

# Visualization and text mining of patent and non-patent data

Anton Heijs

Treparel Information Solutions  
Delft, The Netherlands  
<http://www.treparel.com/>

ICIC conference , Nice , France , 2008

## Outline

**Introduction**

**Applications on patent and non-patent literature**

**Patent mining and visualization**

**Mining and visualization Medline**

**Conclusions and future development**

## Introduction

### Mining + visualization = visual analytics

- ▶ More data - less insight
- ▶ Overview needed

### Text mining approaches controlling over/under-fit

**Rule based** : difficult to control under fit

**Natural language processing** : difficult to control over fit

**Machine learning** : controls over-fit and under-fit

## Document clustering : finding patterns unsupervised

### Unsupervised text mining : clustering

- ▶ **Input** is a vector representation of all documents
- ▶ Separate the data in its natural groups based on a similarity metric
- ▶ **Output** is a similarity matrix

## Document classification : finding patterns supervised

### Supervised text mining : classification

- ▶ **Input** is a vector representation of all documents
- ▶ Determine a hyper plane which separates all the data
  - ▶ Binary classification : classify to 1 class
  - ▶ Multi class classification : classify to multiple classes
  - ▶ Compound classification : combine multiple binary classifiers
- ▶ **Output** is a matrix with classification scores

## Text analytics

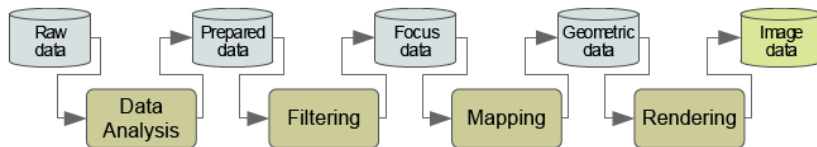
### Text analytics is text mining and visualization

- ▶ Combined use of mining and visualization for analysis
- ▶ Uses text mining to determine trends in the data
- ▶ Use visualization to make complex data understandable

## Applications on patent and non-patent literature

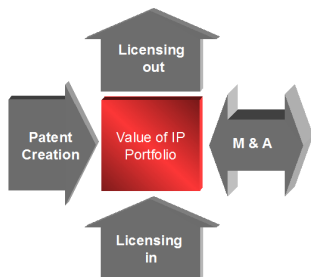
### Text mining and visualization steps

- ▶ management and preprocessing data
- ▶ text mining (filtering the data in the pipeline)
- ▶ mapping filtered data to geometry
- ▶ rendering and interaction



## Patent analytics applications

- ▶ Patent search using classification and clustering algorithms
- ▶ Patent landscaping, spotting new opportunities
- ▶ Patent ranking, utilization optimization and valuation





## Search by query by class

A61

Search Options

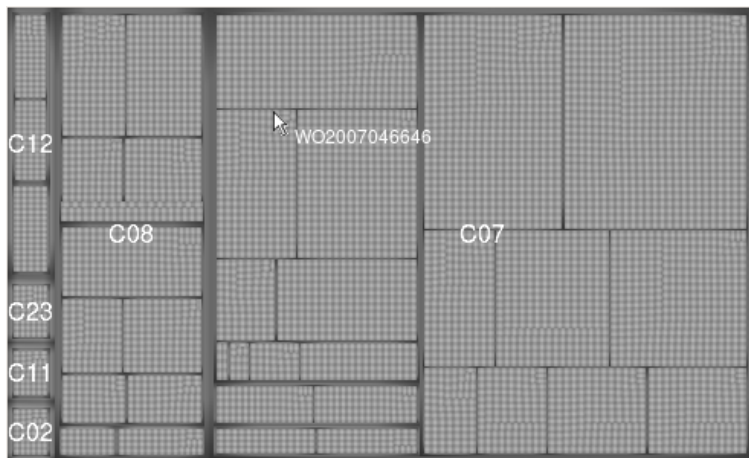
- Case Sensitive
- Allow Wildcard
- Regex
- Create new brush.

Select search columns

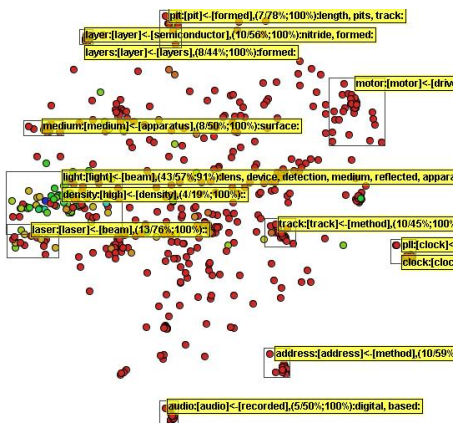
Available Columns	Direction	Selected Columns
CLAIMSEXCER	->	PRECLASS
DATE	<-	
DESCEXCERPT		
DOCNAME		
YEAR		

Navigation: Add All, Remove All

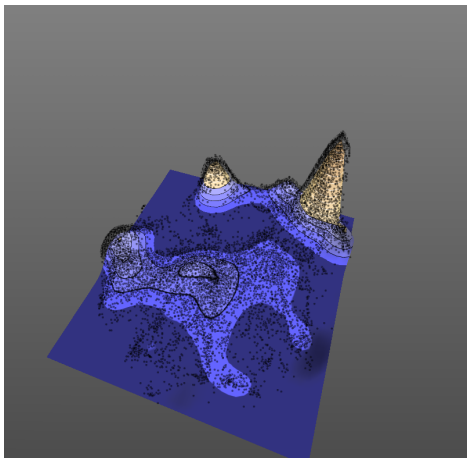
## Tree map visualization of the patent classification



## Clustering of patents of 7 classes in Optical Recording



## Clustering of large patent sets



# Building a classifier

Dataset Features Clustering Classification Postprocessing Iterations View Help

Iteration statistics

**Model performance**

**ebola** Precision: 0.826220650936  
Recall: 0.794117647059

Previous selected: 2  
Current selected: 2

Iteration statistics

DOCNAME	AB
17944642	Handmade influenza is unpredictable, and the risk of an avian flu
17944619	The avian influenza pandemic knowledge about avian H5N1
17955902	The claim that migratory birds are responsible for the long-stay
17952040	Ebola virus is a highly virulent pathogen capable of inducing a f
17944272	SARS-CoV, human coronavirus NL63 (HCoV-NL63) and HCoV
17944271	The identification in 2003 of a coronavirus as the aetiological ag
17944269	Highly pathogenic H5N1 influenza viruses have become endem
17944268	Osteonin phosphate is a product of osteonin carboxylate

Colour	Brush	Items
1	h5n1	76
2	ebola	38
3	sars	58

- \* Initializing 3 clusters
- \* Using cosine distance measure
- \* Using complete linkage
- 14:47:15: Finished clustering
- 14:48:00: Creating classification iteration...
- \* Creating model
- \* Classifying documents
- 14:48:51: Finished creating iteration 1
- 14:55:37: Creating classification iteration...
- \* Creating model
- \* Classifying documents
- 14:55:39: Finished creating iteration 2

ebola h5n1 sars

The renin-angiotensin system (RAS) plays a key role in maintaining blood pressure homeostasis, as well as fluid and salt balance. Angiotensin II, a key effector peptide of the system, causes vasoconstriction and exerts multiple biological functions. Angiotensin-converting enzyme (ACE) plays a central role in generating angiotensin II from angiotensin I, and capillary blood vessels in the lung are one of the major sites of ACE expression and angiotensin II production in the human body. The RAS has been implicated in the pathogenesis of pulmonary hypertension and pulmonary fibrosis, both commonly seen in chronic lung diseases such as chronic obstructive lung disease. Recent studies indicate that the RAS also plays a critical role in acute lung diseases, especially acute respiratory distress syndrome (ARDS).

# Testing a classifier

Dataset Features Clustering Classification Postprocessing Iterations View Help

Optivis - Classification Optimiser

Iteration statistics

Model performance

Precision

Recall

Iteration

Previous selected: 3  
Current selected: 3

show current Compare

Iteration statistics Classification

DOCNAME Search Options

DOCNAME	AB
14735439	Sinai.gov was declared to be eradicated on 8 May 1980, during
14735416	PURPOSE OF REVIEW: Severe acute respiratory syndrome (SARS)
14735504	Of the reemerging transmissible diseases of the post-1945 era,
14735290	The recent epidemic of severe acute respiratory syndrome (SARS)
14735734	The identification of a new coronavirus as the etiologic agent
14715987	Group A streptococci causes hemorrhagic fever with high fever
14685547	Influenza A and B viruses cause seasonal respiratory illness
14677006	Severe acute respiratory syndrome (SARS) is a new disease

Colour	Brush	Items
1	h5n1	210
2	ebola	126
3	sars	257
4	influenza	0

Add Brush

14:48:50 - Finished creating iteration 1...  
\* Creating model  
\* Classifying documents

14:48:51 - Finished creating iteration 1  
14:53:37 - Finished creating iteration 1...  
\* Creating model  
\* Classifying documents

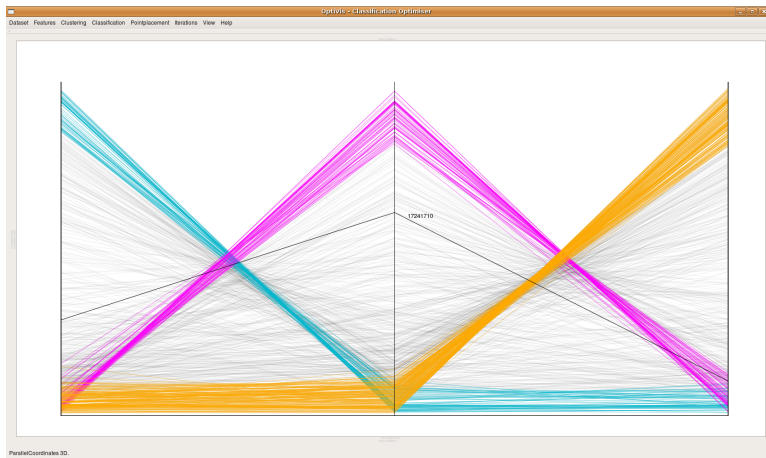
14:55:59 - Finished creating iteration 2  
15:09:26 - Finished creating iteration 2...  
\* Creating model  
\* Classifying documents

15:09:37 - Finished creating iteration 3

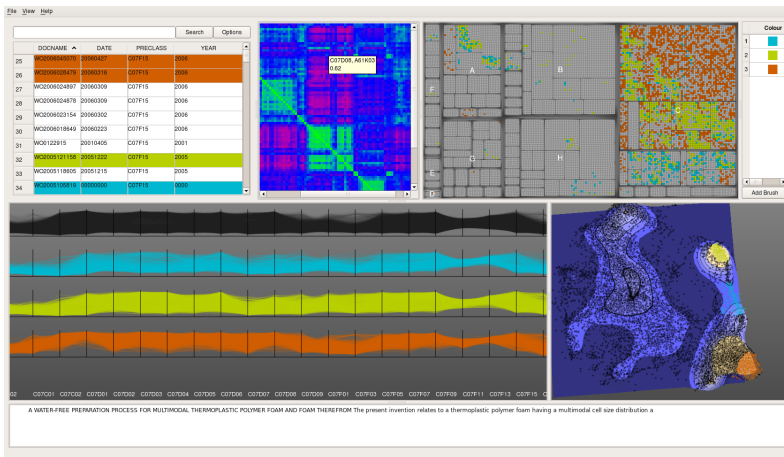
ebola h5n1 sars

Severe Acute Respiratory Syndrome (SARS) caused by new corona-virus SARS-CoV was not identified at humans and animals up to this time. A characteristic feature of this disease is biphasic course. First high fever, parainfluenza syndrome followed by increasing respiratory distress. The main role in transmission is droplet way, things contaminated by excreta, blood transfusion is not excluded. Pathomorphological changes are: bronchial epithelial denudation, loss of cilia, squamous metaplasia, a giant cells infiltrate of macrophages in the alveoli, haemorrhagic oedema, atrophy white pulp of the spleen. Diagnosis is based on clinical picture and epidemiological data supported by positive serology, PCR or presence virus in cell culture. In treatment ribavirin, steroids and mechanical ventilation are

## Visualization of the classifier performance

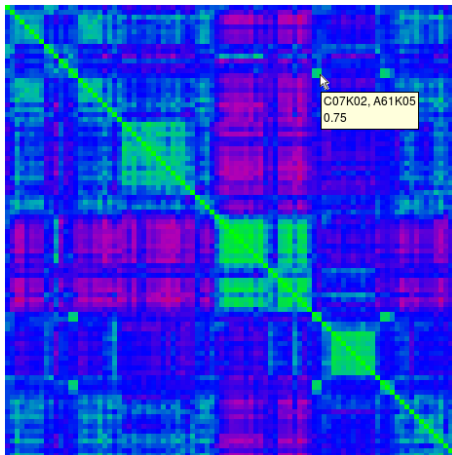


# Patent analytics application combining mining and visualization

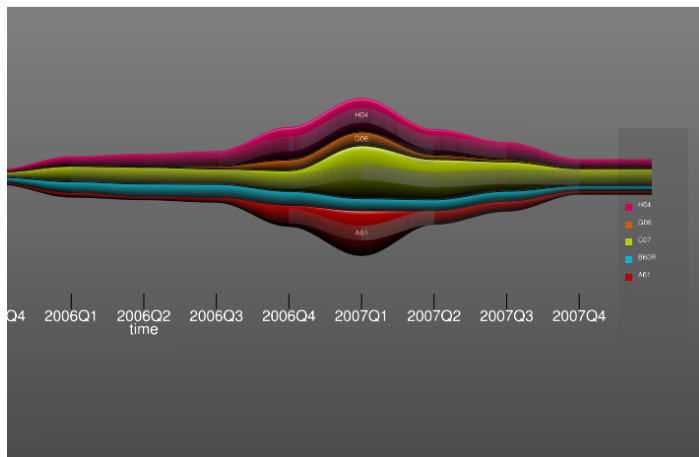




## Correlation between patents



## Patents over time

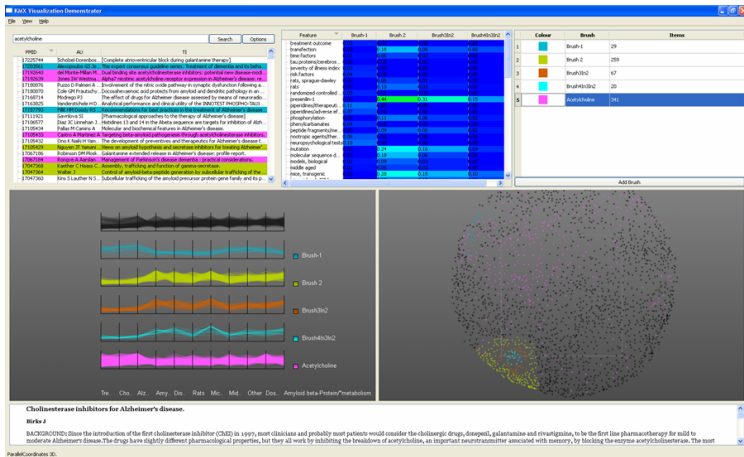




## Text mining and visualizations applications on Medline

- ▶ Search for facts and relationships (using classification and clustering)
- ▶ Disambiguation (using classification)
- ▶ Named entity recognition (using classification)
- ▶ Concept and relationships discovery
- ▶ Knowledge discovery by integrating text mining and visualization with
  - ▶ Data mining : mining table data from experiments
  - ▶ Image mining : micro array analysis
  - ▶ Graph mining : path way analysis

# Mining and visualization Medline



**KAXX Visualizations Demo viewer**

Search Options

acetylcholine

Feature	Brush-1	Brush-2	Brush3n2	Brush4n3n2
treatment outcome	0.00	0.00	0.00	0.00
benefaction	0.00	0.15	0.00	0.00
time factors	0.00	0.00	0.00	0.00
low-probability/ambis.	0.00	0.00	0.00	0.00
severity of illness index	0.00	0.00	0.00	0.00
risk factors	0.00	0.00	0.00	0.00
rats, sprague-dawley	0.00	0.00	0.00	0.00
rats	0.00	0.00	0.00	0.00
randomized controlled	0.00	0.00	0.00	0.00
prevalence	0.00	0.00	0.00	0.00
ependyma/teratop.	0.11	0.00	0.11	0.11
ependyma/dience.	0.00	0.00	0.00	0.00
phenylethanolam.	0.00	0.11	0.00	0.00
phenylalaninase	0.00	0.00	0.00	0.00
peptide fragments	0.00	0.00	0.00	0.00
neuropeptides	0.00	0.00	0.00	0.00
neuropeptide analogs	0.00	0.00	0.00	0.00
neuropeptide analogs	0.00	0.00	0.00	0.00
inhibition	0.00	0.24	0.19	0.00
modular organization	0.00	0.19	0.00	0.00
models, biological	0.00	0.00	0.00	0.00
models, animal	0.00	0.00	0.00	0.00
neurotransmission	0.00	0.25	0.19	0.00

Colour Brush Items

Colour	Brush	Items
1	Brush-1	29
2	Brush-2	289
3	Brush3n2	67
4	Brush4n3n2	30
5	Acetylcholine	191

Add brush

Brush 1  
Brush 2  
Brush3n2  
Brush4n3n2  
Acetylcholine

Tr. Ch. Alz. Amy. Dis. Rats. Mice. Moll. Other. Dis. Amyloid beta-Protein\*metabolism

### Cholinesterase inhibitors for Alzheimer's disease.

**Wicks J**

**BACKGROUND:** Since the introduction of the first cholinesterase inhibitor (DHE) in 1997, most clinicians and probably most patients would consider the cholinergic drugs, donepezil, galantamine and rivastigmine, to be the first line pharmacotherapy for mild to moderate Alzheimer disease. The drugs have slightly different pharmacological properties, but they all work by inhibiting the breakdown of acetylcholine, an important neurotransmitter associated with memory, by blocking the enzyme acetylcholinesterase. The most

PubMedCoordinate: 3D.

## Search Medline by query

animal Search Options

Search options

Case Sensitive

Allow Wildcard

Regex

Create new brush.

Select search columns

AD  
AID  
AU  
CI  
CIN  
CN  
CON  
DA  
DCOM  
DEP  
DP  
EDAT  
ETD

->

<-

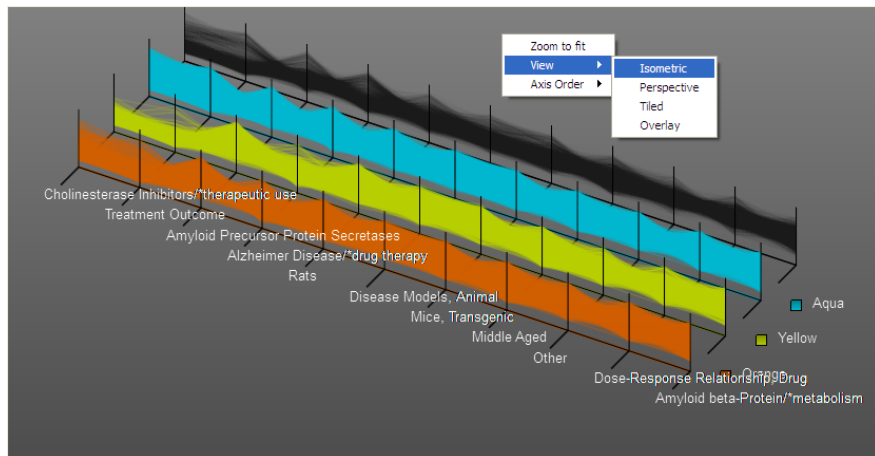
Add All

Remove All

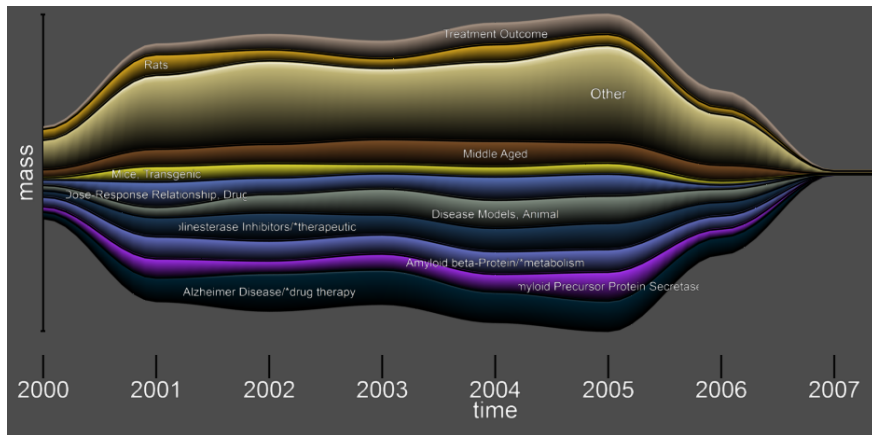
AB

PMID	AU	TI
17225744	Scholzel-Dorenbos...	[Complete atrioventricular block during galantamine therapy]
17203561	Alexopoulos GS Je...	The expert consensus guideline series. Treatment of dementia and its beha...
17192640	del Monte-Millan M...	Dual binding site acetylcholinesterase inhibitors: potential new disease-modi...
17192639	Jones IW Westma...	Alpha7 nicotinic acetylcholine receptor expression in Alzheimer's disease: re...

## Classification of Medline documents to multiple topics



## Visualization of trends in Medline documents over time





## Conclusions

- ▶ Combination of text mining and visualization algorithms enables more application solutions
- ▶ Multiple application solutions are a combination of a small set of text mining algorithms
- ▶ Multiple application solutions require multiple coupled view visualizations to provide complete

## Future development

- ▶ Patent analytics will include patent ranking and valuation
- ▶ Text mining will integrate with data mining , for instance for patent portfolio management
- ▶ Stand alone applications and applications with work flow support will advance

---

## Trends Patterns Relationships

---

<http://www.treparel.com/>

---

Enabling you to learn more and see more

---