RankingAlgorithm scores in two different modes, Document mode and Product mode. In Document mode, the scoring is for relevance and in Product mode, scoring is for occurrence. Document mode is suitable for general purpose searches such as Wikipedia docs, HTML, Word/PDF or similar docs. The Document mode is the default. Product mode is for searches found on retail stores, online store/shopping/comparison/auction websites, etc, including short text sites like tweeter.

Product mode takes a term occurrence into account and scores accordingly. For eg. a search for “wii console” will show titles starting with “wii console” are first, and the others rank lower as the occurrence of “wii console” shifts in the title or gets reversed, see below:

Wii Console and Wii Fit Plus with Balance Board Bundle (Nintendo Wii)
Wii Console System with Wii Sports Resort Game with TWO MotionPlus Attachments
Nintendo **Wii Console** w/ Bonus Wii Sports Resort Bundle, Black
Pelican Accessories **Wii Console** Stand - Nintendo Wii
Grafitti Skin for Nintendo **Wii Console**
NEW AC Adapter Cable Cord Power Supply For NINTENDO **WII** Gaming **Console**
Wii Remote Charging **Console** Stand
Nintendo **Wii** Skin - System **Console** Skin and two Wii Remote Skins - Blue Vortex
CET Domain 10301901 **Console** Stand Station for Nintendo **Wii**

There is also a scan attribute, where the scan can be a fast scan, medium scan or a full scan. Scan is the depth of the search so can be fast, slower or slow. The default is fast scan.

2.1 Enabling RankingAlgorithm

Add

```
<library>rankingalgorithm</library>
```

to solrconfig.xml

RankingAlgorithm can be enabled by adding the above line to solrconfig.xml. To use the SIMPLE algorithm, use:

```
  <library>rankingalgorithm</library>
<rankingalgorithm>
    <algorithm>simple</algorithm>
</rankingalgorithm>
```

To use the SIMPLE1 algorithm, use:

```
  <library>rankingalgorithm</library>
<rankingalgorithm>
    <algorithm>SIMPLE1</algorithm>
</rankingalgorithm>
```

To use the COMPLEX algorithm, use:

```
  <library>rankingalgorithm</library>
<rankingalgorithm>
    <algorithm>COMPLEX</algorithm>
</rankingalgorithm>
```

Default is SIMPLE algorithm.

To enable document mode use:

```
  <library>rankingalgorithm</library>
<rankingalgorithm>
    <algorithm>COMPLEX</algorithm>
    <mode>document</document>
</rankingalgorithm>
```

To enable product mode, enable:

```
  <library>rankingalgorithm</library>
<rankingalgorithm>
    <algorithm>COMPLEX</algorithm>
    <mode>product</document>
</rankingalgorithm>
```

Default is document mode.

To enable scan,

```
  <library>rankingalgorithm</library>
<rankingalgorithm>
    <algorithm>complex</algorithm>
    <mode>product</document>
    <scan>medium</document>
</rankingalgorithm>
```

Default is fast scan.

SIMPLE/SIMPLE1 algorithms functions only in document mode. COMPLEX algorithm is more accurate but a little slow compared to SIMPLE algorithms. SIMPLE is very fast

and can return queries on the wikipedia index in < 50 ms and can also scale to 100 documents. SIMPLE algorithm is also very good and may be well suited than COMPLEX for some type of queries.

2.2 Enabling mode and library dynamically

Document and Product mode can be enabled by adding "mode=product" to the search query. For eg. If the search is for "wii console":

http://localhost:8983/solr/?q=wii console&fl=score&mode=*product*

To use document mode:

http://localhost:8983/solr/?q=wii console&fl=score&mode=*document*

To change scan:

http://localhost:8983/solr/?q=wii console&fl=score&scan=*medium*

To change library to Lucene:

http://localhost:8773/solr/?q=wii console&fl=score&library=lucene

To change the algorithm:

http://localhost:8773/solr/?q=wii console&fl=score&algorithm=complex

To use complex and product mode:

http://localhost:8773/solr/?q=wii console&fl=score&algorithm=complex&mode=product

2.3 Near Real Time Search

Solr with RankingAlgorithm also support Near Real Time Search (NRT). With NRT, any document added immediately becomes searchable without a commit. So documents

can be added in real-time with searches in parallel. As there is no commit, the indexing is very fast. A 10,000 TPS (documents added per sec) has been seen on a dual core intel x86_64 system Linux 11 (2.6 kernel) with 2GB heap, default collectors, parallel collectors enabled for the young generation. See http://solr-ra.tgels.com/wiki/en/Near_Real_Time_Search_ver_3.x for more info.

Steps to enable RT

Add

```
<realtime visible="200">true</realtime>
<library>rankingalgorithm</library>
```

to solrconfig.xml

visible attribute can be a number between 0 to any positive number, and is counted in milliseconds. Visible is a real-time delay that controls how many milliseconds are tolerable before documents are visible in searches. So visible="200" means, any document added will be visible in searches within 200 ms. Setting visible="0" means documents are immediately available in searches but update performance is poor at this granularity.

To enable realtime faceting add:

```
<realtime visible="200" facet="true">true</realtime>
<library>rankingalgorithm</library>
```

Note: Enabling realtime faceting with high document update will lead to performance issues.

2.4 Configuring options in solrconfig.xml

```
<realtime visible="200" facet="false">true</realtime> <!-- true to enable
near real time or false -->
<library>rankingalgorithm</library> <!--rankingalgorithm or lucene -->
<rankingalgorithm>
   <algorithm>simple</algorithm> <!-- simple, simple1 or complex -->
   <mode>document</mode> <!-- document or product mode -->
    <scan>fast</scan> <!-- fast, medium, full works in product mode -->
</rankingalgorithm>
```

3. Installing Solr with the RankingAlgorithm

You can install Solr with RA in two different ways. You can download Solr3.x with RA.zip a bundle of Apache Solr 3.x and Ranking Algorithm (a big download) or just download the solr-ra 3.x.war.zip which is a web archive file and copy it into an existing or new Solr 3.x installation. Below the are steps for both:

3.1 Download Solr 3.5 with RA.zip (bundle)

Installation is the same as Solr. Download Solr 3.x with RA.zip (Solr version 3.5 with the RankingAlgorithm) from solr-ra.tgels.com. Unzip or untar the file, change to examples directory and start Solr as before, java -jar start.jar

Step1:

Download Solr3.5 with RA.zip from http://solr-ra.tgels.com

Step2:

Unzip it to a directory

Step3:

cd unzip directory/apache-solr-3.5.0/examples

Step4:

bash -x start_solr.sh

or

java -Xmx 2gb -jar start.jar.

Step 5:

Configuring options in solrconfig.xml:

```
<realtime visible="200">true</realtime> <!-- true to enable near real time or
false -->
<library>rankingalgorithm</library> <!--rankingalgorithm or lucene -->
```

```
<rankingalgorithm>
     <algorithm>simple</algorithm> <!-- simple, simple1 or complex -->
   <mode>document</mode> <!-- document or product mode -->
    <scan>fast</scan> <!-- fast, medium, full works in product mode -->
</rankingalgorithm>
```

3.2 Download Solr 3.5 with RA.war (war file)

Instead of downloading the Solr 3.5 with RA (a huge file), you can download just the solr_ra.war file. You will still need to download the Solr 3.5 from the Solr website as below. Unzip that first, and then change to the examples directory and follow the steps as below.

Step1:

Download Solr 3.5 from the Apache Solr website, as in here:
http://www.apache.org/dyn/closer.cgi/lucene/solr/

Step2:

Install Solr 3.5 by unzip it to a directory.

Step3:

Download the solr_ra war file from solr-ra.tgels.com, as in here:
http://solr-ra.tgels.com/solr-ra.jsp
(click on download war file link at the bottom of the page)

Step4:

cp solr_ra.war.zip file to the examples directory under unzip directory/apache-solr-3.5.0/examples.

Step5:

unzip solr_ra.war.zip

Step 6:

cp solr.war webapps

Step 7:

bash -x start_solr.sh

or

java -Xmx 2gb -jar start.jar.

Step 8:

Configuring options in solrconfig.xml

```
<realtime>true</realtime> <!-- true to enable near real time or false -->
<library>rankingalgorithm</library> <!--rankingalgorithm or lucene -->
<rankingalgorithm>
    <algorithm>simple</algorithm> <!-- simple, simple1 or complex -->
   <mode>document</mode> <!-- document or product mode -->
    <scan>fast</scan> <!-- fast, medium, full works in product mode -->
</rankingalgorithm>
```

3.5 Installing on Glassfish/Tomcat/JBoss/WebLogic

If you want to deploy Solr on a different container than the default Jetty container, then deploy as before (ie. Deploy examples/webapps/solr.war on Tomcat or Glassfish or Weblogic or Jboss).

4. Using the RankingAlgorithm library

Download the RankingAlgorithm 1.4.1 jar file from here:

http://solr-ra.tgels.com/rankingalgorithm.jsp

(Click on download link)

Add the rankingalgorithm_1.4.1.jar file to your classpath.

Using RankingAlgorithm to search is extremely simple since it make uses of the

Lucene index to access the index. If you already have a Lucene Index, then you can use that as the first argument, see example code below or at http://solr-ra.tgels.com/downloads/code/Example.java:

Example 1:

```
            RankingQuery rq = new RankingQuery();
            IndexSearcher is = new IndexSearcher(index);
            StandardAnalyzer analyzer = new StandardAnalyzer();
            QueryParser parser = new QueryParser(field, analyzer);
            Query query = parser.parse(searchterms);
            RankingHits rh = rq.search(query, is); //is = Lucene IndexSearcher
object
            System.out.println("num hits=" + rh.getNumHits() + "--no docs=" +
is.maxDoc());
            for (int i=0; i<rh.getNumHits() && i<10; i++) {
                  System.out.println("i=" + i + "--" + rh.score(i) + "--docid=" +
rh.docid(i) + "--doc=" + rh.doc(i).get(title) );
            }
```

Example 2:

```
            IndexReader reader = IndexReader.open(FSDirectory.open(new
File(index)));
            RankingQuery rq = new RankingQuery();
            StandardAnalyzer analyzer = new StandardAnalyzer(Version.LUCENE_30);
            QueryParser parser = new QueryParser(Version.LUCENE_30, field,
analyzer);
            Query query = parser.parse(searchterms);
            TopScoreDocCollector tdc = TopScoreDocCollector.create(1000, true);
            rq.search(query, null, reader, tdc); //is = Lucene IndexSearcher object
            int hits = tdc.getTotalHits();
            ScoreDoc sda[] = null;
            if (hits > 0) {
                  sda = tdc.topDocs().scoreDocs;
            }
            System.out.println("num hits=" + hits + "--no docs=" +
reader.maxDoc());
            for (int i=0; i<hits && i<10; i++) {
                  ScoreDoc sd = sda[i];
                  System.out.println("i=" + i + "--" + sd.score + "--docid=" +
sd.doc + "--doc=" + reader.document(sd.doc).get(title) );
            }
            reader.close();
```

Make sure Lucene is also in the classpath since the RankingLibrary uses it to access the index. That is it. Very simple to use but gets you very accurate and relevant

results.

5. Conclusion

Apache Solr with RankingAlgorithm enables very accurate and relevant searches with NRT capability . RankingAlgorithm ranking seems to be comparable to Google site search (see Perl index comparison results) and much better than Lucene.

SIMPLE/SIMPLE1 algorithm returns queries on a 10m wikipedia index in <50 ms. COMPLEX algorithm is more accurate but a little slower and can also return queries in <100ms. In document mode RankingAlgorithm ranks documents relevantly while ranking very accurately and precisely in the product mode. Document mode is very well suited for searching html, wikipedia, pdf/word type documents, while product works very well with short text as in retail websites, product comparison websites, short text messaging like twitter, etc. RankingAlgorithm with document and product mode is very well suited for the enterprise as well as the retail, ecommerce and websites.

The near real time search in Solr-RA works well and allows concurrent search with indexing in parallel without closing the IndexSearchers or clearing the cache providing the ability to offer searches in near real time. The indexing performance observed on a 2 core intel system with Fedora Linux 11 is about 5000 documents in about 498 ms when inserting 1 million documents in batches of 5k to the MbArtists index.

6. References

1. Solr with RA, http://solr-ra.tgels.com/solr-ra.jsp
2. RankingAlgorithm, http://rankginalgorithm.tgels.com/rankingalgorithm.jsp
3. Apache Solr, http://projects.apache.org/projects/solr.html
4. Apache Lucene, http://projects.apache.org/projects/lucene_java.html