SUMOBRAIN
Big, powerful, smart searching

Making Searching Faster and More Complete: Cross-Collection Search and Automated Result Set Analysis

Scientific and Patent Information FIND

ICIC 2011
James Ryley, Ph.D.
james.ryley@sumobrain.com

Federated Search: Is Bigger Better?

Our data includes:

- Patents (and meta data)
- Biomedical literature
- SEC (EDGAR) filings for public companies
- Theses and Dissertations
- Additional full-text articles via large STM publishers

freepatentsonline
all the inventions of mankind

openthesis
Upload and Search Theses & Dissertations

SUMOBRAIN
Big, powerful, smart searching

BioMedSearch.com

GOEDGAR

Combining Biomedical & Patent Docs: A Model For Technical Search

Both large, complex, document sets that are important economically and for research.

- Patents: 27 million docs
  - Includes US, JP, EP, WIPO
  - Biomedical Patents: 3.4 million

freepatentsonline
all the inventions of mankind

- Biomedical: 21 million docs
  - Includes PubMed & other docs

BioMedSearch.com

Assessing the Utility of the Combined Data

Disparate test queries used to determine:

- % contribution of each data set (recall)
- % relevant documents (precision, top 100 only)
- Total relevant contribution from each collection

Other Considerations:

- Duplicates [1]
- Ease/speed of assessment [2]
- Specificity of data found
- Full-text versus abstracts

% Query Responses From Each Collection

Query	Total Docs	Patent	% of total	Non-Patent	% of total
proteasome AND degradation AND ubiquitin AND targeting	13533	9182	67.85%	4351	32.15%
aging AND "oxidative damage" AND lipofuscin	428	305	71.26%	123	28.74%
cancer AND apoptosis AND spindle AND mitosis	4829	4209	87.16%	620	12.84%
antibiotic AND resistant AND "drug design" AND mrsa	600	547	91.17%	53	8.83%
aav AND vector AND "gene therapy" AND dystrophy	3940	3788	96.14%	152	3.86%
lumen AND stent AND coating	35433	35130	99.14%	303	0.86%

% Relevant Docs From Each Collection

Query	% Patent Relevant	% Non-Patent Relevant
proteasome AND degradation AND ubiquitin AND targeting	38%	37%
aging AND "oxidative damage" AND lipofuscin	10%	60%
cancer AND apoptosis AND spindle AND mitosis	46%	84%
antibiotic AND resistant AND "drug design" AND mrsa	50%	65%
aav AND vector AND "gene therapy" AND dystrophy	18%	40%
lumen AND stent AND coating	19%	46%

SUMOBRAIN
Big, powerful, smart searching

Relative Contribution of Relevant Docs

Query	% of Total Relevant Docs from Patents*	% of Total Relevant Docs from Non-Patent*
proteasome AND degradation AND ubiquitin AND targeting	68.43%	31.57%
aging AND "oxidative damage" AND lipofuscin	29.24%	70.76%
cancer AND apoptosis AND spindle AND mitosis	78.80%	21.20%
antibiotic AND resistant AND "drug design" AND mrsa	88.81%	11.19%
aav AND vector AND "gene therapy" AND dystrophy	91.81%	8.19%
lumen AND stent AND coating	97.95%	2.05%

*Extrapolation of top 100 data – should be treated as a rough estimate only

Other Factors in Collection Utility

- Patents tend to have multiple near-duplicate documents due to family members (same patent filed in multiple countries) [1]
- Non-patent biomedical literature tends to have longer, more descriptive titles [2]
- Patents tend to discuss possibly-prophetic examples – the opposite of peer-reviewed literature
- The use of language is quite different between the two collections
- Non-patent documents are harder to index as full-text due to copyright issues

Federated Database Conclusions

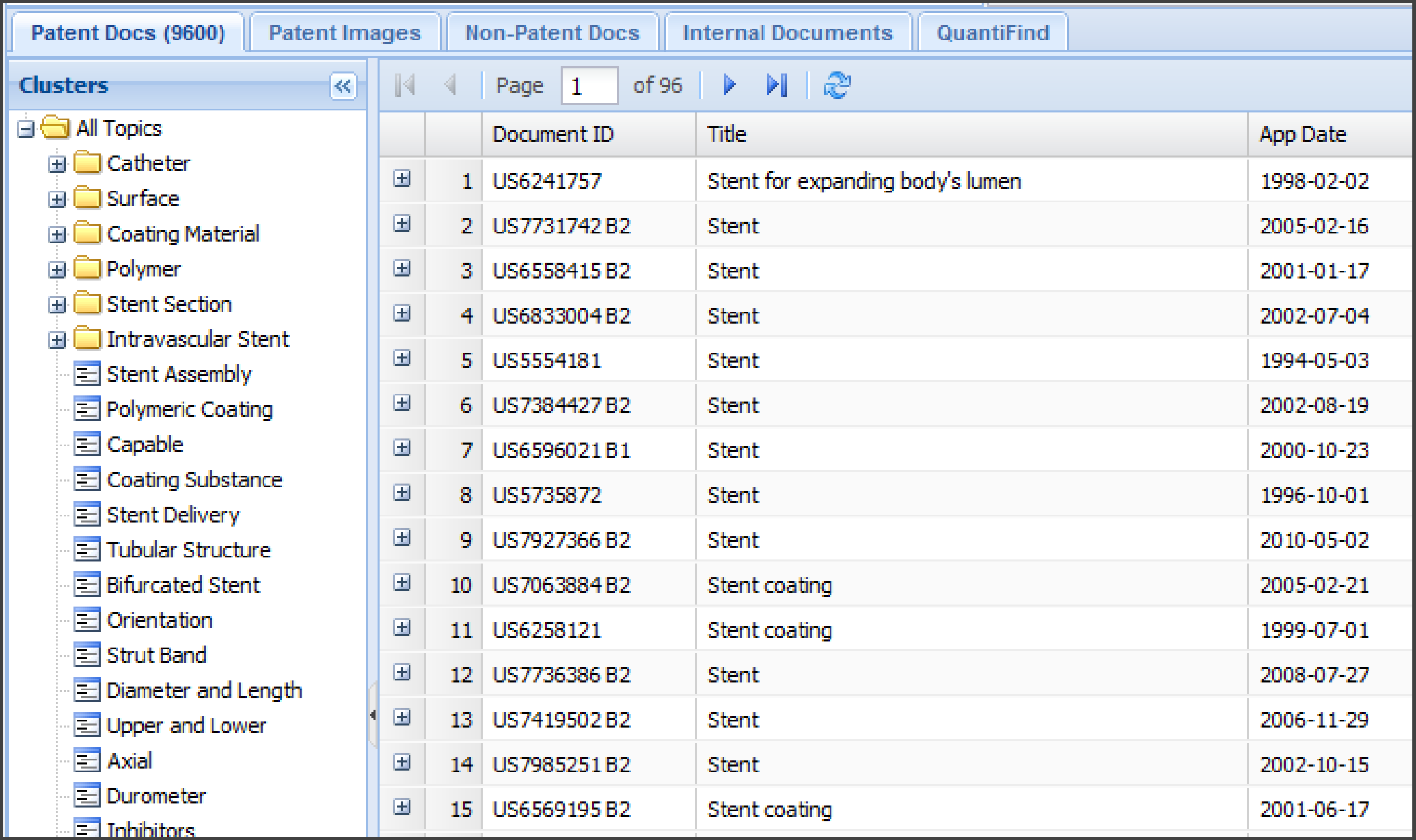
- Bigger is better. In most cases a substantial number of relevant documents would be missed by searching either collection by itself.
- Combine all relevant documents into one database allows for more efficient searching (but, search scope is potentially increased)
- Know your collections. The properties of different document sets can be quite different. This affects search algorithm design and query strategy.
- Most of the concerns noted can be addressed with proper algorithm and tool design. [3]

Larger, More Diverse Collections Benefit From Special-Purpose Tools

- While combining databases enhances efficiency, further efficiency increases demand better ways of sifting through, and analyzing, large result sets, such as:
  - Clustering
  - Image-based results
  - Query refinement based on sample documents
  - Family “roll up”
  - Many other possibilities: no one tool suffices for all situations [3]

SUMOBRAIN
Big, powerful, smart searching

Real-Time Clustering



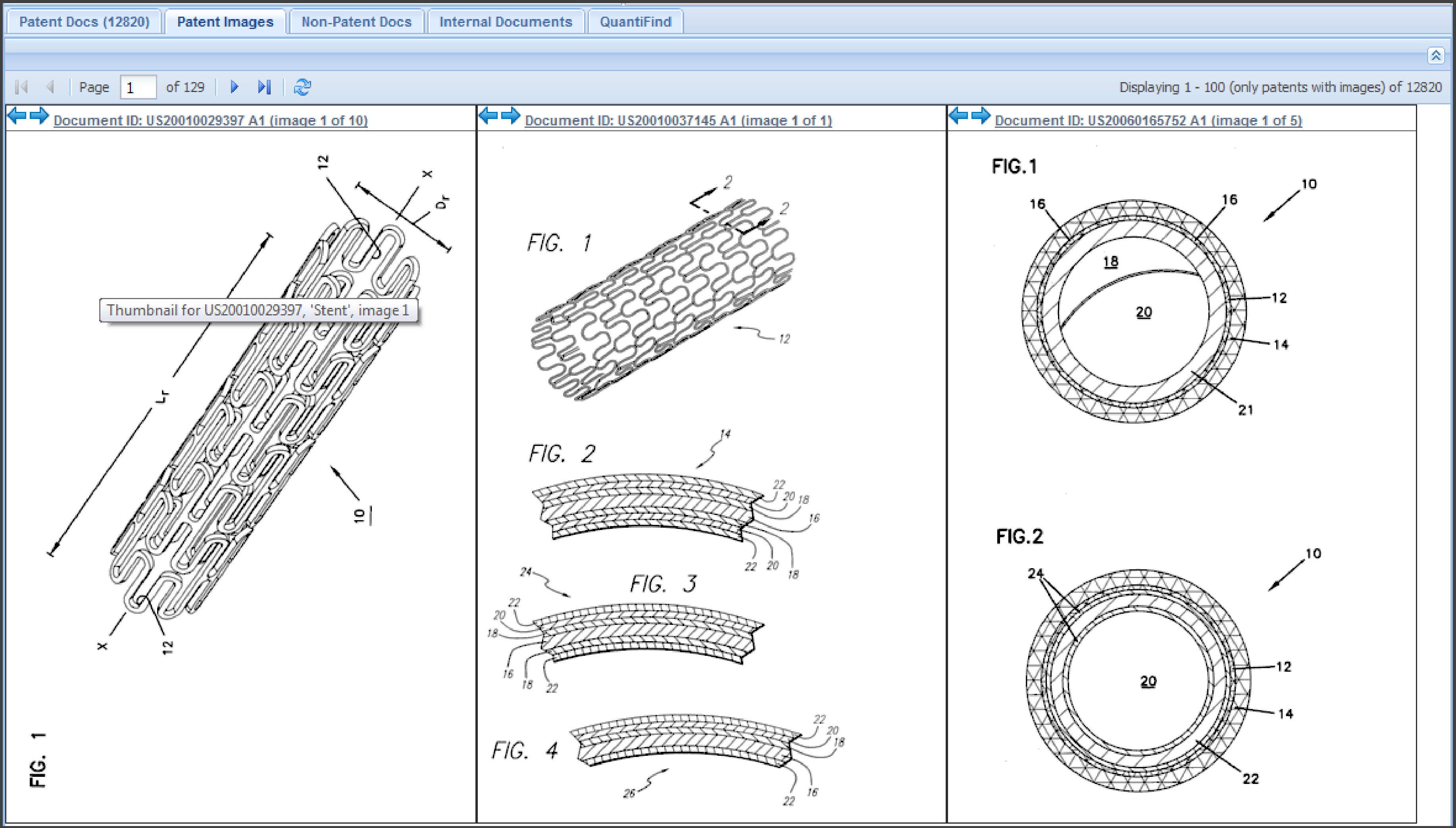
Clusters

- All Topics
  - ⊞ Catheter
  - ⊞ Surface
  - ⊞ Coating Material
  - ⊞ Polymer
  - ⊞ Stent Section
  - ⊞ Intravascular Stent
  - Stent Assembly
  - Polymeric Coating
  - Capable
  - Coating Substance
  - Stent Delivery
  - Tubular Structure
  - Bifurcated Stent
  - Orientation
  - Strut Band
  - Diameter and Length
  - Upper and Lower
  - Axial
  - Durometer
  - Inhibitors

		Document ID	Title	App Date
⊞	1	US6241757	Stent for expanding body's lumen	1998-02-02
⊞	2	US7731742 B2	Stent	2005-02-16
⊞	3	US6558415 B2	Stent	2001-01-17
⊞	4	US6833004 B2	Stent	2002-07-04
⊞	5	US5554181	Stent	1994-05-03
⊞	6	US7384427 B2	Stent	2002-08-19
⊞	7	US6596021 B1	Stent	2000-10-23
⊞	8	US5735872	Stent	1996-10-01
⊞	9	US7927366 B2	Stent	2010-05-02
⊞	10	US7063884 B2	Stent coating	2005-02-21
⊞	11	US6258121	Stent coating	1999-07-01
⊞	12	US7736386 B2	Stent	2008-07-27
⊞	13	US7419502 B2	Stent	2006-11-29
⊞	14	US7985251 B2	Stent	2002-10-15
⊞	15	US6569195 B2	Stent coating	2001-06-17

SUMOBRAIN
Big, powerful, smart searching

Viewing Search Results as Images



Displaying 1 - 100 (only patents with images) of 12820

Document ID: US20010029397 A1 (image 1 of 10)

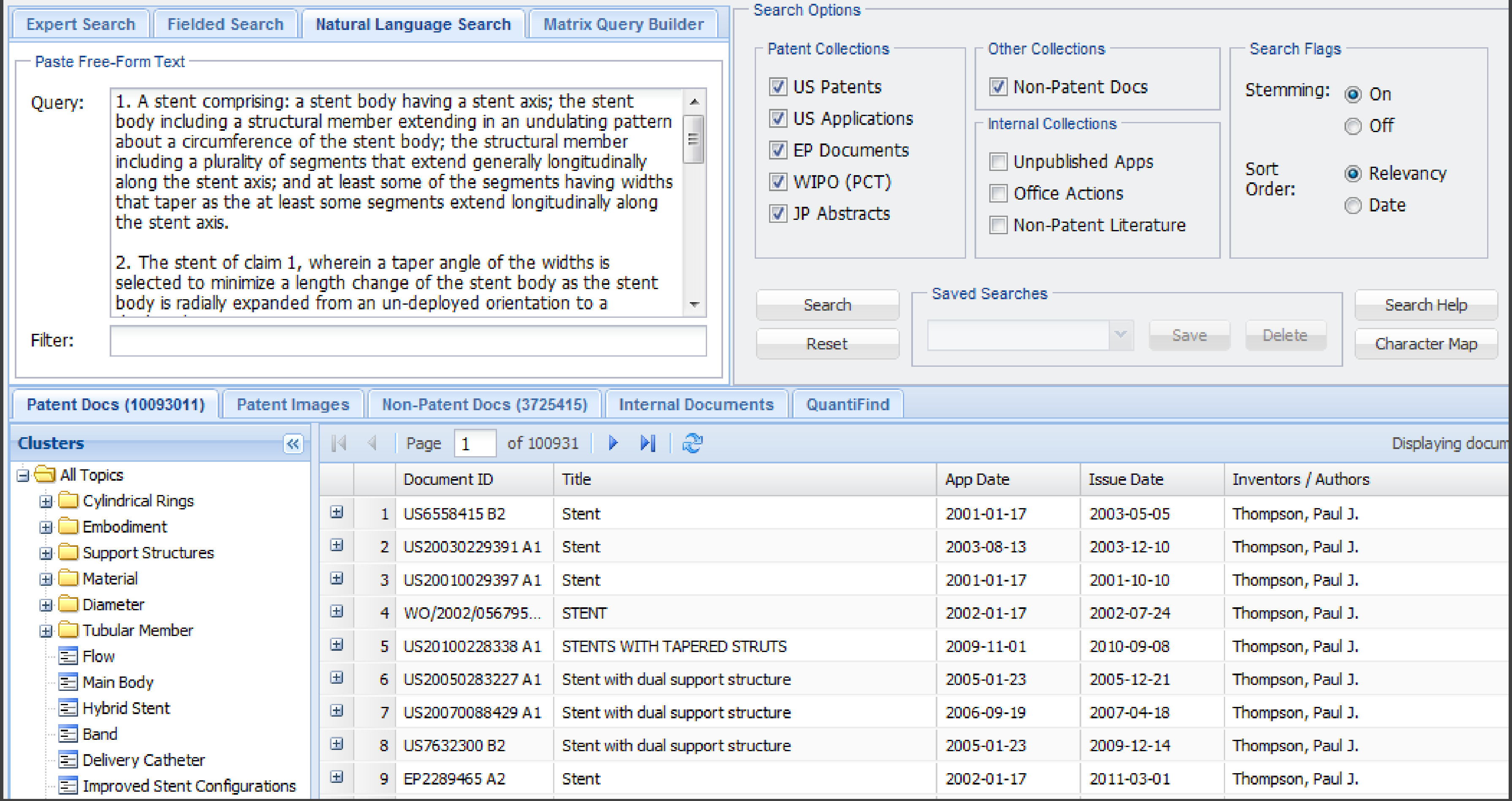
Thumbnail for US20010029397, 'Stent', image 1

Document ID: US20010037145 A1 (image 1 of 1)

Document ID: US20060165752 A1 (image 1 of 5)

SUMOBRAIN
Big, powerful, smart searching

“More Like This”



Paste Free-Form Text

Query: 1. A stent comprising: a stent body having a stent axis; the stent body including a structural member extending in an undulating pattern about a circumference of the stent body; the structural member including a plurality of segments that extend generally longitudinally along the stent axis; and at least some of the segments having widths that taper as the at least some segments extend longitudinally along the stent axis.

2. The stent of claim 1, wherein a taper angle of the widths is selected to minimize a length change of the stent body as the stent body is radially expanded from an un-deployed orientation to a

Filter:

Search Options

Patent Collections
- ☑ US Patents
- ☑ US Applications
- ☑ EP Documents
- ☑ WIPO (PCT)
- ☑ JP Abstracts

Other Collections
- ☑ Non-Patent Docs

Internal Collections
- □ Unpublished Apps
- □ Office Actions
- □ Non-Patent Literature

Search Flags

Stemming: ● On ○ Off

Sort Order: ● Relevancy ○ Date

Saved Searches

Clusters

- All Topics
  - Cylindrical Rings
  - Embodiment
  - Support Structures
  - Material
  - Diameter
  - Tubular Member
  - Flow
  - Main Body
  - Hybrid Stent
  - Band
  - Delivery Catheter
  - Improved Stent Configurations

		Document ID	Title	App Date	Issue Date	Inventors / Authors
⊞	1	US6558415 B2	Stent	2001-01-17	2003-05-05	Thompson, Paul J.
⊞	2	US20030229391 A1	Stent	2003-08-13	2003-12-10	Thompson, Paul J.
⊞	3	US20010029397 A1	Stent	2001-01-17	2001-10-10	Thompson, Paul J.
⊞	4	WO/2002/056795...	STENT	2002-01-17	2002-07-24	Thompson, Paul J.
⊞	5	US20100228338 A1	STENTS WITH TAPERED STRUTS	2009-11-01	2010-09-08	Thompson, Paul J.
⊞	6	US20050283227 A1	Stent with dual support structure	2005-01-23	2005-12-21	Thompson, Paul J.
⊞	7	US20070088429 A1	Stent with dual support structure	2006-09-19	2007-04-18	Thompson, Paul J.
⊞	8	US7632300 B2	Stent with dual support structure	2005-01-23	2009-12-14	Thompson, Paul J.
⊞	9	EP2289465 A2	Stent	2002-01-17	2011-03-01	Thompson, Paul J.

Overall Conclusions

- The combination of patent and non-patent literature in the biomedical space is just one example of a federated database for legal, scientific, or business intelligence search. Many spaces would benefit from such combinations.

- When the data sets and the searcher’s goals are well-understood, specific tools, algorithms, and meta data can be used to substantially increase search efficiency and power.

Thanks!

If next-generation search interests you, please get in touch. Comments, questions and collaboration are welcome. We can only scratch the surface in 30 minutes!

james.ryley@sumobrain.com

Foot Notes

1. The average biomedical patent has 7.29* family members. This phenomena does not exist in the non-patent literature. Family “roll up” (combining all family members into one entry in the search results) can be used to make patent search more efficient.

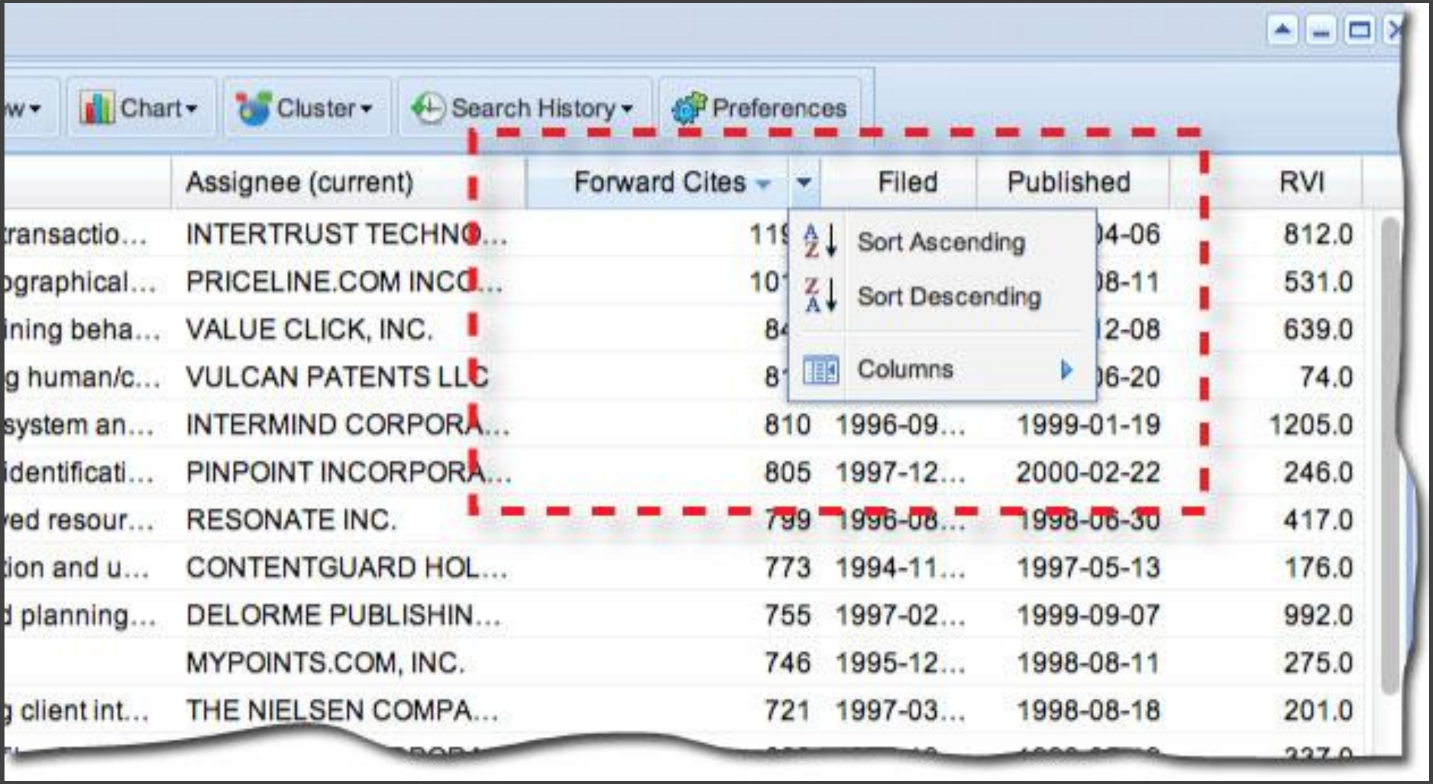
2. The average patent title is 7 words long, while the average non-patent title is 11 words long. These numbers underestimate the difference in information content due to non-informative phrases like “System and method for…” being commonly found in patent titles.

* Using raw EPO data. This data is not always updated post-issuance. We will be analyzing and correcting shortly, but do not expect the number to change substantially.

Footnotes

3. Due to the number of specific search scenarios and the need to address each individually for maximum search efficiency and power, we have created many problem-specific tools, functions, or analytics fields (fields not part of the raw patent data, but calculated from it). Following are screen shots of some of these tools, with brief descriptions. Not all of these tools are publicly available, but collaborations are welcome.

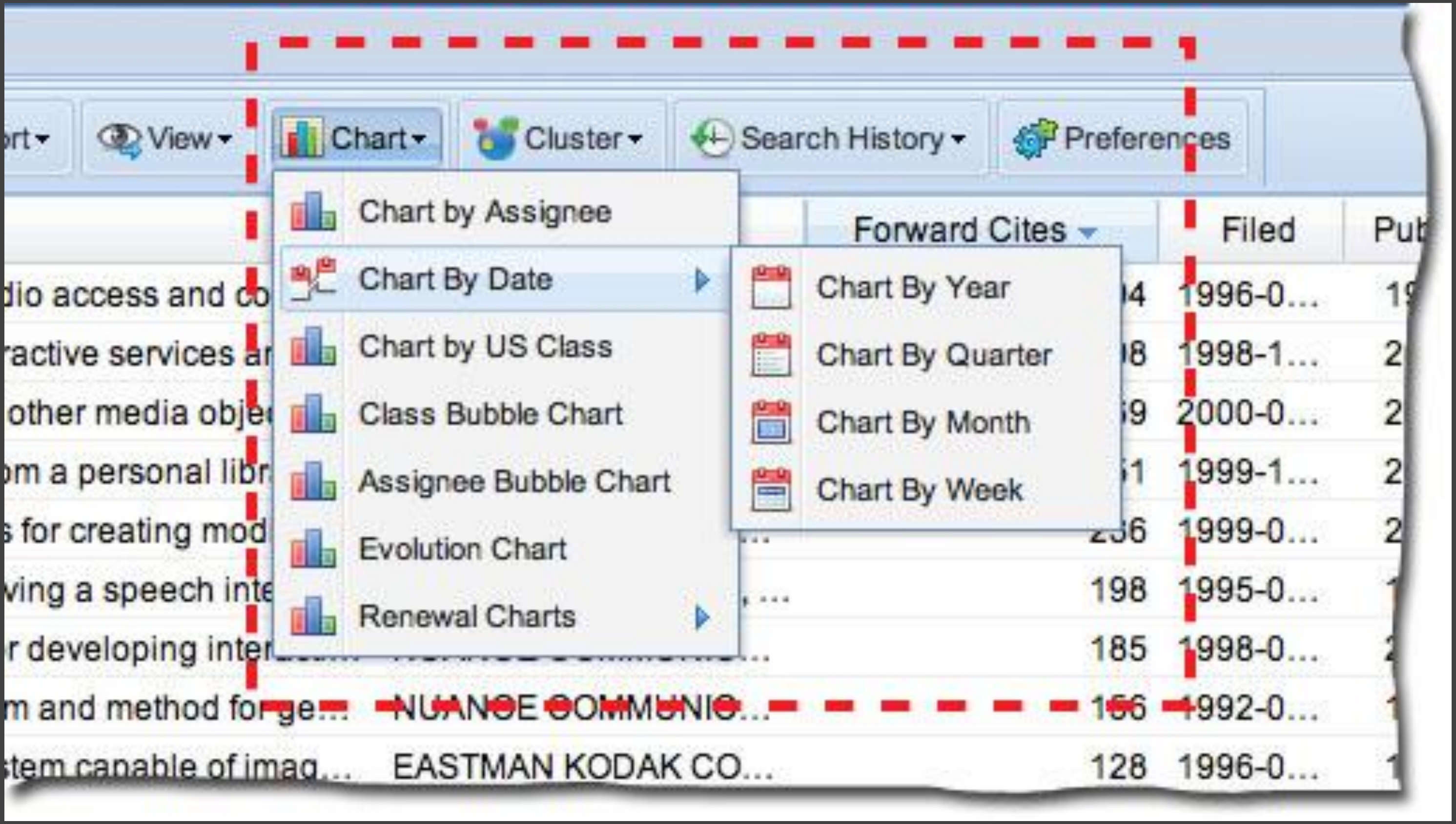
The ability to sort by any column, including calculated analytics fields, is surprisingly handy.



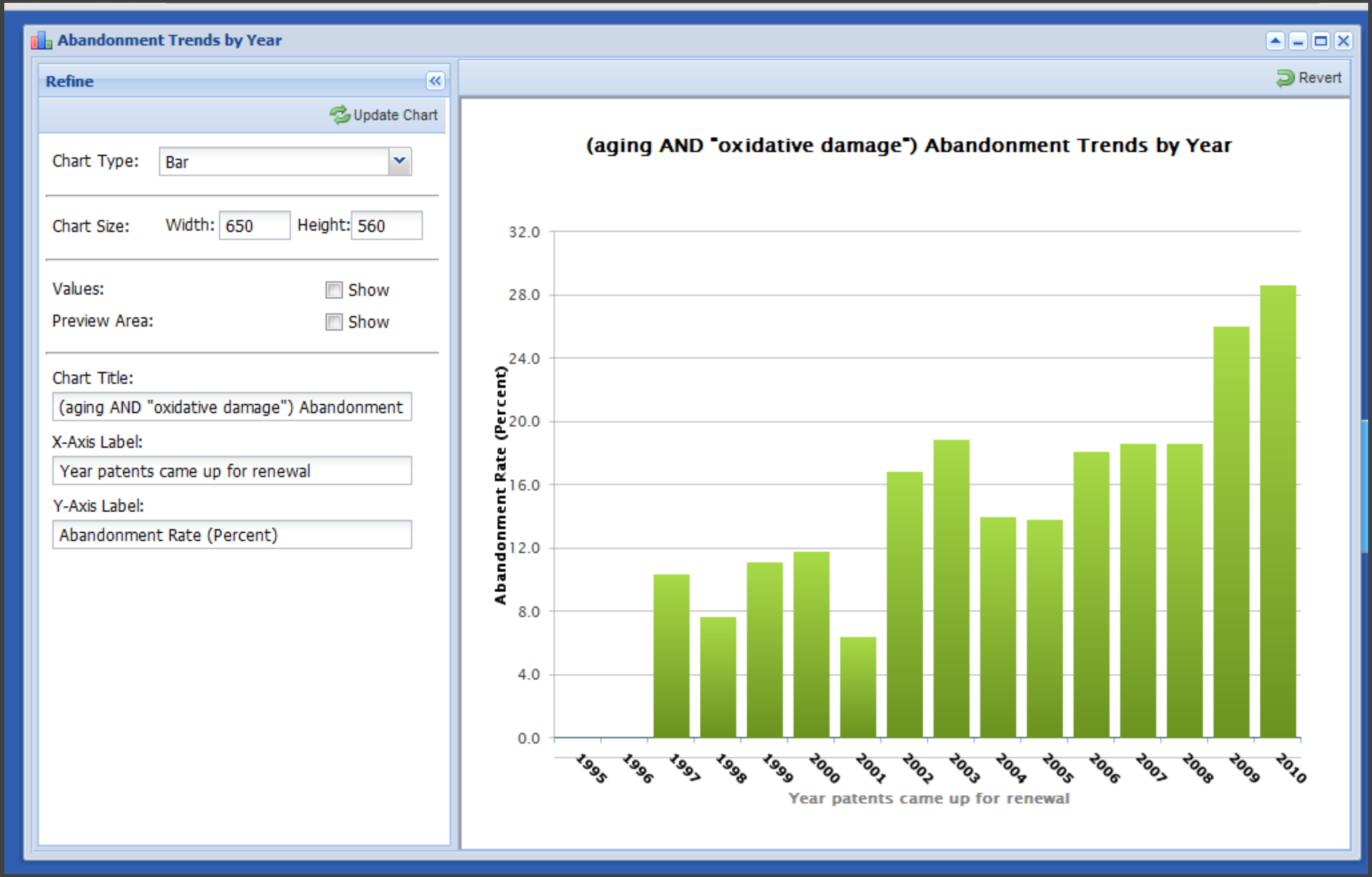
	Assignee (current)	Forward Cites	Filed	Published	RVI
ransactio...	INTERTRUST TECHNO...	11		04-06	812.0
ographical...	PRICELINE.COM INCO...	10		08-11	531.0
ining beha...	VALUE CLICK, INC.	84		12-08	639.0
g human/c...	VULCAN PATENTS LLC	81		06-20	74.0
system an...	INTERMIND CORPORA...	810	1996-09...	1999-01-19	1205.0
identificati...	PINPOINT INCORPORA...	805	1997-12...	2000-02-22	246.0
ed resour...	RESONATE INC.	799	1996-08...	1998-06-30	417.0
ion and u...	CONTENTGUARD HOL...	773	1994-11...	1997-05-13	176.0
d planning...	DELORME PUBLISHIN...	755	1997-02...	1999-09-07	992.0
	MYPOINTS.COM, INC.	746	1995-12...	1998-08-11	275.0
g client int...	THE NIELSEN COMPA...	721	1997-03...	1998-08-18	201.0

(“RVI” is “Relative Value Index” – a metric we use in patent valuation)

Charting is indispensable for Business Intelligence visualization



Charting can show any 2 (or more in some cases) dimensions



Abandonment Trends by Year

Refine

Update Chart

Chart Type: Bar

Chart Size: Width: 650 Height: 560

Values: □ Show

Preview Area: □ Show

Chart Title: (aging AND "oxidative damage") Abandonment

X-Axis Label: Year patents came up for renewal

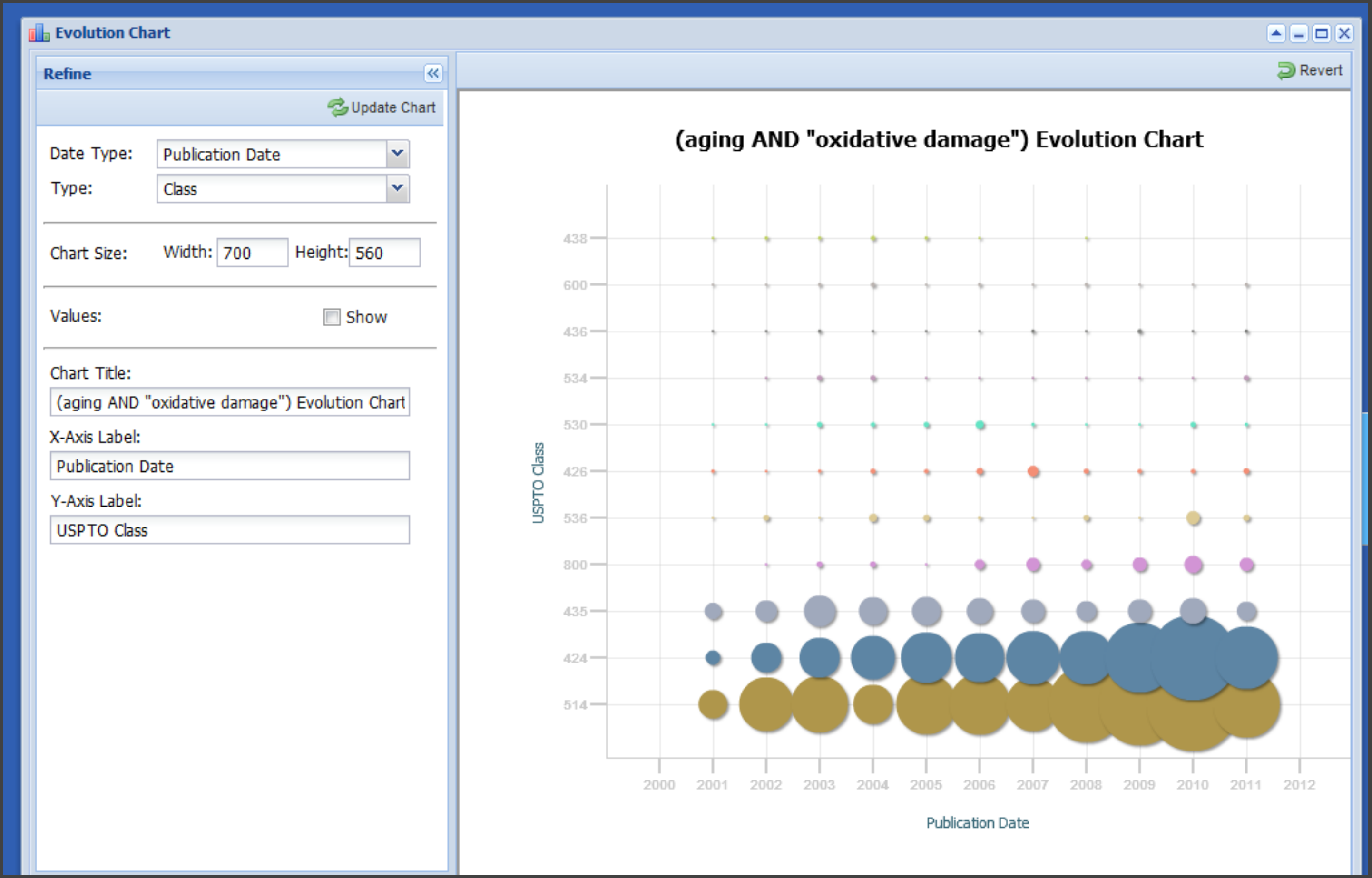
Y-Axis Label: Abandonment Rate (Percent)

Revert

(aging AND "oxidative damage") Abandonment Trends by Year
Abandonment Rate (Percent)
32.0
28.0
24.0
20.0
16.0
12.0
8.0
4.0
0.0
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year patents came up for renewal

SUMOBRAIN
Big, powerful, smart searching

An Evolution Chart Can Show Categorical Change Over Time



Date Type: Publication Date

Type: Class

Chart Size: Width: 700 Height: 560

Values: □ Show

Chart Title: (aging AND "oxidative damage") Evolution Chart

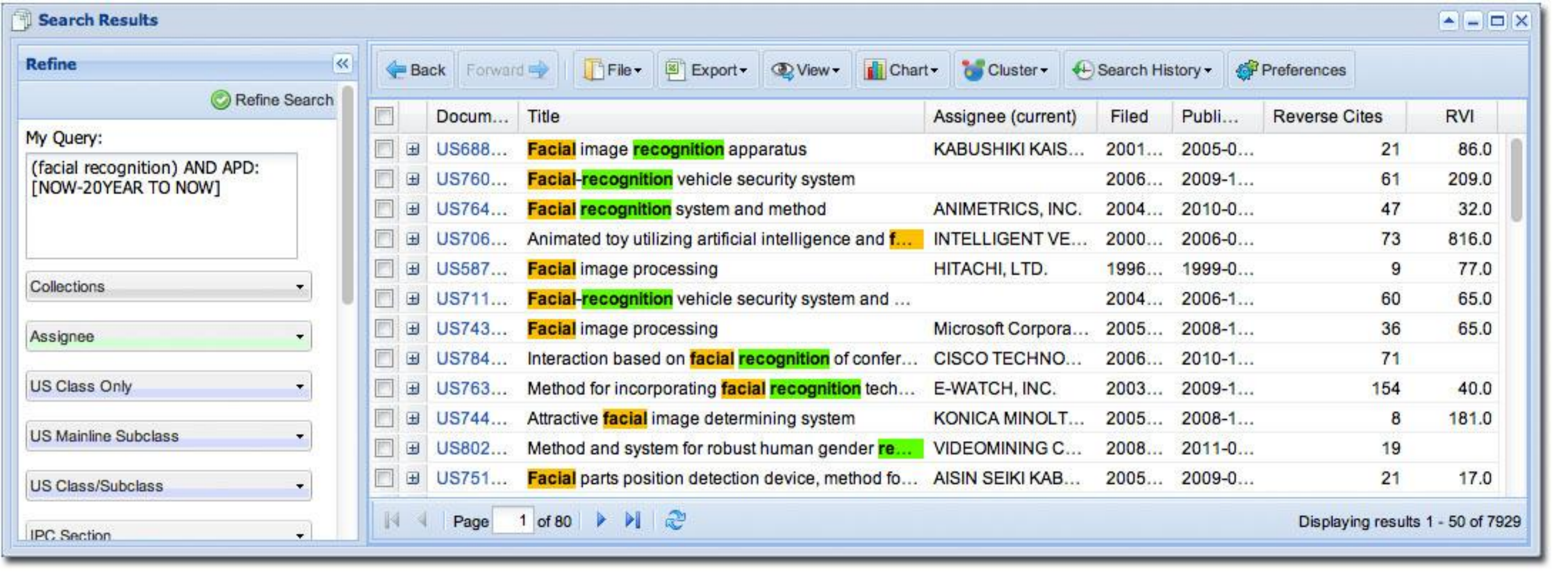
X-Axis Label: Publication Date

Y-Axis Label: USPTO Class

(aging AND "oxidative damage") Evolution Chart

SUMOBRAIN
Big, powerful, smart searching

Search term highlighting: Crucial to fast review of documents



My Query:
(facial recognition) AND APD: [NOW-20YEAR TO NOW]

□		Docum...	Title	Assignee (current)	Filed	Publi...	Reverse Cites	RVI
□	⊞	US688...	Facial image recognition apparatus	KABUSHIKI KAIS...	2001...	2005-0...	21	86.0
□	⊞	US760...	Facial-recognition vehicle security system		2006...	2009-1...	61	209.0
□	⊞	US764...	Facial recognition system and method	ANIMETRICS, INC.	2004...	2010-0...	47	32.0
□	⊞	US706...	Animated toy utilizing artificial intelligence and f...	INTELLIGENT VE...	2000...	2006-0...	73	816.0
□	⊞	US587...	Facial image processing	HITACHI, LTD.	1996...	1999-0...	9	77.0
□	⊞	US711...	Facial-recognition vehicle security system and ...		2004...	2006-1...	60	65.0
□	⊞	US743...	Facial image processing	Microsoft Corpora...	2005...	2008-1...	36	65.0
□	⊞	US784...	Interaction based on facial recognition of confer...	CISCO TECHNO...	2006...	2010-1...	71	
□	⊞	US763...	Method for incorporating facial recognition tech...	E-WATCH, INC.	2003...	2009-1...	154	40.0
□	⊞	US744...	Attractive facial image determining system	KONICA MINOLT...	2005...	2008-1...	8	181.0
□	⊞	US802...	Method and system for robust human gender re...	VIDEOMINING C...	2008...	2011-0...	19	
□	⊞	US751...	Facial parts position detection device, method fo...	AISIN SEIKI KAB...	2005...	2009-0...	21	17.0

Page 1 of 80

Displaying results 1 - 50 of 7929

(The options on the left are “facets” which let you instantly filter docs by many different criteria)