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

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
University of Applied Sciences
Western Switzerland

h e g

BITEM
Bibliomics and Text Mining Group

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

Patent collection
1 million of patents

→

Indexing
Based on the Terrier Platform

→

Retrieval
Rank patents by relevance

→

Re-ranking
e.g. based on co-citations neworks

Experiments

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

Home | Group | Pharma | NV&D | Sandoz | Consumer Health | NIBR

PAtentPilot *mockup*

PATENT NUMBER LOOKUP | SEARCH TERMS | CHEMICAL SIMILARITY SEARCH | WEB SERVICES CATALOG

ENTER SEARCH TERMS :

ccr7 sanofi

- Ad hoc search
- Related article search

SEARCH Clear

Pages : << < 1 2 3 4 5 6 7 8 9 10 > >> Records 1 - 50 of 639 found Page : 1

Export as XML Export Chemical entries Export Patent information

PN	Title	Assignee	First inventor	Pub. date	Score
WO2010120914A1	PREDICTIVE MODELS AND METHOD FOR ASSESSING AGE	CARDIODX, INC.	ROSENBERG, Steve	2010-10-21	13.261
US20020176870	Methods and compositions useful for stimulating an immune response	ChemoCentryx, Inc.	Schall, Thomas J.	2002-11-28	12.154
US20100215674	ENHANCING THE T-CELLS STIMULATORY CAPACITY OF HUMAN ANTIGEN PRESENTING CELLS AND THEIR USE IN VACCINATION	Vrije Universiteit Brussel	Thielemans, Kris Maria Magdalena	2010-08-26	11.207
WO2011015162A2	METHOD OF DETERMINATION OF DIAGNOSIS AND PROGNOSIS IN PATIENTS WITH B-CELL CHRONIC LYMPHOCYTIC LEUKEMIA AND OLIGONUCLEOTIDES FOR USE IN THIS METHOD	MASARYKOVA UNIVERZITA	BRYJA, Vitezslav	2011-02-10	10.359
WO2010136964A1	METHOD OF TREATING SLEEP DISORDERS USING THE COMBINATION OF EPLIVANSERIN AND ZOLPIDEM	SANOFI-AVENTIS	PINQUIER, Jean-Louis	2010-12-02	9.269
WO1988002220A1	METHOD FOR SEPARATING RENNET COMPONENTS	SANOFI BIO INGREDIENTS INC,SANOFI BIO INGREDIENTS INC,SANOFI BIO INGREDIENTS INC	BIRSCHBACH PETER	1988-04-07	9.076
EP2252257A2	SUBSTANCES RAISING THE ACTIVATION THRESHOLD OF IMMUNE CELL	BASF Beauty Care Solutions France SAS	BECHETOILLE, Nicolas	2010-11-24	8.881
WO2011042877A1	USE OF CELIVARONE FOR THE PREPARATION OF A MEDICAMENT FOR USE IN THE PREVENTION OF IMPLANTABLE CARDIOVERTER DEFIBRILLATOR INTERVENTIONS OR DEATH	SANOFI-AVENTIS	GAUDIN, Christophe	2011-04-14	8.756
US20070196867	GPCR expressing cell lines and antibodies	Multispan, Inc.	Mancebo, Helena S.	2007-08-23	8.712
WO1996008180A1	APPLICATOR UNIT FOR A COSMETIC PRODUCT SUCH AS MASCARA AND CORRESPONDING APPLICATOR	SANOFI SA,SANOFI SA	VANDROMME MICHEL	1996-03-21	8.639
US5001844	Method of drying carrageenans	Sanofi,ELF SANOFI	Lhonneur, Jean-Pierre	1991-03-26	8.482
EP0223765A1	Process for the preparation of hydroxy-3 ketone derivatives.	SANOFI SA,CENTRE NAT RECH SCIENT,SANOFI SA,CENTRE NAT RECH SCIENT	VESCHAMBRE HENRI	1987-05-27	8.252

Example

Related patent search

Home | Group | Pharma | NV&D | Sandoz | Consumer Health | NIBR

PAtentPilot *mockup*

PATENT NUMBER LOOKUP | SEARCH TERMS | CHEMICAL SIMILARITY SEARCH | WEB SERVICES CATALOG

ENTER SEARCH TERMS :

EP2277546A1

Ad hoc search

Related article search

SEARCH Clear

Pages : << < 1 2 3 4 5 6 7 8 9 10 > >> Records 1 - 50 of 1000 found Page : 1

Export as XML Export Chemical entries Export Patent information

PN	Title	Assignee	First inventor	Pub. date	Score
EP1889607B1	Injectable liquid paracetamol formulation		Huertas Muñoz, Faustino, Genfarma Laboratorio S.L	2009-06-03	77.669
US20050096393	Environmentally safe fungicide and bactericide formulations		Horst, R. Kenneth	2005-05-05	67.890
EP2289350A2	Pasta compositions comprising natural antimicrobial agent	GENERAL MILLS MARKETING, INC.	Kargel, Colleen, B.,	2011-03-02	60.486
US20080226770	BEVERAGE PRODUCTS HAVING STEVIOL GLYCOSIDES AND AT LEAST ONE ACID	Concentrate Manufacturing Company of Ireland	Lee, Thomas	2008-09-18	56.999
WO1995022908A1	READY-TO-EAT CEREAL PRODUCT FORTIFIED WITH FERRIC EDTA	KELLOG CO,KELLOG CO	HUMBERT ROBERT D	1995-08-31	54.828
EP2164464A2	MULTI-DOSE CONCENTRATE ESMOLOL WITH BENZYL ALCOHOL	Baxter International Inc.,Baxter Healthcare S.A.	TIWARI, Deepak	2010-03-24	53.459
EP1121933A1	Premixed alatrofloxacin injectable compositions	Pfizer Products Inc.	Harper, Nancy Jane	2001-08-08	51.941
EP0397147B1	Stable solutions of rebeccamycin analog and preparation thereof		Venkataram, Ubrani V.	1995-01-18	48.423
US7662419	Sucralose-containing composition and edible products containing the composition	San-Ei Gen F.F.I., Inc	Ojima, Naoto	2010-02-16	48.077
US5804172	Compositions and methods for removing minerals from hair	Vitachlor Corporation	Ault, Frederick K.	1998-09-08	46.740
EP0501381A3	Method for treating and stabilizing alimentary liquids with recovery and regeneration of the stabilizing agent		Pifferi, Francesco	1992-09-02	46.273
US20090192075	Amylin Formulations	Biodel Inc.	Steiner, Solomon S.	2009-07-30	44.364
US20110056409	SETTING RETARDER FOR HYDRAULICALLY SETTING COMPOSITIONS	SIKA TECHNOLOGY AG	Winkler, Maria	2011-03-10	44.345

Example

Ontology-driven metadata

Metadata (automatically generated)

IPC terms	2 Terms mapped 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
Biological process	1 Term mapped GO:0010467 gene expression [9] View in Gene Ontology
Molecular function	N/A
Cellular component	N/A
Chemical from MeSH	N/A
Novartis Product dictionary	N/A
Tissue	N/A
Cell	1 Term mapped TS-0771 Peripheral blood [6] View in NextProt
Anatomy	1 Term mapped D001773 Blood Cells [6] View in MeSH
Disorder	N/A
Gene	2 Terms mapped D005796 Genes [6] View in MeSH D016341 Genes, vif [3] View in MeSH
Procedure	2 Terms mapped D016133 Polymerase Chain Reaction [3] View in MeSH D016002 Discriminant Analysis [3] View in MeSH
Device	N/A
Species	N/A
Population	N/A
Geography	N/A

>> Details << >> Chemicals [0 found] << >> CWU Chemicals [0 found] << >> CWU Sequences [0 found] <<

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 **T**o **Win C**hemathlon can be found here: http://casimir.hesge.ch/ChemAthlon/index.html#