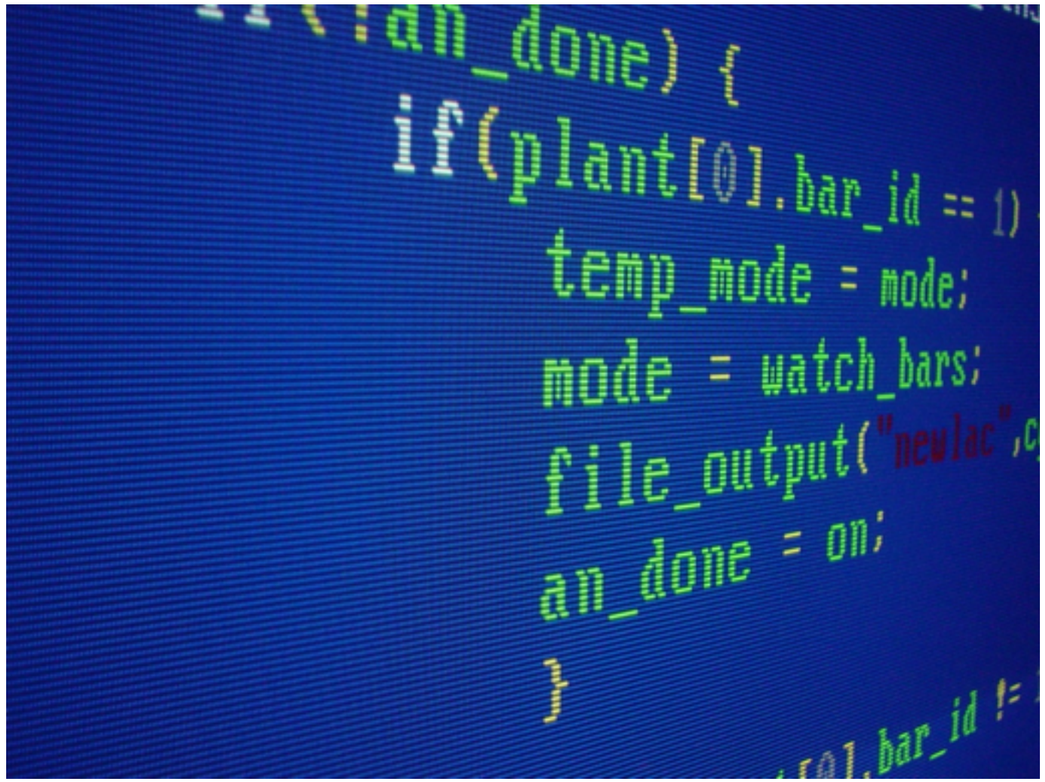
CERN Spring Campus 2014

Information Retrieval and Search Engines: Introduction to Solr

Patrick GLAUNER
CERN
patrick.oliver.glauner@cern.ch
April 14, 2014

About me

- Graduated from Karlsruhe University of Applied Sciences, Germany in 2012
- Software Engineer at CERN since 2011
- R&D focus on document management and search engines

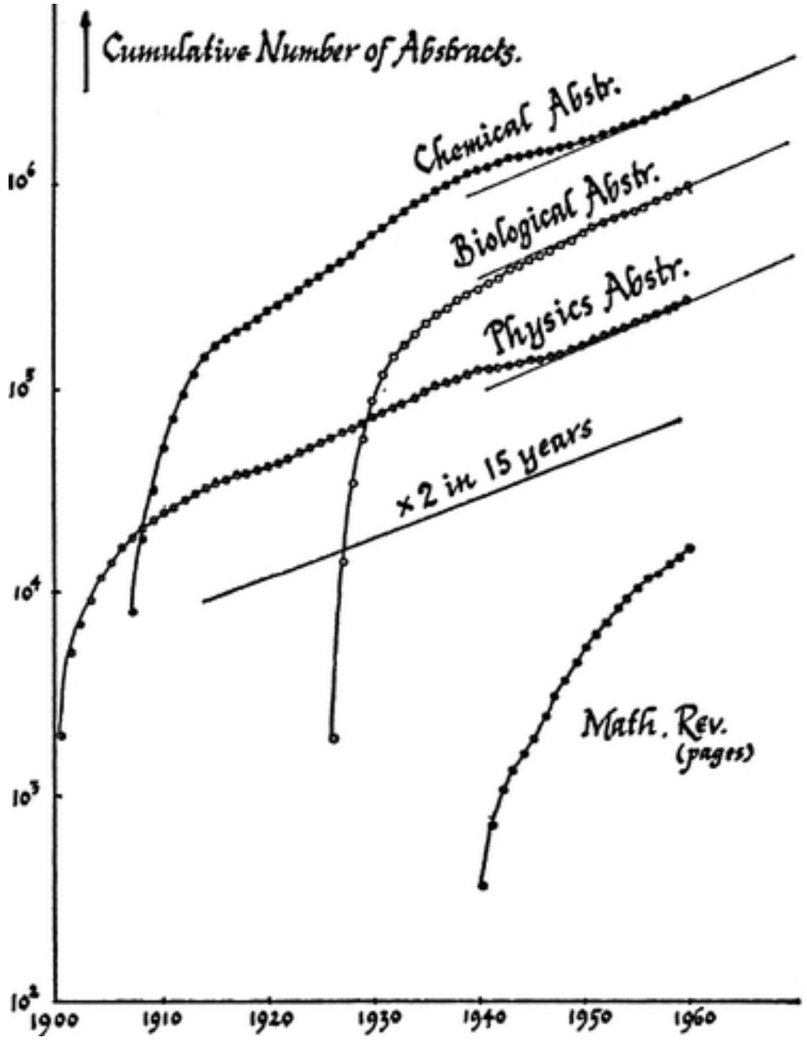


Agenda

- The Importance of Search
- Information Retrieval
- Introduction to Solr
- Solr-based Search and Ranking in Invenio
- Conclusions
- Q&A

The Importance of Search

- Why is search so important?
- Amount of data increases exponentially
- Different kinds of search engines: WWW, desktop, repository, etc.
- Effectivity: finding correct results
- Efficiency: performance, ranking, usability, etc.



P. Glauner: Information Retrieval and Search Engines: Introduction to Solr

The Importance of Search: History

The Importance of Search: History

The Importance of Search: Fundamental Capabilities

- Harvesting of data: submissions, crawling, etc.
- Construction of indexes to speed up search

Indexing —creates/ modifies→

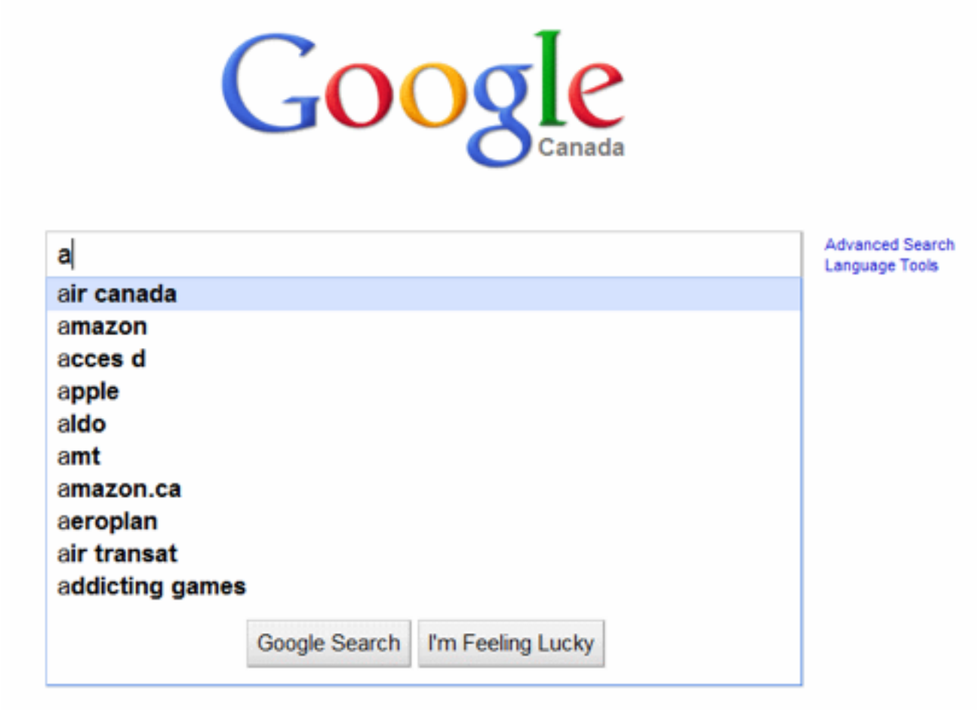
term	records
institute	4, 5, 7, 10, ...
nuclear	5, 6, 8, ...
research	1, 5, 20, ...
...	...

←reads— **Search**

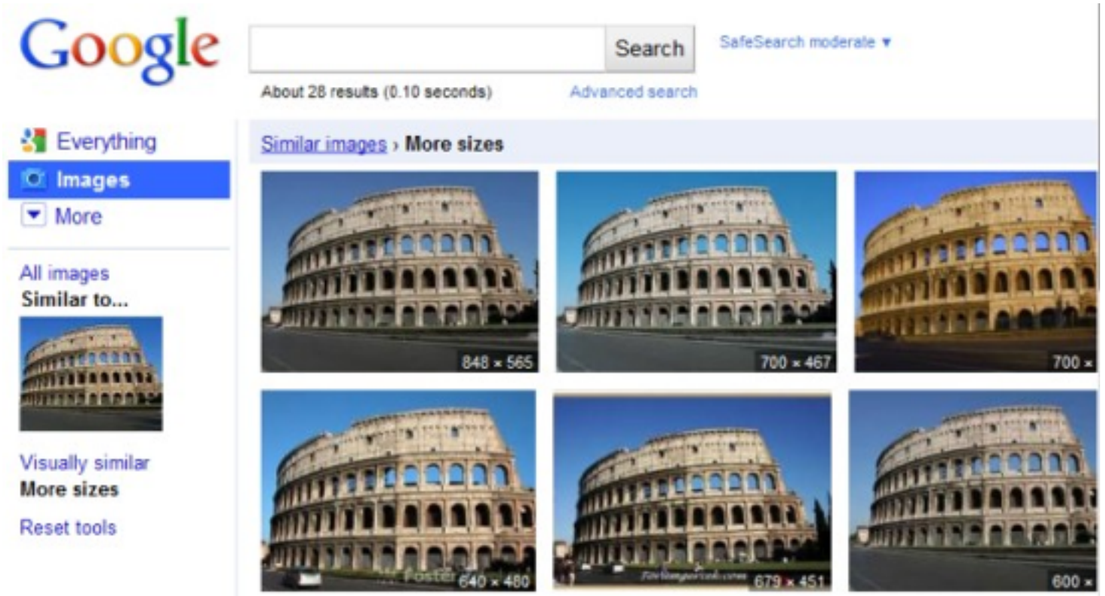
The Importance of Search: Fundamental Capabilities

- Case insensitivity: `computer≡Computer`
- Stemming: `computer≡computers`
- Stop words: removal of terms such as `at`, `of`, `at` among others
- Synonyms: `because≈as`
- ...

The Importance of Search: Advanced Capabilities



The Importance of Search: Advanced Capabilities



Information Retrieval

- "The tracing and recovery of specific information from stored data."
- Relevance ranking based on term frequency (word similarity ranking):

Query	nuclear research	Relevance
Result 1	I like nuclear research, computer science and mathematics	Low
Result 2	Employed in nuclear research	High

- Querying for near words
- Handling of misspellings
- ....

Introduction to Solr

- Wide-spread open source search server
- Written in Java
- Communication through HTTP
- Configuration in XML files
- Flexible data schema, NoSQL-like
- Rich indexing, search and IR capabilities
- http://lucene.apache.org/solr/ (Solr 4.7.1)

Introduction to Solr: Setup

- Prerequisites: a Linux/Mac OS X environment containing Java 6, Python 2.7 and solrpy:
  ```
  $ easy_install solrpy
  ```

1. Download Solr (~150MB):
   ```
   $ curl -O http://mirror.switch.ch/mirror/apache/dist/lucene/solr/4.7.1/solr-4.7.1.zip
   ```

2. Unzip Solr:
   ```
   $ unzip solr-4.7.1.zip
   ```

3. Start Solr:
   ```
   $ cd solr-4.7.1/example
   $ java -jar start.jar
   ```

Introduction to Solr: Indexing

4. Index some documents:

```
import solr

# create a connection to the Solr server
s = solr.SolrConnection('http://localhost:8983/solr')

# add documents to the index
s.add(id=1, title='The best solr book ever written', author='Author1')
s.add(id=2, title='The fastest computer', author='Author2')
s.add(id=3, title='Introduction to Solr', author='P. Glauner')

s.commit()
```

Introduction to Solr: Search and Ranking

5. Search documents:

```
import solr

# create a connection to the Solr server
s = solr.SolrConnection('http://localhost:8983/solr')

# query the index
response = s.query('title:Solr')

for hit in response.results:
    print '%s: %s' % (hit[u'score'], hit[u'title'][0])
```

6. Output:

```
0.5: Introduction to Solr

0.375: The best solr book ever written
```

Introduction to Solr: Configuration

- `schema.xml` defines the document structure and properties:

```
<field name="title" type="text_general" indexed="true" stored="true" multiValued="true"/>
<field name="subject" type="text_general" indexed="true" stored="true"/>
<field name="description" type="text_general" indexed="true" stored="true"/>
<field name="comments" type="text_general" indexed="true" stored="true"/>
<field name="author" type="text_general" indexed="true" stored="true"/>
<field name="keywords" type="text_general" indexed="true" stored="true"/>
```

- `text_general` defines tokenizer, stemmer and synonym filter among others

- Changes require a Solr restart (and reindexing in most cases)

Introduction to Solr: Further Topics

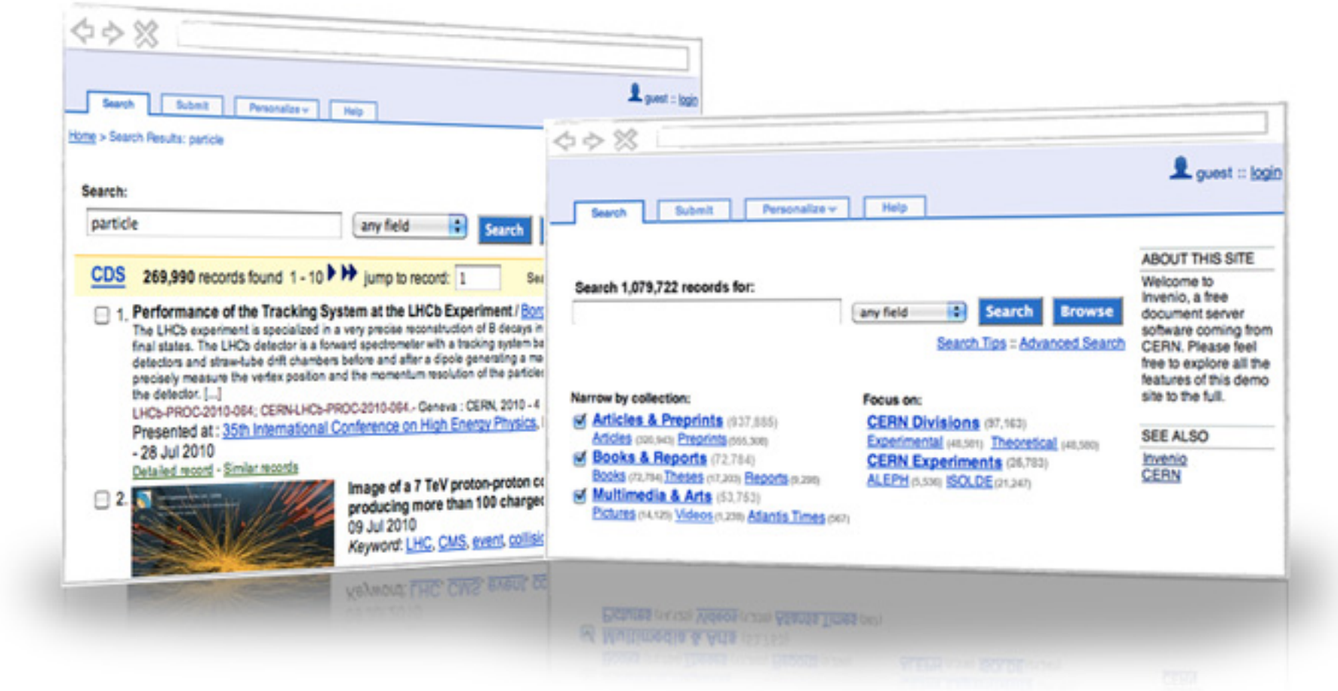
- Delete all data in Solr:

```
$ cd solr-4.7.1/example/exampledocs
$ java -Ddata=args -jar post.jar '<delete><query>*:*</query></delete>'
```

- solrpy tutorial: https://code.google.com/p/solrpy/
- Solr tutorial: https://lucene.apache.org/solr/tutorial.html

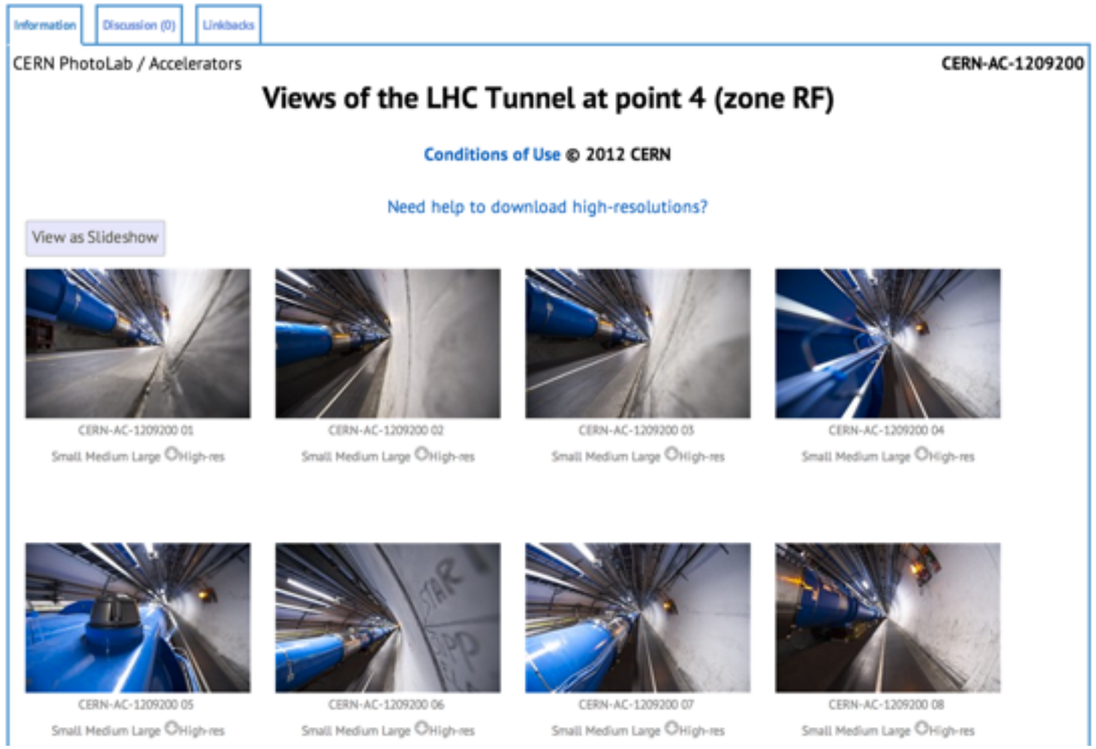
Solr-based Search and Ranking in Invenio

- Invenio is a free document management system
- Originally developed at CERN, nowadays by an international collaboration
- More than 30 instances around the globe
- http://invenio-software.org/

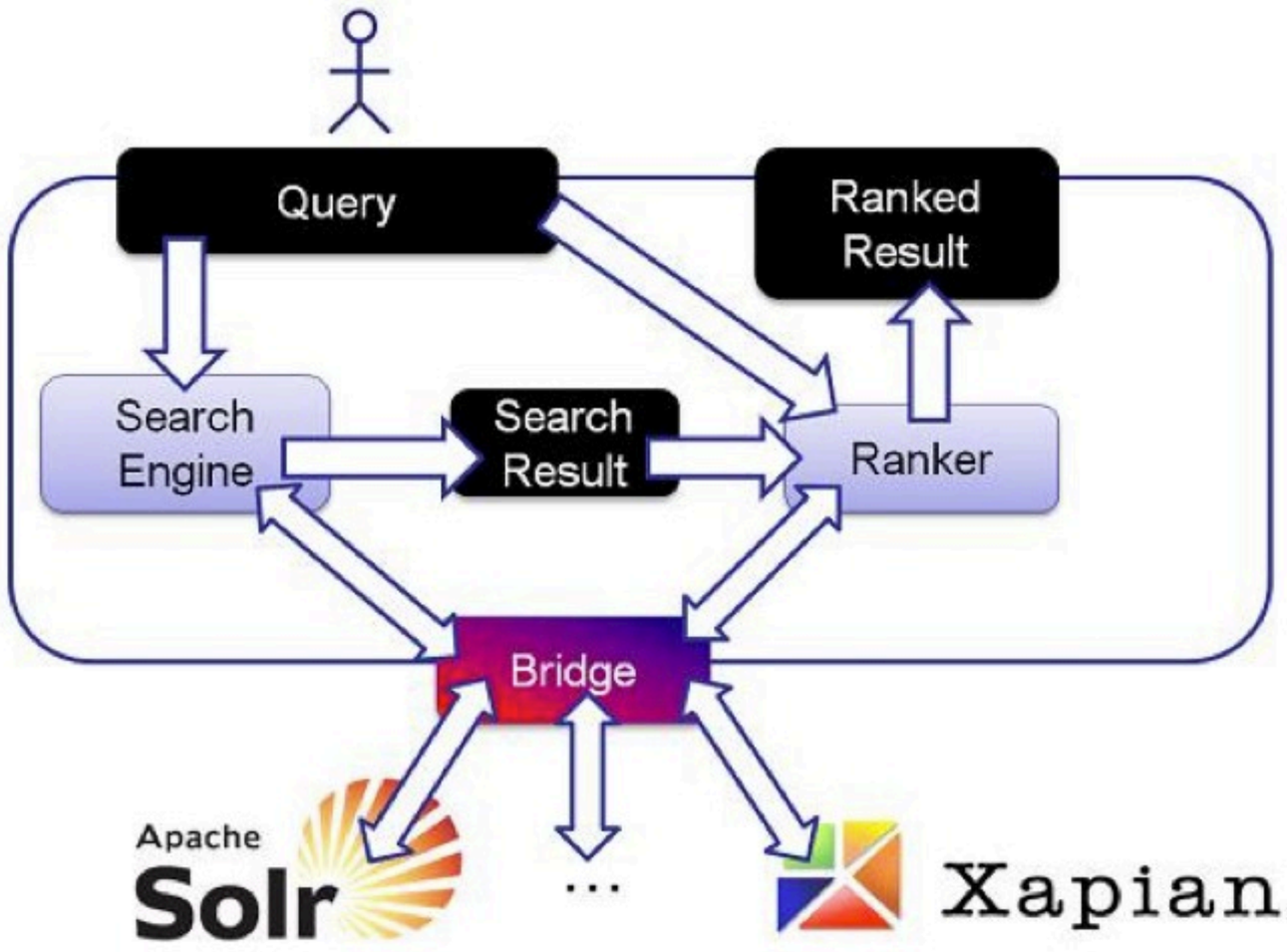


Solr-based Search and Ranking in Invenio

- The CERN Document Server (CDS) is one of the largest Invenio instances
- More than 1.5M records and 400K full-text documents in the high energy physics domain
- Receives more than 10K unique visits per day with occasional peaks of up to 350K visits per day during important physics announcements
- http://cds.cern.ch/



Solr-based Search and Ranking in Invenio



P. Glauner: Information Retrieval and Search Engines: Introduction to Solr

Solr-based Search and Ranking in Invenio

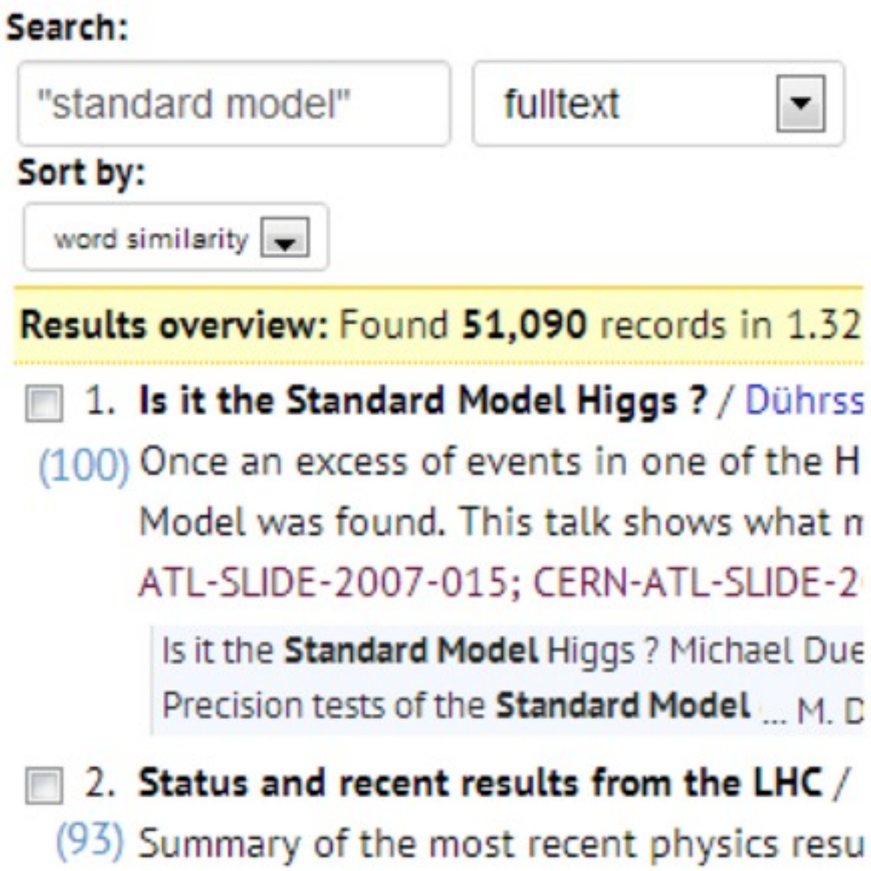
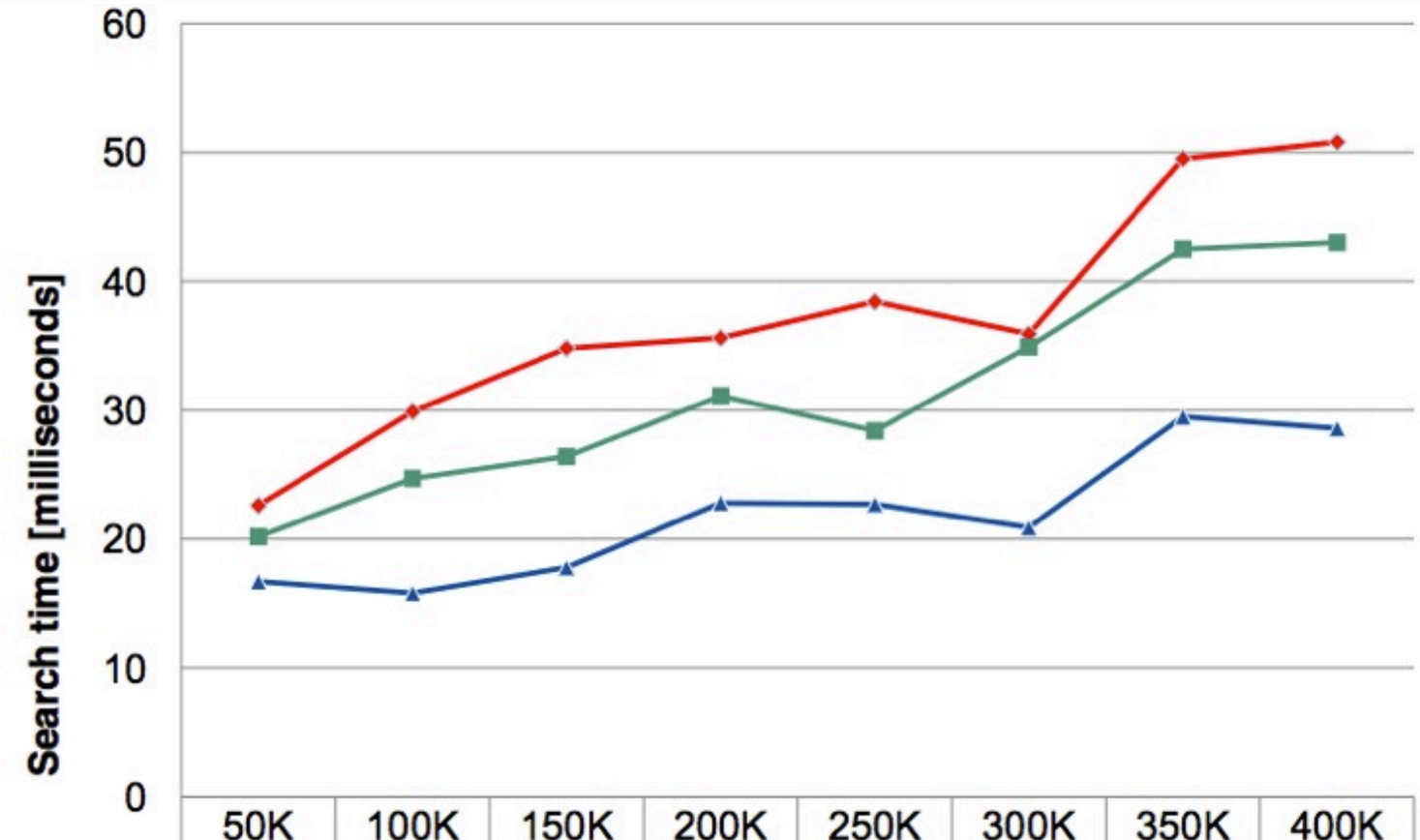


Figure 2: Solr-based word similarity result by percentage

- To benefit from Solr's efficient and scalable full-text search and word similarity ranking, it was integrated into Invenio
- Full-text search
- Word similarity ranking
- Snippet generation
- Finding records similar to a specific one
- Query correction

Solr-based Search and Ranking in Invenio



	50K	100K	150K	200K	250K	300K	350K	400K
of	22.6	29.9	34.8	35.6	38.4	35.9	49.5	50.8
model	20.2	24.7	26.4	31.1	28.4	34.9	42.5	43
boson	16.7	15.8	17.8	22.8	22.7	20.9	29.5	28.6

P. Glauner: Information Retrieval and Search Engines: Introduction to Solr

Conclusions

- Search engines build on top of indexes, algorithms and information retrieval theory
- Solr is a wide-spread open source search server
- It can be easily integrated into applications, e.g. Invenio
- Solr offers a rich set of features including search, ranking, snippets and finding records similar to a specific one
- It scales well for large corpuses

Q&A

References

- http://lucene.apache.org/solr/
- http://invenio-software.org/
- P. Glauner et al. Use of Solr and Xapian in the Invenio document repository software. Open Repositories Conference 2013. Charlottetown, Canada. arXiv:1310.0250.
- P. Glauner, T. Simko. Enhancing Invenio Digital Library With An External Relevance Ranking Engine. BSc thesis, Karlsruhe University of Applied Sciences, Germany. 2012. arXiv:1211.0689.