Text Mining Patents at Scale

David Milward

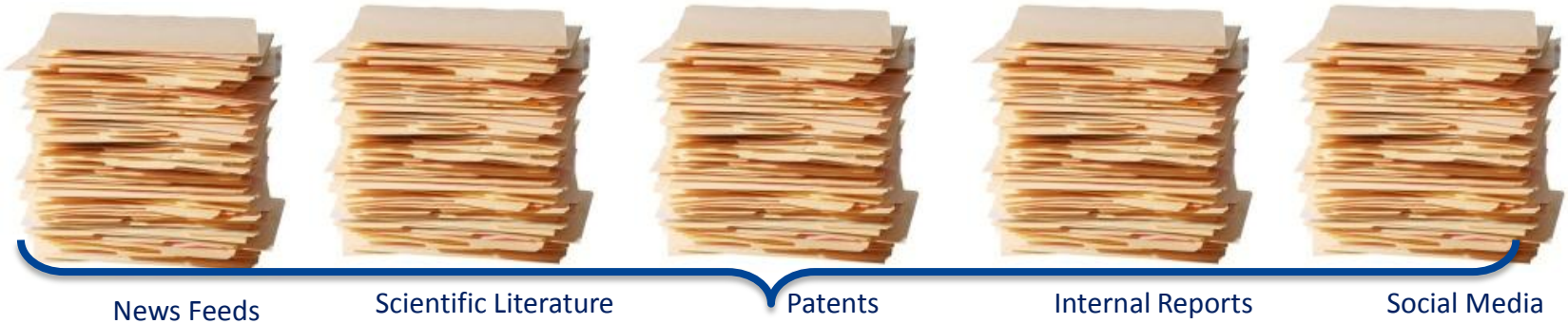
Linguamatics

ICIC 2012

Overview

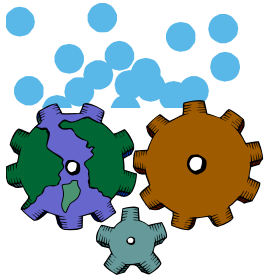
- Introduction to Agile Text Mining
- Applications for Patent Search and Mining
- Strategies to find the most relevant documents
- Accelerating search strategy development
- Challenges
  - Linking across patent documents
  - Chemical naming
  - Multilingual processing

Search vs. Text Mining



Search Engine

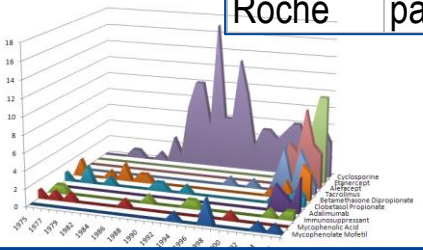
Filter to find most relevant documents, then read



Text Mining

Manipulate the text to discover what is there

company	activity	company
Sanofi	bid	Aventis
Roche	partner	Antisoma



Natural Language Processing (NLP) to understand meaning

Statistics to provide trends

Linguamatics

Agile Text Mining

- *Text mining* provides ability to **discover**
  - but typically queries have to be programmed in, and processing is slow
- *Search* provides ability to **filter** quickly to relevant documents
  - but poor at answering open questions e.g. “what are biomarkers for breast cancer”
- Combine *text mining* with *search* to **discover** within **specific contexts** e.g.

Discover what is available
Smoking
alcohol intake
Diet
Excess adiposity
Heavy alcohol consumption
heavy alcohol intake
lower birth weight

Filter to the context of interest
is a risk factor for type 2 diabetes
is a potential risk factor for incident diabetes
is a potential risk factor for newly diagnosed type 2 diabetes mellitus
is a potentially modifiable risk factor for diabetes
is a risk factor for diabetes
is a risk factor for diabetes mellitus
is an established risk factor for diabetes

Wide Variety of Strategies to Find Information

Terminologies

- Search for e.g. cancer and get synonyms ***and*** children:
  - Malignant neoplasms, Malignant tumor …
  - Leukaemia, Lymphoma, Astrocytoma …

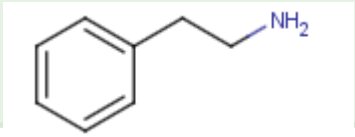
NLP

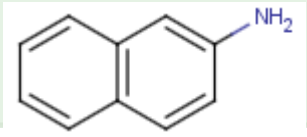
- Precise linguistic relationships, sentence co-occurrence
- Precise negation e.g. “pressure” but not “blood pressure”

Regular Expressions

- e.g. microRNA: let-?\d+.* mirn?a?-?\d+.*

Chemical Substructure



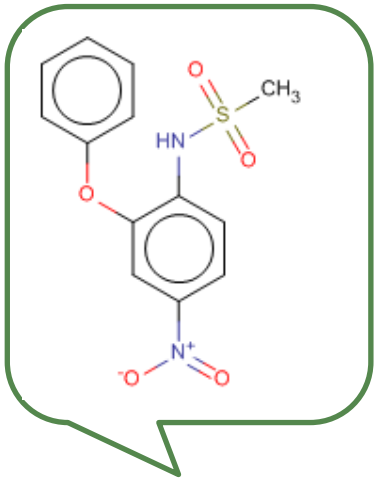


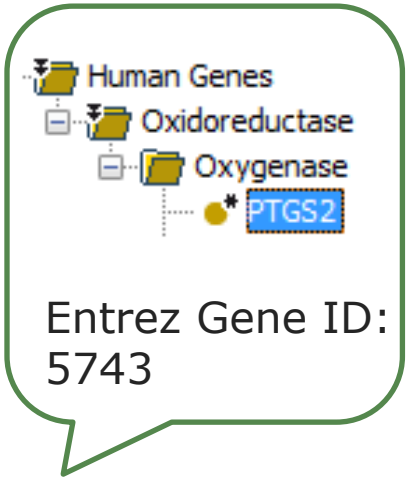
High Throughput

- Simultaneous processing of large numbers of items e.g.
  - 500 genes from microarray experiment

From Words to Meaning

Identifying entities and relations

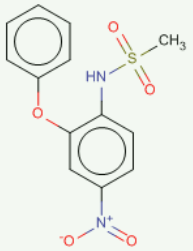




"Among them, nimesulide, a selective COX2 inhibitor, …"

inhibits

Linguistics to establish relationships

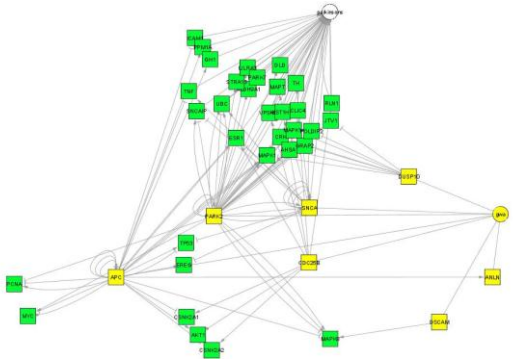


inhibits → Entrez Gene ID: 5743

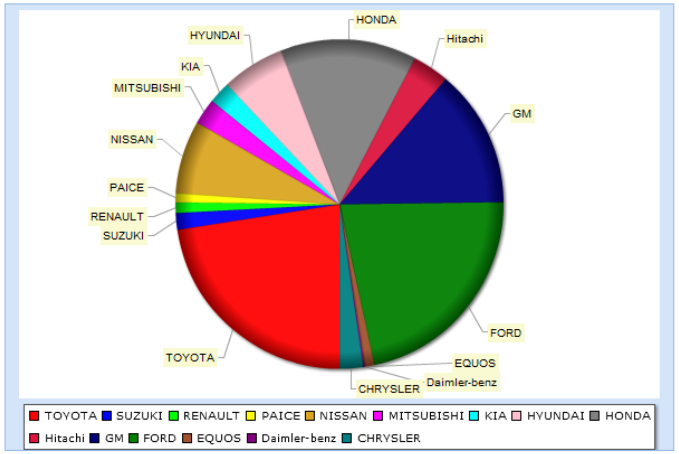
Linguamatics

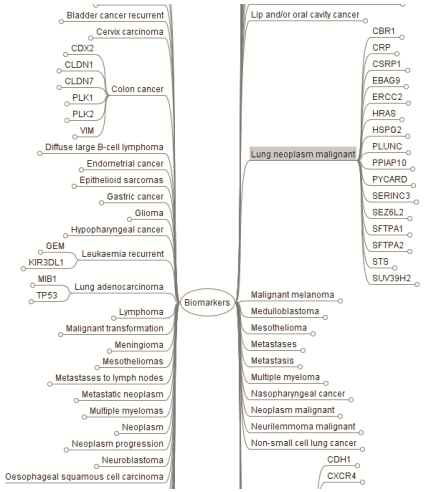
Presenting Results for Efficient Review

- Focus on documents with the necessary information
- Don’t need to read every document supporting commonly known information



Entrez Genes		Doc
ERBB2	470	15870086
BRCA1	248	15564800
BRCA2	160	15986445
PGR	129	15272277
EGFR	101	16280056
VEGFA	76	15897560
TP53	73	15583825
INS	68	15805581

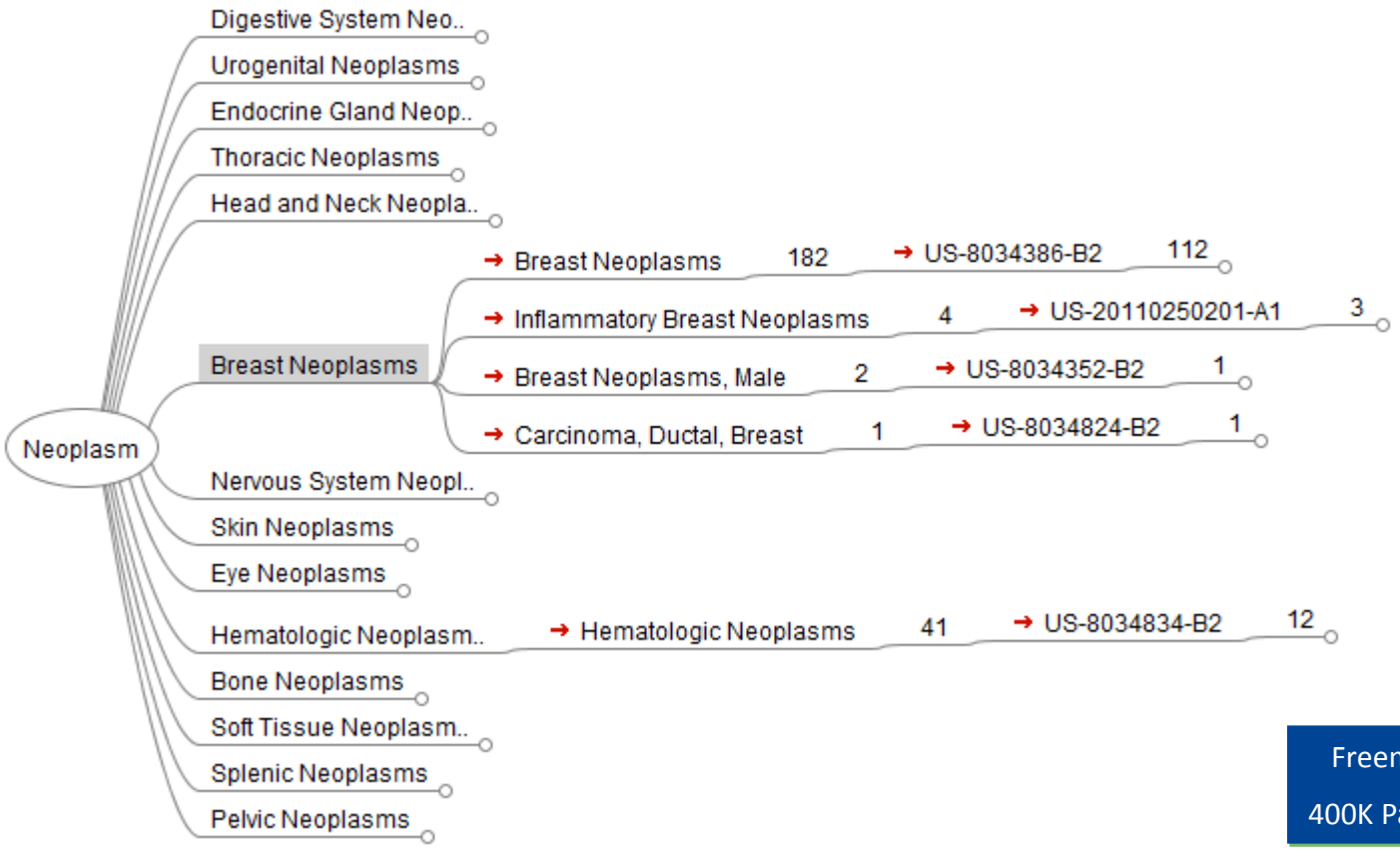




Linguamatics

Disease Areas with the most Activity

- What cancers are receiving the most patent activity this year?

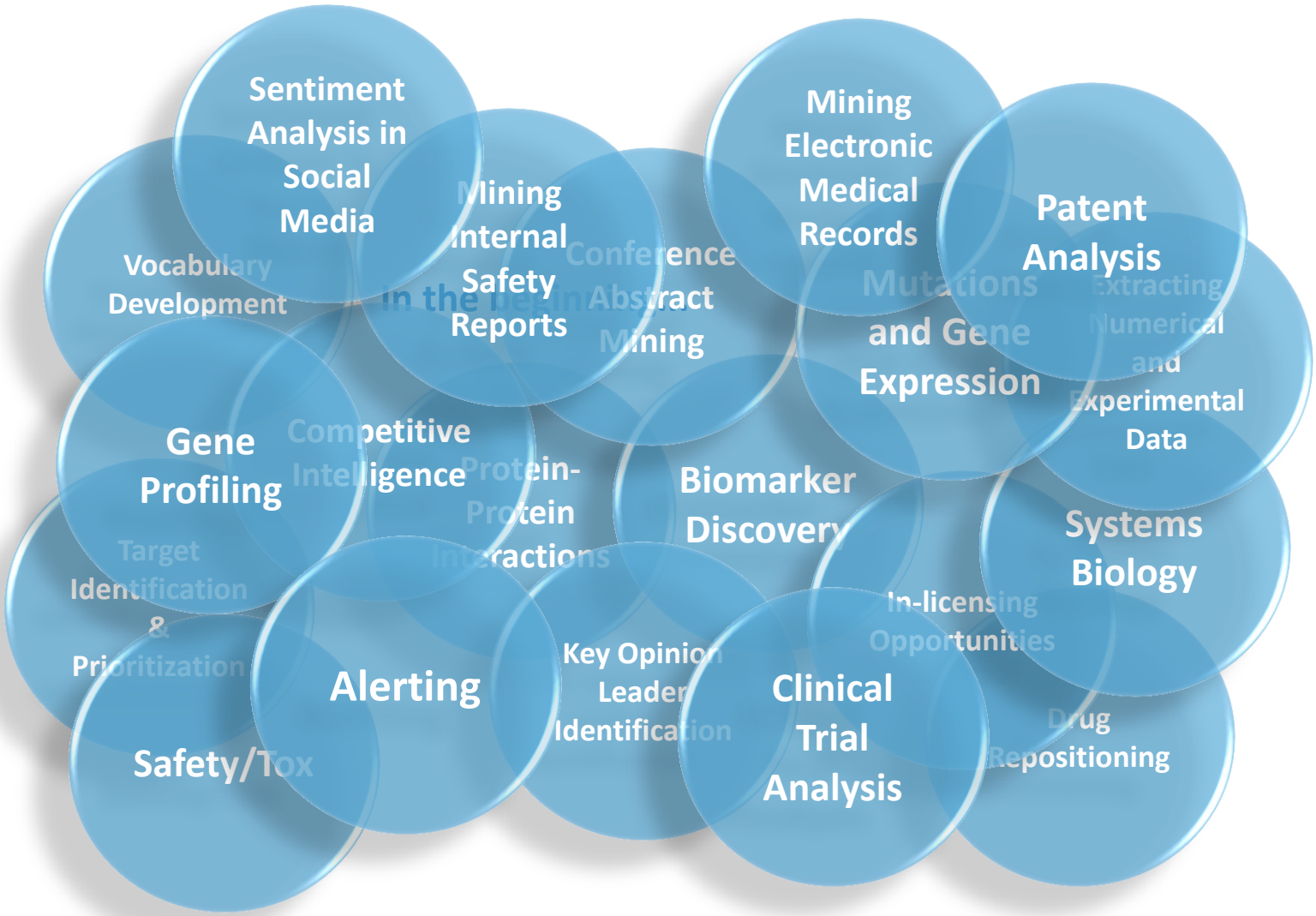


Freemind

400K Patents

Linguamatics

Applications of I2E Agile Text Mining



Linguamatics

Example Applications for Patents

- New strategies for novelty/validity/freedom to operate searches
- Analysis of sets of patents e.g. for Competitive Intelligence
- Knowledge discovery e.g. scientific or engineering information reported in patents
- Linking knowledge silos e.g. between patents and scientific literature, internal documents or structured data
- Integration with workflow tools e.g. Pipeline Pilot for:
  - Automated analysis e.g. specialised email/RSS alerting
  - Dashboards e.g. monthly trend analysis
- Automatic terminology discovery

Agile Text Mining of Patent Data

- I2E has been used for several years to mine selected patent content extracted from databases such as:
  - MicroPatent
  - PatBase
  - IBM
- Customers do an initial search of their database, ensure this provides good recall, and then do further filtering via text mining
- However, text mining provides different filters, so there are advantages in mining the full patent content

Just Launched: I2E Patent Solution

I2E text mining

- High performance NLP querying
- Fact extraction, knowledge synthesis
- Use of domain thesauri
- Ability to cross document silos

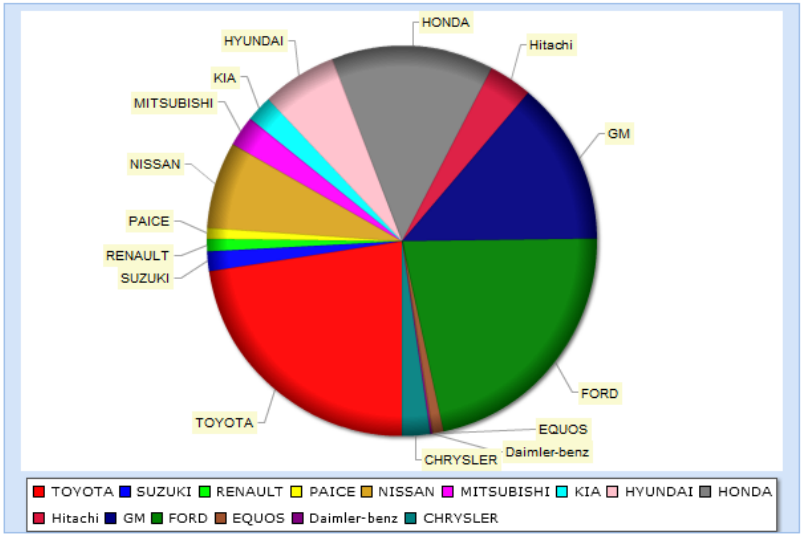
Hosted platform

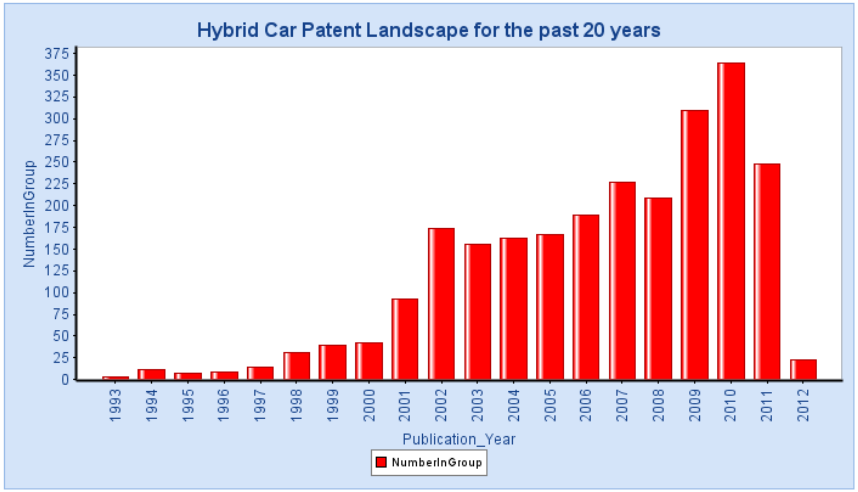
- Reduced Cost of Ownership
- No installation delay

Full text patents

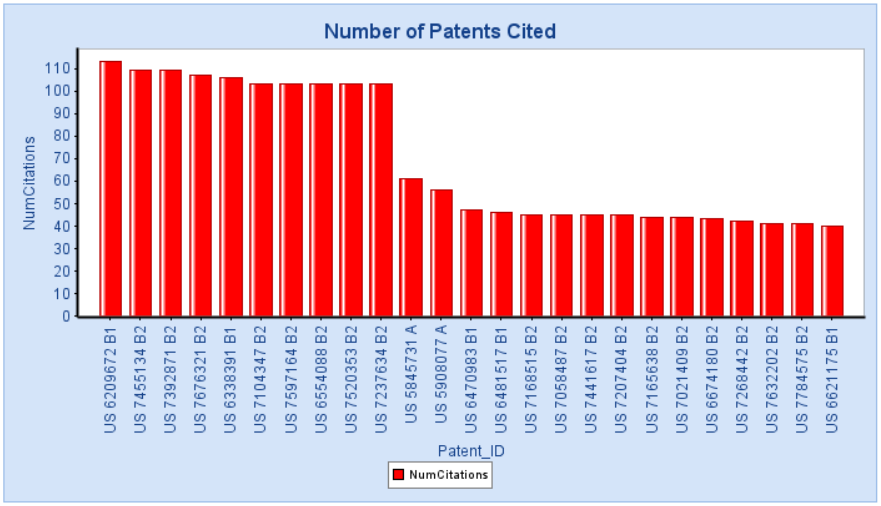
- 17 million documents (USPTO, EPO, WIPO)
- Applications and Grants
- From IFI Claims Direct

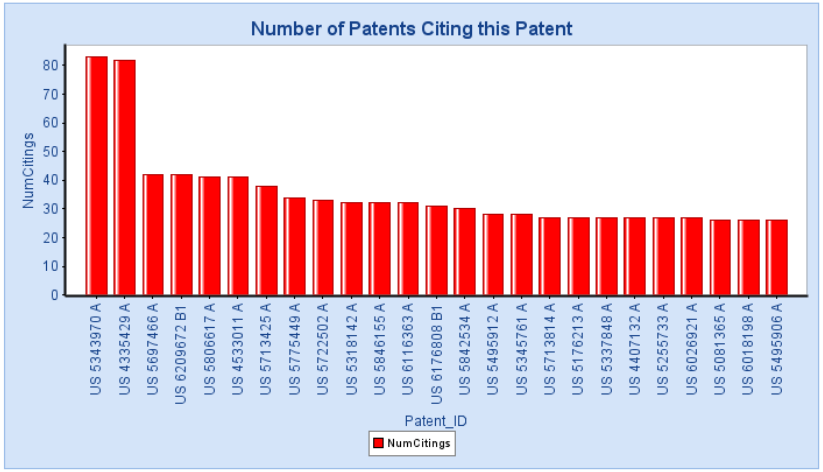
Hybrid Car Patent Landscape for the Last 20 Years





2478 patents from USPTO





Linguamatics

Infringement Search for Controllers in Hybrid Cars

Company	Document Id	Claim#	Controller relation
PAICE	US-5343970-A	11.	a controllable torque transfer unit adapted to receive torque
			a controllable torque transfer unit adapted to receive torque from two sources

Quickly compare claimed subject matter for all in-force patents

Company	Document Id	Claim#	Controller relation	Electric Engine	Combustion Engine	Propulsion
TOYOTA	US-7832512-B2	1.	a control device for driving the drive device	the second electric motor	the internal combustion engine	the output shaft
NISSAN	US-6026921-A	1	a control unit being configured to be electrically connected to said propulsion mechanism	the first electric motor	an internal combustion engine	an output shaft

Hit
... the hybrid vehicle travels using the internal combustion engine as a main drive power ... by the electricity generated from the second electric motor; a drive device for ... device for being coupled to the output shaft of the internal, combustion ... by a driver; and a control device for driving the drive device to rotate the output shaft ...

Example Highlighted Document

CONTROL APPARATUS FOR HYBRID VEHICLE

Abstract (EN)

In a hybrid vehicle having a locking mechanism in which a play elimination process is required in the locking, a torque shock in the play elimination is reduced. In a hybrid vehicle **1** having a locking mechanism **700** which is a cam-lock type engaging apparatus, an ECU **100** performs MG**1** locking control. In the control, play is formed between a cam **710** and a clutch plate **720** of the locking mechanism **700**. The formed play is gradually reduced such that the torque shock in the play elimination does not occur due to the phase control of the cam **710**, on the basis of an initial value of the amount of the play when the clutch plate **720** is bought into contact with a friction part **733** and a play elimination amount G.

Inventors: Ebuchi, Hiroaki; Ideshio, Yukihiko; Kimura, Hiromichi; Kitabatake, Hirotatsu; Ono, Tomohito (Format: Original)
Assignees: TOYOTA JIDOSHA KABUSHIKI KAISHA (Toyota-shi Aichi Country: JP) (Format: Original)
Document Id: US-20120028757-A1 (Country: US Document Number: 20120028757 Document Kind: A1 Document Language: EN)
Patent Family Id (DOCDB): -1 (Unassigned)
Publication Date: 20120202 (Publication Year: 2012 Publication Month: 02 Publication Day: 02)

International Convention Data
PCT or Regional Filing Data:
- Country: JP Doc Number: 0957697 Date: 20090416

Classifications

IPCR
- B60W 10/02 20060101ALI20120202BHUS
- B60W 20/00 20060101AFI20120202BHUS
- B60W 10/08 20060101ALI20120202BHUS
- B60W 10/06 20060101ALI20120202BHUS

US National
- 477005
- 180065265
- 18006522
- 903930

Publication Reference: Country: US Doc Number: 20120028757 Kind: A1 Date: 20120202 Language: EN
Application Reference: Country: US Doc Number: 26401109 Kind: A Date: 20090416 Language: EN

Claims (EN)

[00001] **1**. A control apparatus for a hybrid vehicle, the hybrid vehicle comprising:power supplying elements including at least a first electric motor, a second electric motor, and an internal combustion engine;a power transmission mechanism comprising a plurality of rotational elements which can mutually perform differential rotation and which includes a first rotational element coupled with the first electric motor, a second rotational element coupled with the second electric motor, and a third rotational element coupled with the internal combustion engine, the power transmission mechanism performing power transmission between a drive shaft coupled with an axle and the power supplying elements in accordance with a power transmission mode determined in accordance with a state of each of the plurality of rotational

Linguamatics

Novel Filters to Get to Relevant Documents

- Different filters and combinations of filters will home in on different sets of documents
- Agile text mining provides a wider range of filters:
  - Precise negation e.g. pressure but not preceded by blood
  - Regular Expressions
  - Terminologies
  - Linguistic constraints
  - High-throughput
  - Chemical substructure
  - Restrictions to a particular field e.g. Claims
  - Nested fields e.g. within the same Claim within Claims, or Table within the Background

Accelerating a Search Strategy

- Search is good if you know what you are looking for. However, this often needs to be an iterative process
- Text mining allows you to discover as well as search e.g.
  - The kinds of chocolate mentioned in patents
  - The most common IPC codes assigned to patents by companies of interest
- This means you iterate much faster to an improved, more complete search strategy

IPCR	Group	Subgroup	Organization		Doc
A61K	31	506	Abbott	5	US-20120003291-A1
C07D	403	12	Abbott	3	US-8101754-B2
A61P	25	00	Abbott	6	US-20120022103-A1
A61M	25	00	Abbott	4	US-8100856-B2
G01N	33	53	Abbott	3	US-20120020972-A1
C12N	5	16	Abbott	3	US-20120020972-A1
A61F	2	06	Abbott	5	US-8088157-B2
C07K	14	00	Abbott	2	US-8101565-B2

dark chocolate	20
milk chocolate	17
chocolate products	14
hot chocolate	14
chocolate bars	13
chocolate milk	13
chocolate bar	12
white chocolate	12
chocolate liquor	9
chocolate product	9
chocolate production	8
chocolate mass	7
chocolate flavoring	7
chocolate flavor	7
chocolate chips	6
molten chocolate	6

Terminology Development

- Agile text mining provides a way of discovering actual language use e.g.
- What terms are synonyms
- What relationships do you get between entities of interest
- What terms are in the same class

word		Doc
effective	51	19744410
associated	51	20959578
treat	45	19821411
reduced	21	20419513
used	19	21044451
improves	18	12717587
prevent	16	20234184
reduces	16	20851204
increased	15	20838997
inhibited	14	19825949
induce	13	19734442
inhibit	12	20819433

Synonym		Doc
CsA	10	21095450
ciclosporin	2	20158284
CSA	2	11669169
Cys	2	17008059
CYA	1	7822939
CyA	1	21169912
CYC	1	20130383
Neoral	1	11260540
Sandimmun	1	2190313
Sandimmune	1	2096278

Challenge 1: Linking across a Patent Document

- Information required is often distributed across the patent
- The title compound may be defined at the top of the patent
- Exemplified compound properties may appear in a later table
- Interpreting a claim may depend on understanding claims it is dependent upon

Linking Exemplified Compounds to their Properties

- Linking information in one part of a patent to another e.g.
  - Finding examples where there is a value reported

EXAMPLE 12

2-(R)-2-(2-Amino-2-methylpropionylamino)-3-(2,4-difluorophenyl)methoxy propionic acid N-[5-(4-chlorophenyl)-3,3-dimethyl-1,1-dioxo-2,3-dihydroisothiazol-4-ylmethyl]-N-ethylamide Hydrochloride (Z=2,4-F_2)

Example 4	0.9
Example 5	0.6
Example 7	1.7
Example 12	8.0
Example 15	2.1
Example 16	1.1

Chemical	Inhibition Metric	Value	Units		Doc
(2R)-2-(2-amino-2-methylpropanamido)-N-{[5-(4-chlorophenyl)-3,3-dimethyl-1,1-dioxo-2H-1λ⁶,2-thiazol-4-yl]methyl}-3-[(2,4-difluorophenyl)methoxy]-N-ethylpropanamide hydrochloride	EC50	8.0	nM	2	US-7396846-B2

Patent Data from IFI Claims Direct

Linguamatics

Claim Chain Information

- For information in claims, often want to work back along the chain of claims, to see what the current claim is dependent upon

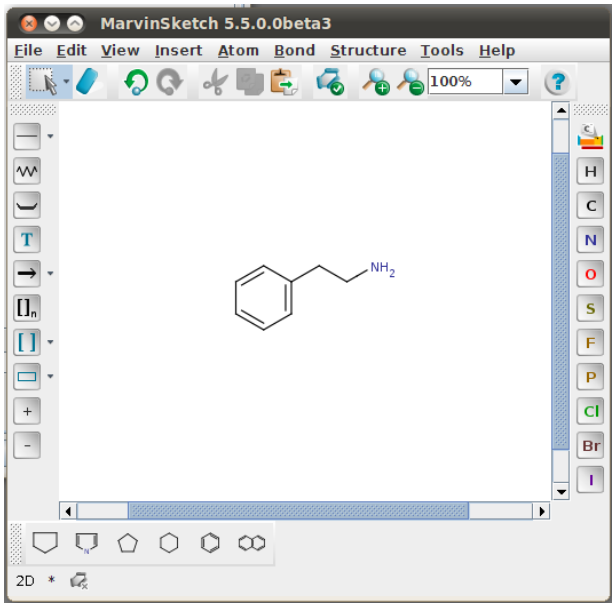
14. The method of claim 12, wherein the chemotherapeutic agent is selected from the group consisting of bleomycin, chlorambucil, epirubicin, 5-fluorouracil, ifosfamide, mitomycin, methotrexate, vincristine, cisplatin and vinblastine.	12. The method of claim 10, wherein the cervical cancer cells are treated with radiation in conjunction with the polypeptide.	10. The method of claim 5, wherein the polypeptide comprises amino acid residues 1-176 of SEQ ID NO:2.	5. A method for inhibiting the growth and or proliferation of cervical cancer cells comprising bringing a polypeptide comprising amino acid residues 33-176 of SEQ ID NO:2 into contact with the cervical cancer cells.
7. The method of claim 5, wherein the cervical cancer cells are treated with one or more chemotherapeutic agents in conjunction with the polypeptide.	5. A method for inhibiting the growth and or proliferation of cervical cancer cells comprising bringing a polypeptide comprising amino acid residues 33-176 of SEQ ID NO:2 into contact with the cervical cancer cells.		

Challenge 2: Chemistry

- Chemicals in patents cannot be found purely using dictionary matching
- Many of the more interesting chemicals are being described for the first time using systematic names e.g.
  - 4-hydrazino-5-methyl-1H-pyridin-2-one
- There may also be defined within supplementary MOL files
- Sometimes all we have is an image of a structure

Integration of ChemAxon Name to Structure with I2E

- Chemicals found by name to structure as well as by dictionary matching
- Terminology created on the fly, with different matches brought together as a single chemical concept via an ID, either
  - SMILES
  - InChi
- Can search via SMILES/SMARTS
- Can output SMILES/InChi or preferred names in results
- Preferred names use either:
  - the common name if available
  - systematic name via structure to name



ChemAxon

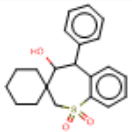
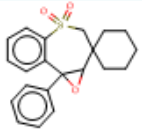
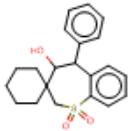
Linguamatics

Melting Points for Exemplified Compounds

Find properties

Chemical	Feature	Value	Doc		Hit
4-hydroxy-5-phenyl-4,5-dihydro-2H-spiro[1λ⁶-benzothiepine-3,1'-cyclohexane]-1,1-dione	mp	209-210°	US-6943189-B2	1	Example 31 cis-4-Hydroxy-5-phenyl-2,3,4,5-tetrahydro spiro(benzothiepine-3,1′-cyclohexane)-1,1-dioxide (60) ... mg of white crystal, mp 209-210° C. Proton and carbon NMR ...
2'-phenyl-3'-oxa-7'λ⁶-thiaspiro [cyclohexane-1,5'-tricyclo [6.4.0.0²,⁴]dodecane]-1'(8'),9',11'-triene-7',7'-dione	mp	154-155°	US-6642268-B2	1	8b-Phenyl-1a,2,3,8b-tetrahydrospiro (benzothiepino[4,5-b]oxirene-2,1′-cyclohexane)-4,4-dioxide (58) ...) of yellow solid, mp 154-155° C. Proton and carbon NMR ...
4-hydroxy-5-phenyl-4,5-dihydro-2H-spiro[1λ⁶-benzothiepine-3,1'-cyclohexane]-1,1-dione	mp	99-100°	US-6642268-B2	1	trans-4-Hydroxy-5-phenyl-2,3,4,5-tetrahydro spiro(benzothiepine-3,1′-cyclohexane)-1,1-dioxide (59) ... as a white solid, mp 99-100° C. Proton NMR showed this ...

Output to e.g. Excel

Chemical	Feature	Value	Hit
	mp	209-210°	Example 31 cis-4-Hydroxy-5-phenyl-2,3,4,5-tetrahydro spiro(benzothiepine-3,1′-cyclohexane)-1,1-dioxide (60) ... mg of white crystal, mp 209-210° C. Proton and carbon NMR ...
	mp	154-155°	8b-Phenyl-1a,2,3,8b-tetrahydrospiro(benzothiepino[4,5-b]oxirene-2,1′-cyclohexane)-4,4-dioxide (58) ...) of yellow solid, mp 154-155° C. Proton and carbon NMR ...
	mp	99-100°	trans-4-Hydroxy-5-phenyl-2,3,4,5-tetrahydro spiro(benzothiepine-3,1′-cyclohexane)-1,1-dioxide (59) ... as a white solid, mp 99-100° C. Proton NMR showed this ...

Patent Data from IFI Claims Direct

Linguamatics

Generating an Evaluation Corpus using Linguistics

- Use high precision linguistic patterns to pull out likely systematic names e.g.

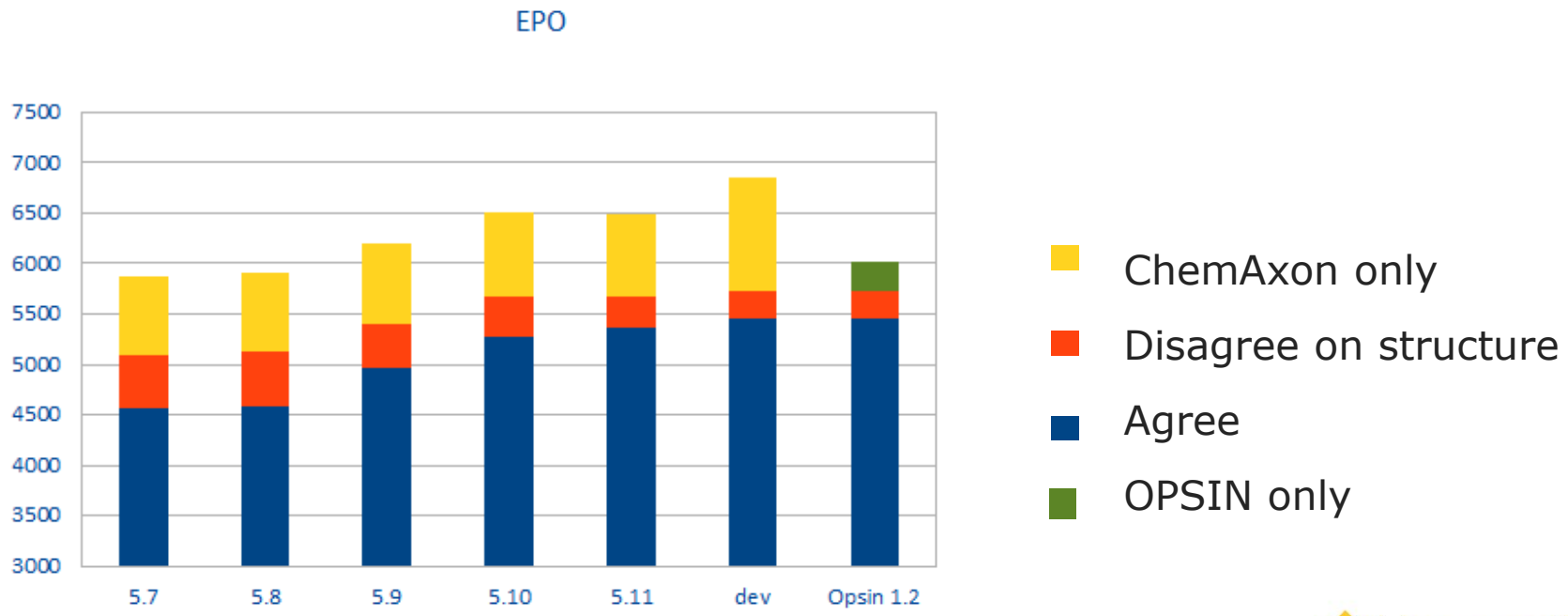
Example 4 3-Butyl-3-ethyl-5-phenyl-2,3-dihydrobenzothiepine-1,1-dioxide (7)

- Extracted 70K names from 100K patents. Sampling suggests >90% precision

2-[(3-triethoxysilylpropyl)aminocarbonyl]-benzoic acid-(6-methylene-1,4-dithiepane-2-yl)methylester
2-Amino-3-(4-hydroxy-phenyl)-propionic acid methyl ester
3-Butyl-3-ethyl-5-phenyl-2,3-dihydrobenzothiepine-1,1-dioxide
3-Butyl-3-ethyl-5-phenyl-2,3-dihydzobenzothiepine-1,1-dioxide
5-(2-(4-(4-Fluoro-2-methyl-1H-indol-5-yloxy)-6-methoxyquinolin-7-yloxy)ethyl)-5-azaspiro[2.4]-heptan-7-one
5-Fluoro-3-methyl-benzo[b]thiophene-2-sulfonic acid(4-bromo-2-trifluoromethyl-phenyl)-amide
(4S,5S)-2-(3-bromophenyl)-3,4-dimethyl-5-phenyl-1,3-oxazolidine

Evaluation of Name to Structure

- Evaluated past systems on the extracted set of compounds
  - confirmed that good progress had been made
- Set of extracted compounds has been used to drive recent improvements
- Dataset extracted by I2E from EPO data is challenging for name to structure, in particular due to OCR errors
  - sampling suggests currently converting 86% of systematic chemical names into structures



Linguamatics

Including MOL Files

- Many modern patents are associated with supplementary files
- Can process MOL files, and generate SMILES and IUPAC names

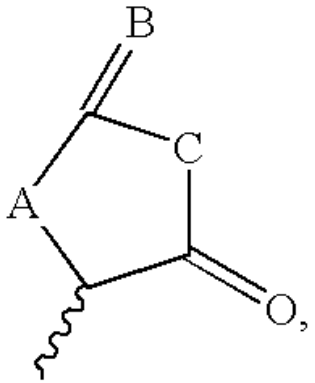
MOL file ID: **US07314881-20080101-C00103**
SMILES from MOL: CCC(CC)n1c2nc(NCCO)nc(-c3ccccc3Cl)c2ccc1=O
Name from MOL: 4-(2-chlorophenyl)-2-[(2-hydroxyethyl)amino]-8-(pentan-3-yl)pyrido[2,3-d]pyrimidin-7-one]

- Retain the correct position within the document
  - Can ask for properties in proximity to a structure defined by a MOL file e.g. information in the same table row

CCOC(=O)C(\C#N)=C1/CC(C)(C)c2cc(OC)ccc12	0.09%
CC[Si](CC)(CC)\C=C\CCOC(=O)C(\C#N)=C1/CC(C)(C)c2cc(OC)c(OC)cc12	20%
CC[Si](CC)(CC)CCOCCCCOC(=O)C(\C#N)=C1/CC(C)(C)c2cc(OC)c(OC)cc12	20%
CCOC(=O)C(\C#N)=C1/CC(C)(C)c2cc(OC)c(OC)cc12	0.35%
COc1cc2\C(CC(C)(C)c2cc1OC)=C(/C#N)C(=O)NCC#C	0.04%
COc1cc2\C(CC(C)(C)c2cc1OC)=C(/C#N)C(=O)OCC(C)C[Si](C)(C)O[Si](C)(C)C	20%
COc1cc2\C(CC(C)(C)c2cc1OC)=C(/C#N)C(=O)OCCC[Si](C)(C)O[Si](C)(C)C	20%

Challenges in Processing MOL files

- Structures containing generic groups defined elsewhere in the text
- Elided parts of structures
- Parts of reaction mechanisms
- "Using a chemical editor as a drawing package"
- "Using a chemical editor to draw something not even chemical"



- Can successfully remove unusual structures to gain high precision

Challenge 3: Multilingual Processing

- Linguistic differences
  - Word and sentence breaks are different in e.g. Asian languages
  - Linguistic entities require different rules/training
- Noise from existing terminologies
  - Need to avoid adding noise in results from English synonyms hitting inappropriately in non-English text
- Coverage of terminologies outside English
  - Most terminologies are either English only, or have very sparse coverage of other languages relative to English
  - I2E can be used to expand terminologies from a monolingual corpus
  - We can also exploit corpora where there are parallel documents in different languages

Multilingual Terminology Acquisition

m<an>tra

- If we are to directly text mine other languages it would be useful to have more comprehensive multilingual terminologies
- Mantra is an EU FP7 project looking at how to automatically adapt biomedical terminologies from one language to another using parallel multilingual corpora

Universität Zürich[UZH]

Linguamatics

averbis

Erasmus MC

EMBL-EBI

seit 1558
Friedrich-Schiller-Universität Jena

Conclusions

- Agile Text Mining provides a wider range of filters for patent searching
  - new strategies for finding the most relevant documents
  - a systematic approach for developing search strategies
- Ability to slice and dice the data to summarize, show trends e.g. for competitive intelligence, and to categorize
- Extraction of information for new knowledge
  - extract precise information from within documents
  - link information within one document to another via use of terminologies and structured output
  - link across knowledge silos e.g. patents and scientific literature