Localizing Integrated Search
Developments in integrated search

A story about business-centered IT development

Jens Hofman Hansen
jh@statsbiblioteket.dk
State and University Library, Denmark
Six years with the development of Search

Virtual integrated search. How we use Summon today

Spin-offs

Should you put everything behind one search field just because you can?
<table>
<thead>
<tr>
<th>Status</th>
<th>Titel</th>
<th>Forfatter</th>
<th>År</th>
<th>Type</th>
<th>Bestill</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hjemme</td>
<td>Foo Fighters</td>
<td>Foo Fighters</td>
<td>2000</td>
<td>node</td>
<td>Tilføj</td>
<td></td>
</tr>
<tr>
<td>Hjemme</td>
<td><strong>Best of Foo Fighters</strong></td>
<td>Troy Stetina</td>
<td>2000</td>
<td>node</td>
<td>Tilføj</td>
<td></td>
</tr>
<tr>
<td>Hjemme</td>
<td>Foo Fighters - Everywhere but home</td>
<td>Foo Fighters</td>
<td>2003</td>
<td>film</td>
<td>Tilføj</td>
<td></td>
</tr>
<tr>
<td>Hjemme</td>
<td>The colour and the shape</td>
<td>Foo Fighters</td>
<td>1999</td>
<td>node</td>
<td>Tilføj</td>
<td></td>
</tr>
<tr>
<td>Hjemme</td>
<td>Foo Fighters</td>
<td>David Grohl</td>
<td>2000</td>
<td>musik og lyd</td>
<td>Tilføj</td>
<td></td>
</tr>
<tr>
<td>2-8 uger</td>
<td>There is nothing left to lose</td>
<td>Foo Fighters</td>
<td>1999</td>
<td>musik og lyd</td>
<td>Tilføj</td>
<td></td>
</tr>
<tr>
<td>2-8 uger</td>
<td>In your honor</td>
<td>Foo Fighters</td>
<td>2005</td>
<td>musik og lyd</td>
<td>Tilføj</td>
<td></td>
</tr>
</tbody>
</table>

**Afgræns din søgning**

- Emner
  - **rock** (22)
  - **vokal** (21)
  - **usa** (20)
  - 2000 (12)
  - 1990 (11)
- **Vis alle (33)**

- **Forfattere**
  - foo fighters (29)
  - mondal nate (11)
  - grohl david (10)
  - hawkins taylor (9)
  - shiflet chris (6)
- **Vis alle (50)**

- **Materialtyper**
  - Musikfyldt (22)
  - Lyd på cd (20)
  - Noder (10)
  - Bog (2)
  - Film på dvd (2)
- **Vis alle (9)**
CONTENT: map and normalize metadata
INDEX: single unified search index including full text
RIGHTS: search appropriate resources
RESULTS: relevance ranked quick reliable
API: integrates with other systems
INTERFACE: clean simple easy to use
foo fighters

Limit your search

Subjects
- rock (25)
- vokal (23)
- usa (22)
- 2000 2009 (14)
- 1990 1999 (13)
Show more (13)

Authors
- foo fighters (32)
- mendel nate (12)
- greti david (11)
- hawkins taylor (10)
- nielsen chris (6)
Show more (15)

Material types
- music/sound (25)
- sound on CD (22)
- sheet music (10)
- movie on DVD (3)
- movie (3)
Show more (9)

Individual libraries
- statsbiblioteket (33)
- drh (1)

Full-text on "foo fighters" | See all 11192 articles

Books, journals and other physical materials about "foo fighters"
1 to 20 of 34

<table>
<thead>
<tr>
<th>Title</th>
<th>Author/creator</th>
<th>Year</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foo Fighters : guardian angels</td>
<td>Foo Fighters</td>
<td>2005</td>
<td>dvd</td>
<td>Available</td>
</tr>
<tr>
<td>Best of Foo Fighters</td>
<td>Troy Stetina</td>
<td>cop. 2000</td>
<td>sheet music</td>
<td>Available</td>
</tr>
<tr>
<td>Foo Fighters</td>
<td>Foo Fighters</td>
<td>[200-]</td>
<td>sheet music</td>
<td>Available</td>
</tr>
</tbody>
</table>
Limit your search
Subjects
- united states (230895)
- studies (178424)
- chemistry, physical (138508)
- materials science, multidisciplinary (130185)
- chemistry (128641)
Show more (30)

Authors
- aubert b (583)
- jodc (547)
- international energy agency (468)
- the babar collaboration (298)
- massachusetts institute of technology dept of physics (273)
Show more (15)

Tilgængelighed
Materialer der er tilgængelige
- Materialer der er tilgængelige inkl. internationale aviser og tidskrifter
- Alle materialer

Material types
- article (5059344)
- journal article (4822547)

Get help and guidance
Ask the library

Loans from other libraries
- Didn't find what you are looking for?
  - We'll get it for you

Twitter
- Statsbiblioteket is transcoding 6 years of radio - 350,000 hours. The process will run for 7 days on 4 computers, each with 12 CPU cores #39perta

Others are looking at

Energy (Oxford)
Multiple access points e-journal
Online

Energy
journal
Copy request

Energy
Gerard M. Crawley
1975 book
Available
Do you need the title right away?

ENERGY
Anonymous
2010 Journal Article, Boulder County Business Report
Local energy companies span from numerous entrepreneurial startups working with solar panels and biofuels to industry giants such as ConocoPhillips, GE Energy...
Online
from Summon

from local database
What is relevance ranking? (according to Lucene)

Purely mathematical:
- Documents have a base value
- Rare words have higher value
- Precise match has higher value
- Score is relative
- Fields can be boosted (title^2)
- Queries can be boosted (whale^0.5 dolphin^2)

\[ \text{coord}(d, q) \times \text{queryNorm}(q) \times \sum_{t \in q} (\text{tf}(t \in d) \times \text{idf}(t)^2 \times t\text{.getBoost()} \times \text{norm}(t)) \]

http://sbdevel.wordpress.com/
“dolphin”

<table>
<thead>
<tr>
<th>Repository A</th>
<th>Repository B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Red</td>
</tr>
<tr>
<td>Blue</td>
<td>Red</td>
</tr>
<tr>
<td>Blue</td>
<td>Red</td>
</tr>
<tr>
<td>Blue</td>
<td>Red</td>
</tr>
<tr>
<td>Blue</td>
<td>Red</td>
</tr>
<tr>
<td>Blue</td>
<td>Pink</td>
</tr>
<tr>
<td>Blue</td>
<td>Pink</td>
</tr>
<tr>
<td>Blue</td>
<td>Pink</td>
</tr>
<tr>
<td>Blue</td>
<td>Pink</td>
</tr>
<tr>
<td>Blue</td>
<td>Pink</td>
</tr>
<tr>
<td>Repository A</td>
<td>Repository B</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“dolphin”

Repository A

Repository B

Repository A+B
<table>
<thead>
<tr>
<th>“dolphin”</th>
<th>Repository A</th>
<th>Repository B</th>
<th>Repository A+B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

The values of documents (and the order of search results) are different because of the context they are in, field boosting and query boosting.

...this means...

Merging of two search results is not simple!
By exchanging term stats (*tf-idf weights*) a true merging with a technically perfect relevance ranking is possible.
energy

Energy (Oxford)
Multiple access points e-journal
Online

Energy
journal
Copy request

Energy
Gerard M. Crawley
1975 book
Available
Do you need the title right away?

ENERGY
Anonymous
2010 Journal Article, Boulder County Business Report
Local energy companies span from numerous entrepreneurial startups working with solar panels and biofuels to industry giants such as ConocoPhillips, GE Energy...
Online
User evaluation
1-2 weeks ago
There are still many improvements to be done:

Better ranking integration
True facets integration
Author and publication title index
Deeper insight into Summon architecture
Spin offs from controlling the service layer
Scientific articles for everyone
You can put everything in one search field, but does it mean that you should?
Manden, de kaldte hest 1. december 2007
Lord John Morgan bliver taget til fange af Sioux-indianerne og sættes på hårde manddom...

Bergerac: Ukendt størrelse 12. maj 2010
Charlie Hungerford står i spidsen for et syndikat, der ønsker...

- og så kom hævnens time 9. september 2007
Manuel Artiguez kæmpede i Den Spanske Borgerkrig på republikanernes side og har i 15 å...
Thank you!

Questions? :-)

Jens Hofman Hansen
jhh@statsbiblioteket.dk
State and University Library, Denmark