Near Real Time Search With Apache Solr 3.x and RankingAlgorithm 1.2

By
Nagendra Nagarajayya
http://solr-ra.tgels.com
Updated 2011/08/18

Summary

Apache Solr 3.x is a very popular open source search platform which uses Lucene as the underlying search library. Solr 3.x with RankingAlgorithm (Solr-RA) uses RankingAlgorithm 1.2 as the underlying search library. RankingAlgorithm uses Lucene indexing to read and write documents but scores and ranks on its own. Solr-RA enables adding documents to the index in Near Real Time (NRT) with concurrent searches. Updating or adding a document does not need commit, nor closing of the Index Searchers or clearing of the caches. As there is no commit, the indexing is very fast. A 10,000 TPS (document adds) has been observed with the MbArtists index (MbArtists index is the example index discussed in the Solr 1.4 Enterprise Search Server book by David Smiley and Eric Pugh).

Steps to enable RT

Add

```
<realtime visible="200">true</realtime>
<library>rankingalgorithm</library>
<rankingalgorithm>
     <mode>document</mode>
     <algorithm>simple</simple>
</rankingalgorithm>
```

to solrconfig.xml

Adding documents

No changes to adding documents except, you don't need to call commit after you

add a document. Commit is only needed if the index is empty and to create the first document. After that no commits are needed. See below example:

Example:

Example:

```
curl "http://localhost:8983/solr/update/csv?stream.file=/tmp/x1.csv&encapsulator=
%1f"

Note:
1. 500 mbartists records are updated at a time (/tmp/x1 contains 500 records)
2. A script generated the 500 martists records and called curl as above to load
3. The performance measurements are done after 10000 records are added.
```

Note: you need to add the commit parameter only for the first document when starting indexing with an empty index

Search concurrently while the indexing is going on.

As before, no changes.

http://localhost:8983/solr/select/?q=john&fl=score

Performance

Indexing:

```
Indexing about 10000 mbartist entries with curl, visible attribute set to 200
(after server has been warmed up)

time:

real    0m45.794s
user    0m8.908s
sys     0m24.674s


Time measure of shell script running time without curl [create mbartist entries,
etc]:

real    0m44.860s
user    0m9.172s
sys     0m26.045s

So time to load 10000 documents = 45.794 - 44.860 = ~1 secs (10000 docs/sec)
```

Concurrent search during load:

```
http://192.168.1.126:8983/solr/twitter/select/?fl=score&q=john ab180027&fl=score
```

Eliminating duplicates in an update:

If documents have unique id and multiple documents with the same unique id are added and if only the last document updated should be visible in searches add the following to solrconfig.xml :

Search for `<indexDefault>` and then under `<indexDefaults>`, look for `<maxBufferedDocs>`. Add below `<maxBufferedDocs>`,

```
<maxBufferedDeleteTerms>1</maxBufferedDeleteTerms>
```

Implementation

The Near Real Time has been implemented by retrieving the IndexReader from the IndexWriter.getReader() method after a document has been added to the index. The addDoc function in DirectHandlerUpdate2.java has been modified so that retrieved IndexReader is stored in a HashMap in SolrCore.java. To avoid locking, a non locking concurrent time managed access is used to make available the IndexReader to SolrIndexSearchers. The SolrIndexSearchers access this IndexReader instead of the SolrIndexReader and pass this as a parameter to RankingAlgorithm for the search. RankingAlgorithm uses the reader to access the index and returns the results which are in near real time as it is using the updated IndexReader.

The NRT implementation supports faceting, filter queries, etc. The faceting count can be seen changing as documents are added in the screenshots below Fig 1 and Fig2. Fig 1 shows a facet query for “john” from the mbartists index (from the book Solr-14-

Enterprise-Search-Server). Fig 2 shows the same query after adding a new artist to the index as below:

```
curl "http://localhost:8990/solr/mbartists/update/csv?stream.file=/tmp/x.csv&encapsulator=%1f"
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">163</int></lst>
</response>
```

```
cat /tmp/x:
id,type,a_name,a_name_sort,a_alias,a_type,a_begin_date,a_end_date,a_member_name,a_member_id,a_release_d
ate_latest,a_spell,a_spellPhrase,r_name,r_name_sort,r_name_facetLetter,r_a_name,r_a_id,r_attributes,r_t
ype,r_official,r_lang,r_tracks,r_event_country,r_event_date,r_event_date_earliest,l_name,l_name_sort,l_
type,l_begin_date,l_end_date,t_name,t_duration,t_a_id,t_a_name,t_num,t_r_id,t_r_name,t_r_attributes,t_r
_tracks,t_trm_lookups,word,includes
Artist:3991866,Artist,John Ab Davis,John Ab Davis,,person,1942-12-29T00:00:00Z,1999-12-
10T00:00:00Z,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

Fig 1, shows numFound as 13261, and the facet count for “john” as 13261. Fig 2 after adding a doc with curl shows 13262, and the facet count for “john” as 13262. The facet counts for Ab and Davis also change to 10038 and 10020. The Solr query is as below:

http://192.168.1.126:8990/solr/mbartists/select/?q=john&facet=on&facet.field=a_name&facet.field=a_type&fl=score

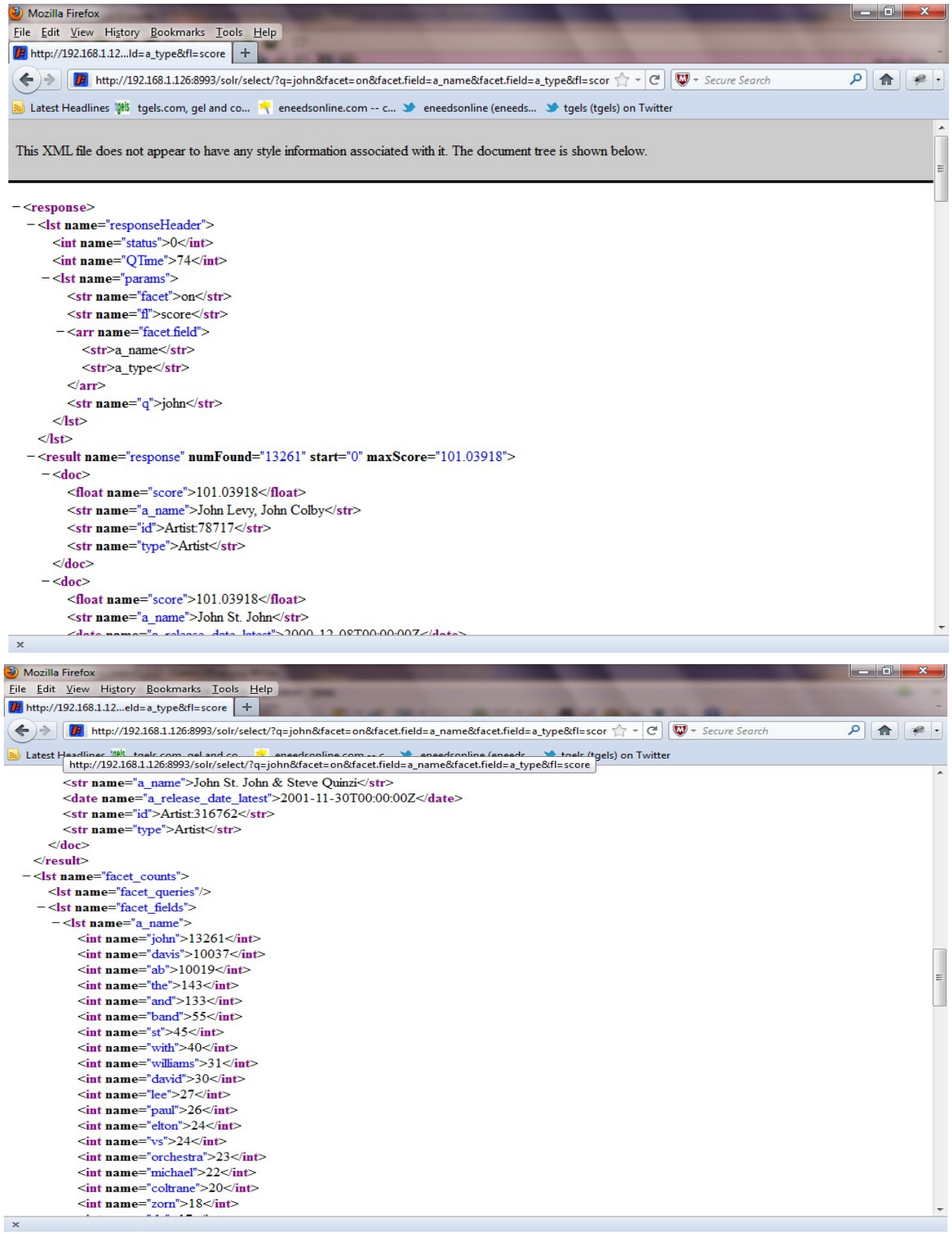
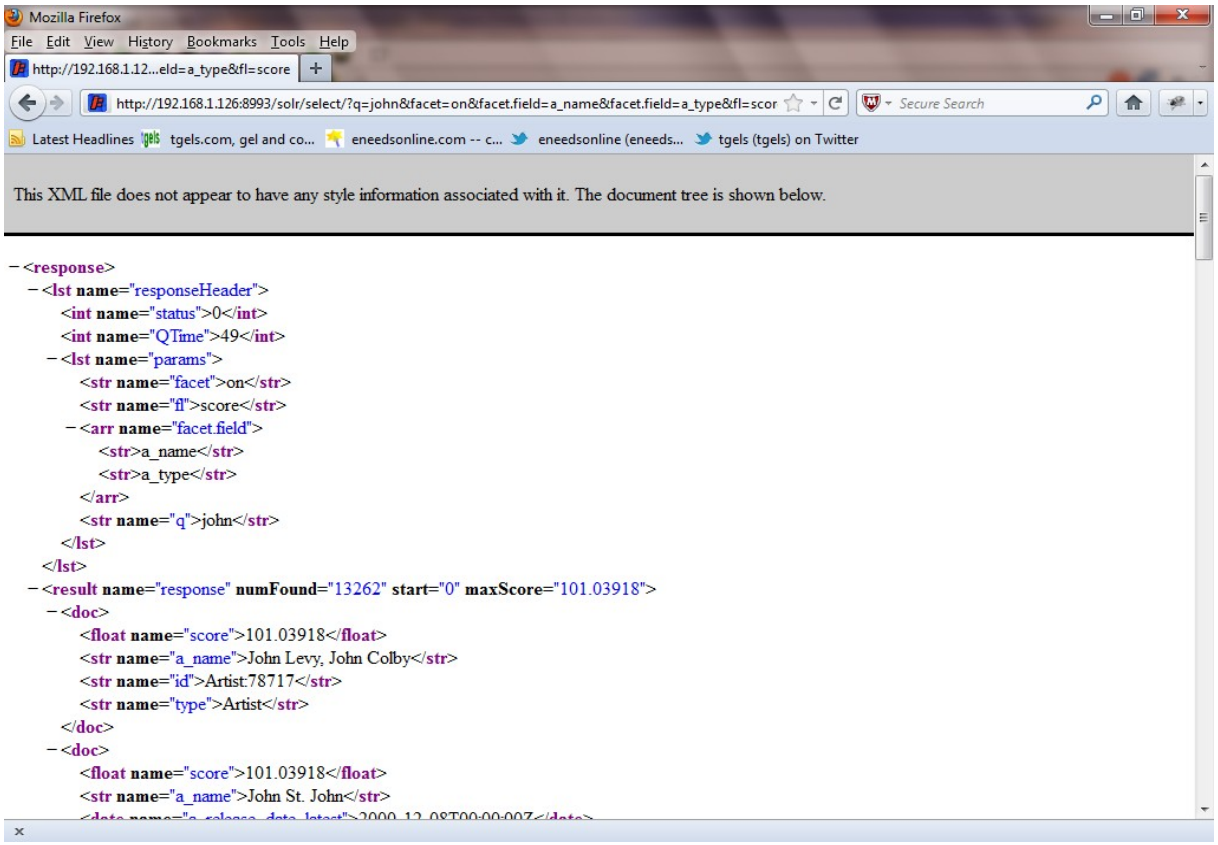

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://192.168.1.12...ld=a_type&fl=score
http://192.168.1.126:8993/solr/select/?q=john&facet=on&facet.field=a_name&facet.field=a_type&fl=scor
Secure Search
Latest Headlines tgels.com, gel and co... eneedsonline.com -- c... eneedsonline (eneeds... tgels (tgels) on Twitter
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">74</int>
<lst name="params">
<str name="facet">on</str>
<str name="fl">score</str>
<arr name="facet.field">
<str>a_name</str>
<str>a_type</str>
</arr>
<str name="q">john</str>
</lst>
</lst>
<result name="response" numFound="13261" start="0" maxScore="101.03918">
<doc>
<float name="score">101.03918</float>
<str name="a_name">John Levy, John Colby</str>
<str name="id">Artist:78717</str>
<str name="type">Artist</str>
</doc>
<doc>
<float name="score">101.03918</float>
<str name="a_name">John St. John</str>
Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://192.168.1.12...eld=a_type&fl=score
http://192.168.1.126:8993/solr/select/?q=john&facet=on&facet.field=a_name&facet.field=a_type&fl=scor
Secure Search
http://192.168.1.126:8993/solr/select/?q=john&facet=on&facet.field=a_name&facet.field=a_type&fl=score
<str name="a_name">John St. John & Steve Quinzi</str>
<date name="a_release_date_latest">2001-11-30T00:00:00Z</date>
<str name="id">Artist:316762</str>
<str name="type">Artist</str>
</doc>
</result>
<lst name="facet_counts">
<lst name="facet_queries"/>
<lst name="facet_fields">
<lst name="a_name">
<int name="john">13261</int>
<int name="davis">10037</int>
<int name="ab">10019</int>
<int name="the">143</int>
<int name="and">133</int>
<int name="band">55</int>
<int name="st">45</int>
<int name="with">40</int>
<int name="williams">31</int>
<int name="david">30</int>
<int name="lee">27</int>
<int name="paul">26</int>
<int name="elton">24</int>
<int name="vs">24</int>
<int name="orchestra">23</int>
<int name="michael">22</int>
<int name="coltrane">20</int>
<int name="zorn">18</int>

Fig 1


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://192.168.1.12...eld=a_type&fl=score
http://192.168.1.126:8993/solr/select/?q=john&facet=on&facet.field=a_name&facet.field=a_type&fl=scor
Secure Search
Latest Headlines tgels.com, gel and co... eneedsonline.com -- c... eneedsonline (eneeds... tgels (tgels) on Twitter
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">49</int>
<lst name="params">
<str name="facet">on</str>
<str name="fl">score</str>
<arr name="facet.field">
<str>a_name</str>
<str>a_type</str>
</arr>
<str name="q">john</str>
</lst>
</lst>
<result name="response" numFound="13262" start="0" maxScore="101.03918">
<doc>
<float name="score">101.03918</float>
<str name="a_name">John Levy, John Colby</str>
<str name="id">Artist:78717</str>
<str name="type">Artist</str>
</doc>
<doc>
<float name="score">101.03918</float>
<str name="a_name">John St. John</str>

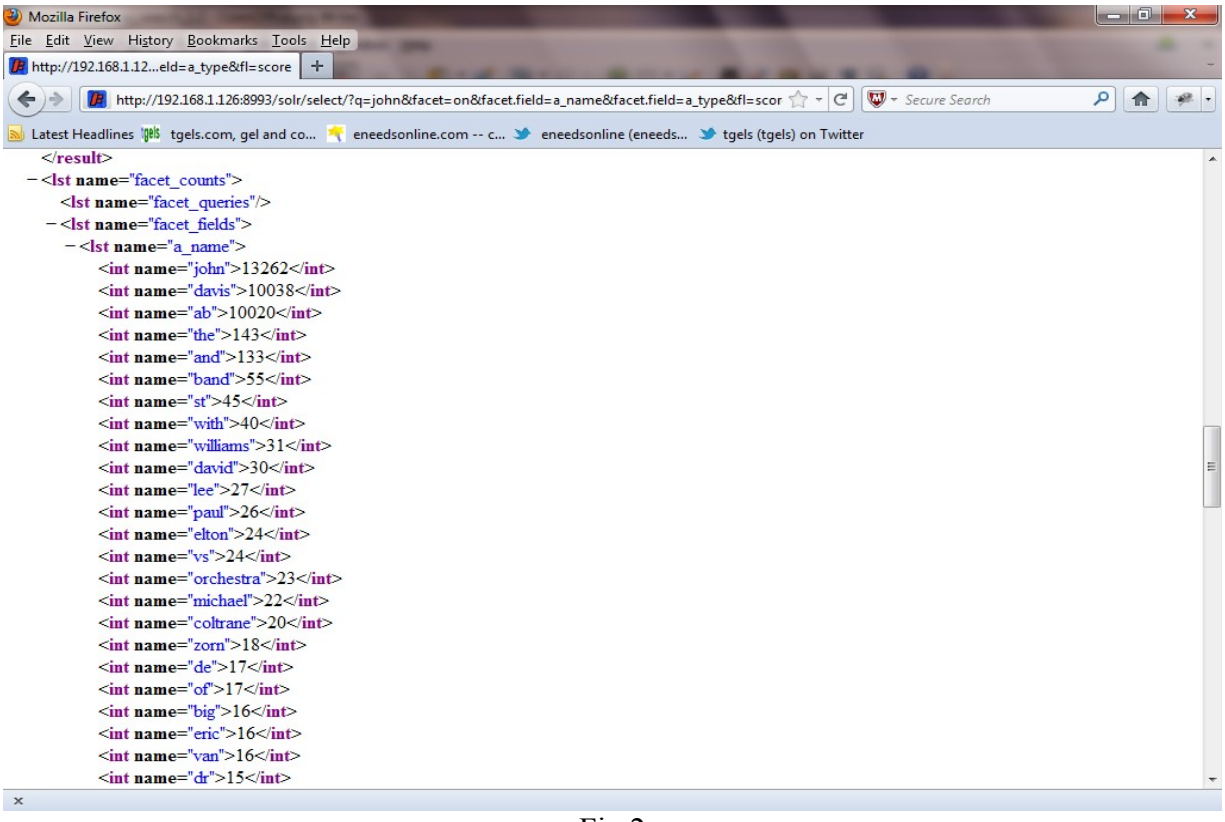

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://192.168.1.12...eld=a_type&fl=score
http://192.168.1.126:8993/solr/select/?q=john&facet=on&facet.field=a_name&facet.field=a_type&fl=scor
Secure Search
Latest Headlines tgels.com, gel and co... eneedsonline.com -- c... eneedsonline (eneeds... tgels (tgels) on Twitter
</result>
<lst name="facet_counts">
<lst name="facet_queries"/>
<lst name="facet_fields">
<lst name="a_name">
<int name="john">13262</int>
<int name="davis">10038</int>
<int name="ab">10020</int>
<int name="the">143</int>
<int name="and">133</int>
<int name="band">55</int>
<int name="st">45</int>
<int name="with">40</int>
<int name="williams">31</int>
<int name="david">30</int>
<int name="lee">27</int>
<int name="paul">26</int>
<int name="elton">24</int>
<int name="vs">24</int>
<int name="orchestra">23</int>
<int name="michael">22</int>
<int name="coltrane">20</int>
<int name="zorn">18</int>
<int name="de">17</int>
<int name="of">17</int>
<int name="big">16</int>
<int name="eric">16</int>
<int name="van">16</int>
<int name="dr">15</int>

Fig 2

Caveat

1. The performance is limited by how fast the IndexWriter.getReader() returns. This seems to take the most time between 2ms to 70ms avg. The faster this goes, the faster the index time.

2. Caching needs to be disabled at the moment to see NRT updates, as with cache enabled, the first time a search is executed, the results are cached and for the next matching search the results are retrieved directly from cache. The solution here is to look at the docs added and to invalidate/update the cache as needed based on the cache query but at 10000 docs / sec this will become the new bottleneck and may limit scalability.

3. Setting maxBufferedDeleteTerms=1 will slow down update performance.

4. Setting maxBufferedDeleteTerms=1 will remove duplicates with unique ids but search results may still show the old document content, if the new document added has changed content, even though the new document content is searchable ie.
<pre>
if the most recent doc has afield set to

```
<doc><afield>abc</afield></doc>
```

and this is updated, and the old docs were

```
<doc><afield>xyz</afield>,
```

at query time, q=afield:abc matches, but the results show may show

```
<doc><afield>xyz</afield>
```

Download

Download Solr-RA including tweet file and try it out yourself.

You can download Solr-RA from here:

http://solr-ra.tgels.com

(You can get the MbArtists schema with the download)

Conclusion

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 12 is about 10000 tps (new document adds) with visible set to 200ms.

Note:

1. solr and lucene are registered trandemarks of apache software foundation.