Exploring Automated Patent Search with KNIME
Possibilities, Limits, Future

Alexander Klenner-Bajaja, PhD
aklenner@epo.org
European Patent Office

**Offices:** Berlin, Vienna, Munich, The Hague (Rijswijk), Brussels

**Staff:** over 7000

**Mission:** Granting European Patents

New building in Rijswijk, finished approx. 2018
Why Search – European Patent Convention

Article 4
European Patent Organisation

(1) A European Patent Organisation, hereinafter referred to as the Organisation, is established by this Convention. It shall have administrative and financial autonomy.

(2) The organs of the Organisation shall be:

(a) the European Patent Office;

(b) the Administrative Council.

(3) The task of the Organisation shall be to grant European patents. This shall be carried out by the European Patent Office supervised by the Administrative Council.

Article 92
Drawing up of the European search report

The European Patent Office shall, in accordance with the Implementing Regulations, draw up and publish a European search report in respect of the European patent application on the basis of the claims, with due regard to the description and any drawings.

Information Management`s Task: Support Search
Patent Search – Filing numbers increase

2005 -- 2015
Introduction – What do we want, where are we?

Fully manual

Application

Query Formulation

Search

PriorArt

relevant documents

Fully automated
The EPO`s Tool landscape

- Translation API
- Boolean Search Engine
- Lucene Search Engine
- MetaData Neo4J
- GoldStandard SQL
- ImageDB
- Fulltext MongoDB
- Search Evaluation
- Concept Extraction
The EPO's Tool landscape

- Translation API
- Boolean Search Engine
- Lucene Search Engine
- Fulltext MongoDB
- ImageDB
- MetaData Neo4J
- GoldStandard SQL
- Search Evaluation
- Concept Extraction
The EPO`s Tool landscape

Translation API

Lucene Search Engine

Boolean Search Engine

GoldStandard SQL

Search Evaluation

Concept Extraction

ImageDB

Fulltext MongoDB

MetaData Neo4J

Translation

Evaluator Search Results

Get Gold Standard

OscarKeywordExtractor

Epoqe

Node 2015

Node 2017

Node 2018

Node 1298

Normal Search

Downloading Application Query Test

GetKimeDrawings

Node 2014
The EPO's Tool landscape

- Translation API
- Boolean Search Engine
- Lucene Search Engine
- MetaData Neo4J
- Fulltext MongoDB
- ImageDB
- Search Evaluation
- GoldStandard SQL
- Concept Extraction
- Translation
- Epoque
- Get Gold Standard
- Evaluate Search Results
- ImageDB
- Neo4J
- MQA
- MySQL
- Node 2018
- Node 2015
- Node 2014
- Node 2017
- Node 1
- Node 1298
- Normal Search
- AnseraQuery MQL
- Downloading Application Query Test
One platform to allow rapid prototyping and evaluation
The current Search System

- A Lucene elastic search based system, documents are returned as ranked lists
Patent Gold Standards

- We have “manually” curated search reports for about 40 million simple patent families

- The relevant documents are mentioned in the search report as either

  – \( X(I,N), A, Y, \ldots \) documents

![Graph showing median citations]

median: 5 citations in search reports
Setting up a benchmarking environment

- We need to move away from anecdotal evidence to statistically meaningful facts
- TAPAS

* Exploiting real queries
Using KNIME to enhance automated search

Application

Query Formulation

Search

PriorArt

relevant documents

Evaluation and Feedback

Get Gold Standard

Translation

AnseraQuery MLT

Evaluate Search Results

German Files

Node 1298

Translated Search

TAPAS
Use Case: Does Translation help to find relevant Prior Art?
Evaluating the results – Does Translation help to find relevant Prior Art?

<table>
<thead>
<tr>
<th>%-relevant retrieved</th>
<th>Relevant Expected</th>
<th>Combined results</th>
<th>Translated Queries</th>
<th>Original Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>%-relevant retrieved</td>
<td>100</td>
<td>36,92307692</td>
<td>27,69230769</td>
<td>11,53846154</td>
</tr>
</tbody>
</table>
Patents have multi-modal information content: Images

- Images
  - Chemical Formulas
  - Flow Diagrams
  - Circuits
  - Technical Drawings
Image Search the Google way

- Standard Google image search (very strong on real world images) is currently not suited for technical drawings
- Extremely complicated endeavour
Use Case: Image Search

**Search Space**

- Table Creator
- Transpose
- Create Collection Column
- Cross Joiner
- Compare Images
- Joiner
- Inverter
- Node 34
- Node 35
- Node 49
- Node 50
- Table Cell Viewer
- Row Filter
- Node 36
- Images, Title, Abstract, ...
- Input Patent
- Predefined Patent Pool
- Rows to Columns
- Download Patent Information
- Create Collection Column of Patents
- Retrieve the most relevant patents

**Query**

- Table Creator
- Transpose
- Create Collection Column
- Cross Joiner
- Compare Images
- Inverter
- Node 34
- Node 35
- Node 49
- Node 50
- Table Cell Viewer
- Row Filter
- Node 36
- Images, Title, Abstract, ...
- Input Patent
- Predefined Patent Pool
- Rows to Columns
- Download Patent Information
- Create Collection Column of Patents
- Retrieve the most relevant patents

**Filtering and Visualisation**

- Table Cell Viewer
- Node 36
- Images, Title, Abstract, ...
- Input Patent
- Predefined Patent Pool
- Rows to Columns
- Download Patent Information
- Create Collection Column of Patents
- Retrieve the most relevant patents

**State of the art**

- Image processing

---

19
Image Search  Adaptive Hierarchical Density Histogram (AHDH)

- Exploit distribution of image points

New results

AHDH

closest match
Can we optimize our Search Tool Parameters?

Parameter Optimization Loop Start

Parameter To Optimize

Retrieving Application Information

Get Gold Standard

ANSERA Prod

ANSERA Query

Parameter Optimization Loop End

TAPAS Evaluation

TAPAS Eval

CSV Writer

save optimization parameter

Node 981

CSV Wr

Save comple

Dialog - 6:1072 - Parameter Optimization Loop Start...

File

Standard settings | Flow Variables | Memory Policy

Parameter | Start value | Stop value | Step size | Integer?

Percentage of ... | 0.1 | 1 | 0.1 | 

Occurrence | 1 | 20 | 1 | 

Terms considered | 1 | 500 | 20 | 

Add new parameter

Search strategy | Hillclimbing

Random seed | 1437146946056

OK | Apply | Cancel
Use Case: Parameter Optimisation

Swarm Optimization
Can we optimize our Search Tool Parameters?
What are the possibilities?

- We were able to create a system where all components sharing the same (table based) interface.

- We make full use of many existing components and combine them successfully with internal and external additional developments
  
  - Text mining
  - Image Similarity
  - Machine Learning based document similarity
  - Influence of translation to patent search

Rapid Prototyping
Evaluation and Feedback
What are the limits?

- Very practical issues like memory freezes of the eclipse Environment
- Sometimes extreme over head introduction e.g. text mining nodes (String to Document)
- Not suitable from our experience for full corpus analytics (100 million full text documents and more)
- Getting lost in “yellow” nodes (ungroup, pivot, regroup, filter, missing value,...)
Future

- Exposing stable services as web service through the KNIME server web-interface
- Exposing workflow creation to a wider range of user (right now its IM only)
- Connecting more services
- Using of streaming to overcome some of the limits
Thank you for your attention

aklenner@epo.org
- Supplement