Development of a text search engine for medicinal chemistry patents

Emilie Pasche, Julien Gobeill, Fatma Oezdemir-Zaech, Thérèse Vachon, Christian Lovis, Patrick Ruch

Presented by Patrick Ruch

November 14-16, 2012 NETTAB 2012, Como







Motivations

Our objective

 Development of a search engine dedicated to patent retrieval in the pharmaceutical domain

What is the interest of patent collections?

- Important source of knowledge (> 50 millions)
- Unique and validated information

What is the status of search engines for patent collections?

- Search engines for biomedical patent collections are rare.
- Evaluation campaigns (TREC) have encouraged such research.



Data

Patent collection:

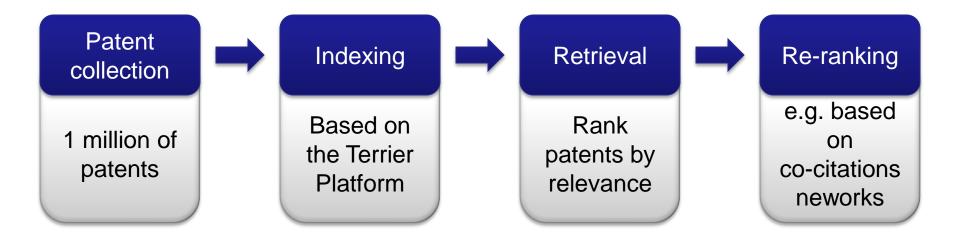
Random subset of about 1 millions of patents

Evaluation:

- Benchmark 1
 - Task: related patent search
 - Topics: 96 long queries
 - Relevance judgment: patents cited as prior-art
- Benchmark 2
 - Task: ad hoc search
 - Topics: 24 short queries
 - Relevance judgment: provided by TREC evaluators
- Benchmark 3
 - Task: know-item search
 - Topics: 514 short queries
 - Relevance judgment: the patent from which the query came



Methods



1) Impact of the description field

Aims

Use only the most content-bearing sections of the patent.

Methods

Indexing with and without the description.

Results

Description does not improve results (p<0.01)

Conclusion

Description will not be indexed in our search engine.

Settings	Benchmark 1	Benchmark 2	Benchmark3
With description	2.20%	15.87%	23.63%
Without description	2.87 (+30.0%)	19.51 (+22.9%)	33.59 (+42.2%)



2) Impact of the ontology-driven normalization of the patent content

Aims

Add metadata to patent contents.

Methods

Use of 3 terminologies: MeSH, GO and Caloha.

Results

Metadata based on the title, abstract and claims increase the results.

Conclusion

Normalization of the patent content (but not description) will be done.

Settings	Benchmark 1	Benchmark 2	Benchmark3
Metadata on title, abstract, claims and description	2.20%	15.87%	23.63%
Metadata on title, abstract and claims	3.63%	30.30%	35.02%

3) Impact of the search model

Aims

Determine the best model for patent retrieval.

Methods

Retrieval with 2 search models: PL2 and BM25.

Results

BM25 performs better than PL2.

Conclusion

BM25 will be used for retrieval.

Settings	Benchmark 1	Benchmark 2	Benchmark3
PL2	2.87%	19.51%	33.59%
BM25	5.36%	20.05%	40.86%

4) Impact of the co-citation networks

Aims

Patents that are the most cited should be favored.

Methods

Construction of a co-citation matrix to re-rank results.

Results

Co-citation networks improve results, mainly for related patent search.

Conclusion

Results will be re-ranked based on the citations.

Settings	Benchmark 1	Benchmark 2	Benchmark3
Without re-ranking	5.36%	20.05%	40.86%
With re-ranking	6.76%	21.24%	40.87%

5) Impact of the IPC classification

Aims

Evaluate if IPC codes improve quality of retrieval.

Methods

IPC codes are added to the query.

Results

Only ad hoc searches are improved.

Conclusion

An interactive IPC classifier could be used for ad hoc search.

Settings	Benchmark 1	Benchmark 2	Benchmark3
Without IPC classification	6.76%	21.24%	40.87%
With IPC classification	5.88%	23.28%	46.02%



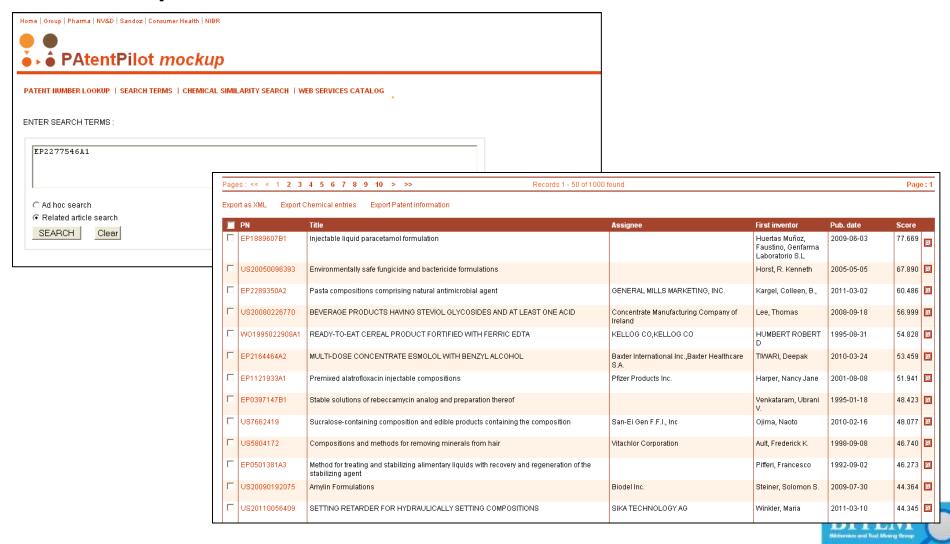
Example

Ad hoc search



Example

Related patent search



Example

Ontology-driven metadata

```
■ Metadata (automatically generated)
                     ■ 2 Terms mapped
IPC terms
                        G06F_00700 Methods or arrangements for processing data by operating upon the order or content of the data handled (logic circuits H03K0019000000) [2] View in Wipo
                        C12Q_00168 involving nucleic acids [2]
                                                                                                                                                                    View in Wipo
Inchikey
                     N/A
                     ■ 1 Term mapped
Biological process
                       GO:0010467 gene expression [9] View in Gene Ontology
Molecular function
                     N/A
Cellular component
                     N/A
Chemical from MeSH N/A
Novartis Product
                      N/A
dictionary
Tissue
                     N/A
                     ■ 1 Term mapped
Cell
                       TS-0771 Peripheral blood [6] View in NextProt
                      ■ 1 Term mapped
Anatomy
                       D001773 Blood Cells [6] View in MeSH
Disorder
                      ■ 2 Terms mapped
Gene
                       D005796 Genes [6] View in MeSH
                       D016341 Genes, vif [3] View in MeSH
                     ■ 2 Terms mapped
Procedure
                        D016133 Polymerase Chain Reaction [3] View in MeSH
                       D016002 Discriminant Analysis [3]
Device
                     N/A
Species
                      N/A
Population
                     N/A
Geography
                      N/A
                                                                             >> CWU Sequences [0 found] <<
>> Details <<
                >> Chemicals [0 found] <<
                                            >> CWU Chemicals [0 found] <<
```

Conclusion

Conclusion

- Development of a search engine dedicated to patent search
 - Based on the state of the research methods
 - Tested in a pharmaceutical industry
- Different tuning supports different use cases
 - Related patent search
 - Ad hoc search
- Future works
 - Evaluate impact of normalization by entity types

Questions?

Acknowledgements: This study has been fully supported by Novartis Pharma AG, Basel Campus, NIBR IT.

The **TWinC** prototype designed **To Win C**hemathlon can be found

here: http://casimir.hesge.ch/ChemAthlon/index.html#

