



# *SearX-Engine and INEX*

XML enabled probabilistic retrieval

**Holger Flörke**

doctronic GmbH & Co. KG



# doctronic GmbH & Co. KG



## ✘ Multi Channel Publishing

technology, tools, know-how

products  

## ✘ customers

professional publishers

technical documentation

...



# Query Language I



## ✘ *Role*

grouping structures with common semantics  
(eg. author <a>, <au>, <author>, ...)

## ✘ Abstraction for

the end-user

searching within heterogeneous collection





# Query Language III



- ✘ Term-Operators

  - must have +

  - must not have -

  - phrase ` . . . `

- ✘ Filter

- ✘ Scenario (named set of weightings)

  - ...



# Query Language IV



## × Example

```
<query name="INEX">
  <retrieval-role><constant value="article"/></retrieval-role>
  <filter/>
  <query-item>
    <role><constant value="article"/></role>
    <terms><constant value=" 'digital library' "/></terms>
  </query-item>
  <query-item>
    <role><constant value="article//p"/></role>
    <terms><constant value="+authorization 'access control'+security"/></terms>
  </query-item>
</query>
```



# Ranking I



$$\text{Score}_d = \text{TF}_{t,d} * \text{IDF}_t$$

$\text{Score}_e =$  „all weighted occurrences of  $t$  in  $e$ “

$$* \\ \text{IEF}_{t,s(e)}$$



# Ranking II



- ✘ Continue heuristics
- ✘ Filters
  - evaluated at runtime
- ✘ Structural inheritance
  - processed in a second step
- ✘ Index structures



# INEX Topics and Queries



```
<inex_topic ct_no="35" query_type="CO" topic_id="100">  
  <title>+association +mining +rule +medical</title>  
  <description> Retrieve information about association rule mining in medical databases </description>  
  <narrative> We have a medical data mining ... </narrative>  
  <keywords>association, mining, rule, medical</keywords>  
</inex_topic>
```



```
<query name="INEX">  
  <retrieval-role><constant value="article"/></retrieval-role>  
  <filter/>  
  <query-item>  
    <role><constant value="article"/></role>  
    <terms><constant value="+association +mining +rule +medical association mining rule medical"/> </terms>  
  </query-item>  
</query>
```



# INEX Topics and Queries



```
<inex_topic ct_no="14" query_type="CAS" topic_id="63">
  <title>
    //article[about(., "digital library") AND about(./p, '+authorization +access control' +security')]
  </title>
  [...]
</inex_topic>
```



```
<query name="INEX">
  <retrieval-role><constant value="article"/></retrieval-role>
  <filter/>
  <query-item>
    <role><constant value="article"/></role>
    <terms><constant value=" 'digital library' "/></terms>
  </query-item>
  <query-item>
    <role><constant value="article//p"/></role>
    <terms><constant value="+authorization +'access control' +security"/></terms>
  </query-item>
</query>
```



# Summary



- ✘ *Role* abstracts from the collection structure
- ✘ *SearX-Engine* integrates structure based on roles into TF\*IDF
- ✘ Overview of the index structures
- ✘ CAS topics can be mapped





# doctronic

Information Publishing + Retrieval

Adenauerallee 45-49  
53332 Bornheim  
<http://www.doctronic.de>

Fon (0 22 22) 92 92 90  
Fax (0 22 22) 92 92 99  
E-Mail [info@doctronic.de](mailto:info@doctronic.de)