Visualization of Statistical and Text Mining Results from Large Document Collections

Anton Heijs

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

ICIC conference , Sitges , Spain , 2009

Outline

Treparel

- Treparel = TREnds PAtterns and RELationships
- KMX = Knowledge Mapping and EXploration
- KMX integrates text-mining with visualization
- Treparel developed KMX Patent Analytics in collaboration with Philips IP&S
- Focus on university collaborations : Delft, Paris, Sao Paulo
- Focus on powerful algorithms for large data sets
- Focus on analysis of patents and non-patent literature
- Focus on text intensive solutions for Pharma/Biotech

Introduction

The need for data mining & visualization approaches

- The amount and complexity of data is increasing rapidly
- There is an increasing need to obtain insight and overview
- There is an increasing need to make better decisions faster
- More analysis tasks become information critical

Treparel focuses on a new visual analytics approach

- Text mining : extracting information
- Statistics : providing descriptive information
- Visualization : visualizing patterns and trends

Text analytics

Text analytics is text mining and visualization

- Combined use of mining and visualization for analysis
- Uses text mining to determine classes or clusters in the data
- Uses statistics to determine characteristics of the data
- Use visualization to identify patterns or trends in the data

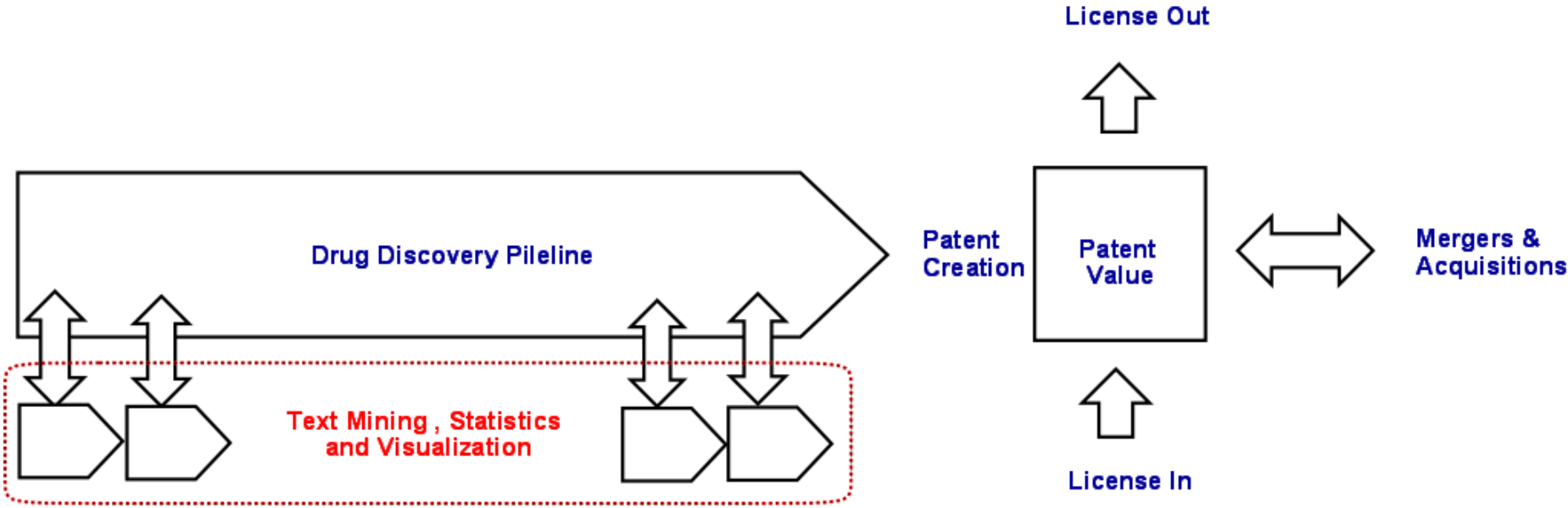
Combined analysis of patent and non-patent literature

The drivers for patent and non patent literature analysis

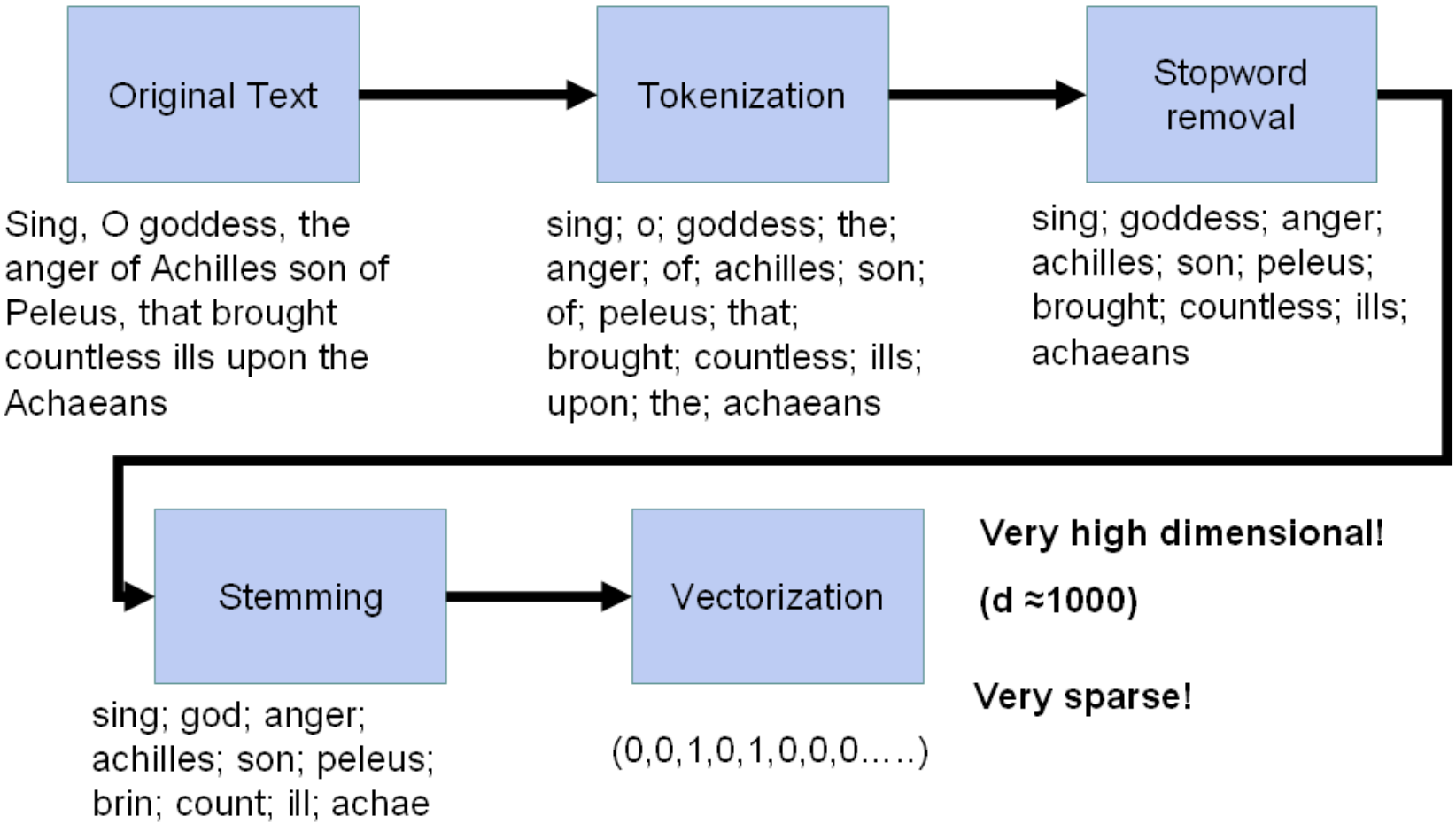
- All relevant documents need to be taken into account
- The amount of documents is growing fast over time
- The utilization of documents will decrease over time
- The need to process larger document sets more often will increase over time
- A combined approach can be more effective

Patent analytics must support patent value creation

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



Text preprocessing to a vector space model



Document clustering and classification

Document clustering : unsupervised text mining

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

Document classification : supervised text mining

- Input is a vector representation of all documents
- Determine a hyper plane which separates the data in classes
- Output is a matrix with doc-class classification scores

Why use classification?

Benefits of classification / drawbacks of keyword search

- Deals better with context: words depending on context or even ambiguity in a single word (homonym)
- Deals better with synonyms: several words or expressions can describe the same concept
- Increased precision: high occurrence of keywords is no guarantee of relevance
- Deals with vague composition: concepts that are hard to describe in a set of relevant keywords
- Able to retain expert knowledge: classifiers highlight relevance and can rank documents

How does automated classification work?

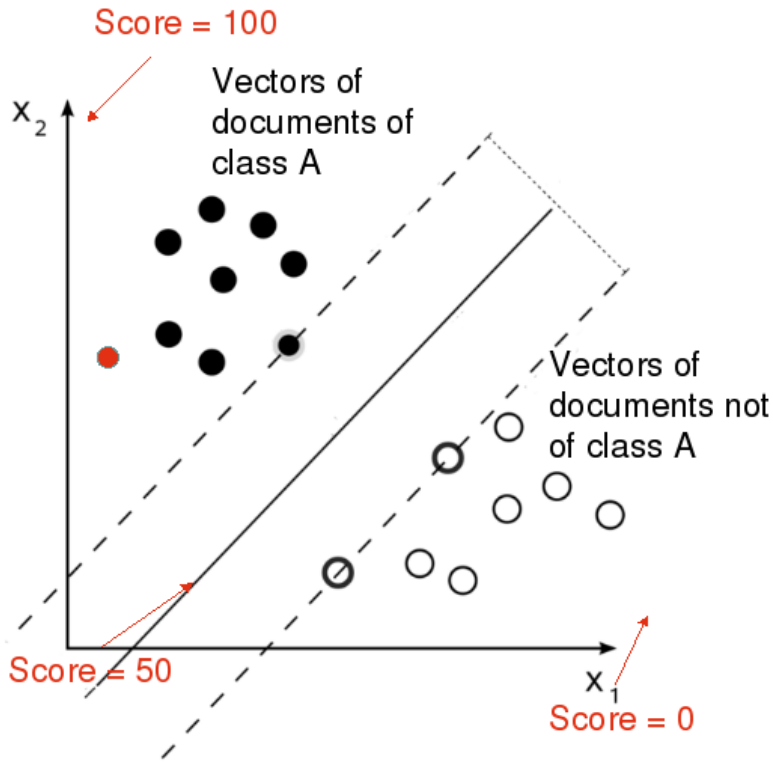
We use the Support Vector Machine algorithm (SVM).

Why SVM?

- SVM enables us to find a good trade-off between precision and recall
- SVM is well-suited for sparse data, text data is often sparse
- SVM is well-suited for high-dimensional data
- SVM is generally applicable, performing very well on a wide variety of tasks

How does automated classification work?

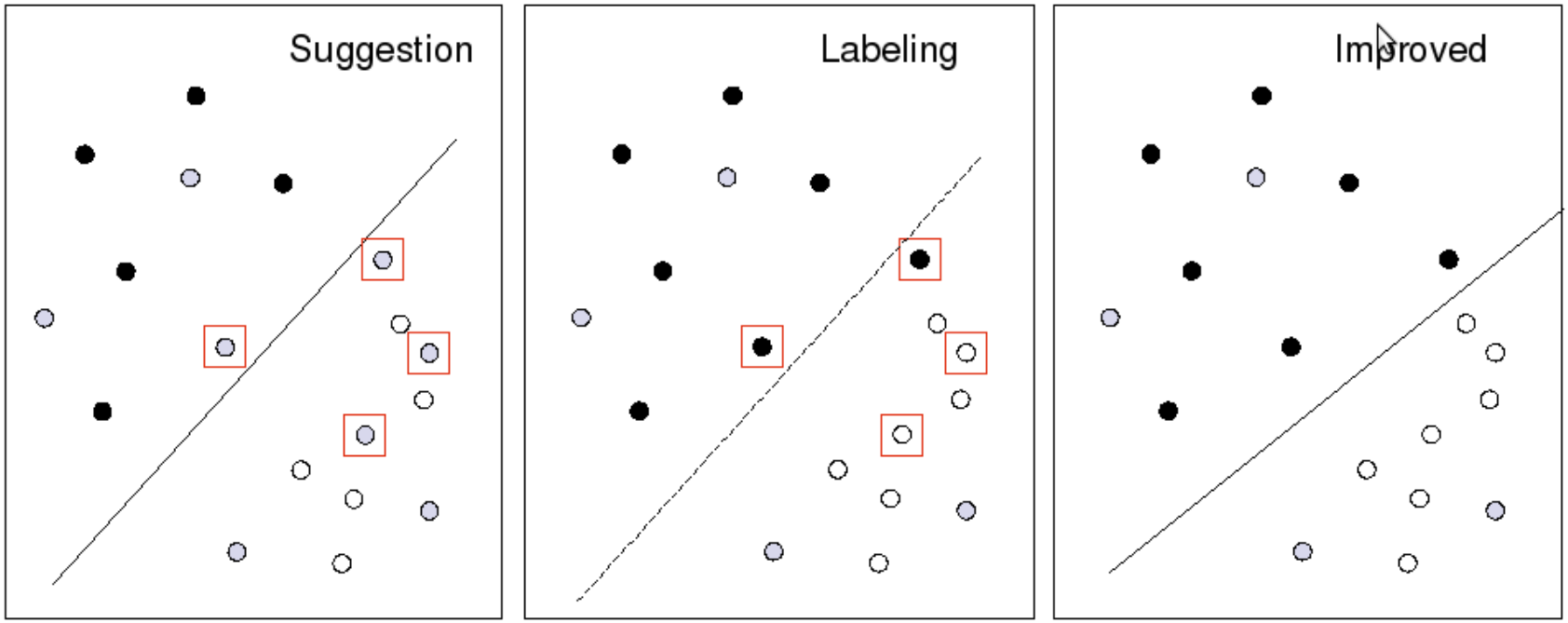
- Classes are separated by a line (d=2) a plane (d=3) or a hyperplane (d>3).
- The SVM algorithm is used to determine the optimal separating hyperplane
- Unknown examples are classified according to their position to the hyperplane.



How does automated classification work?

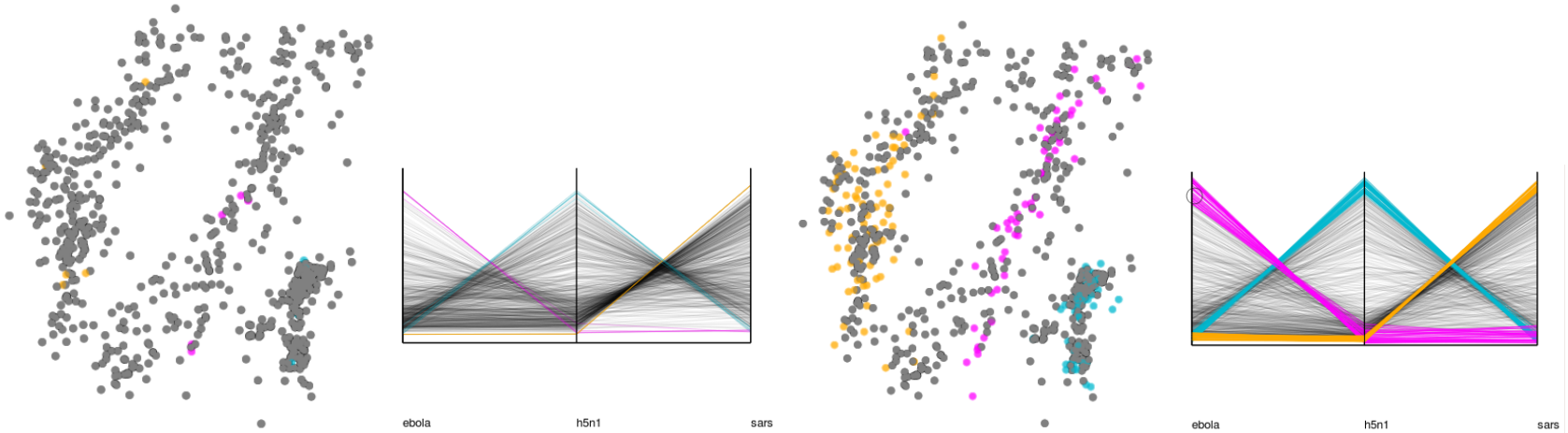
To help create classifiers we added the suggestion system.

After creating the first classifier the suggestion system proposes documents for labeling:

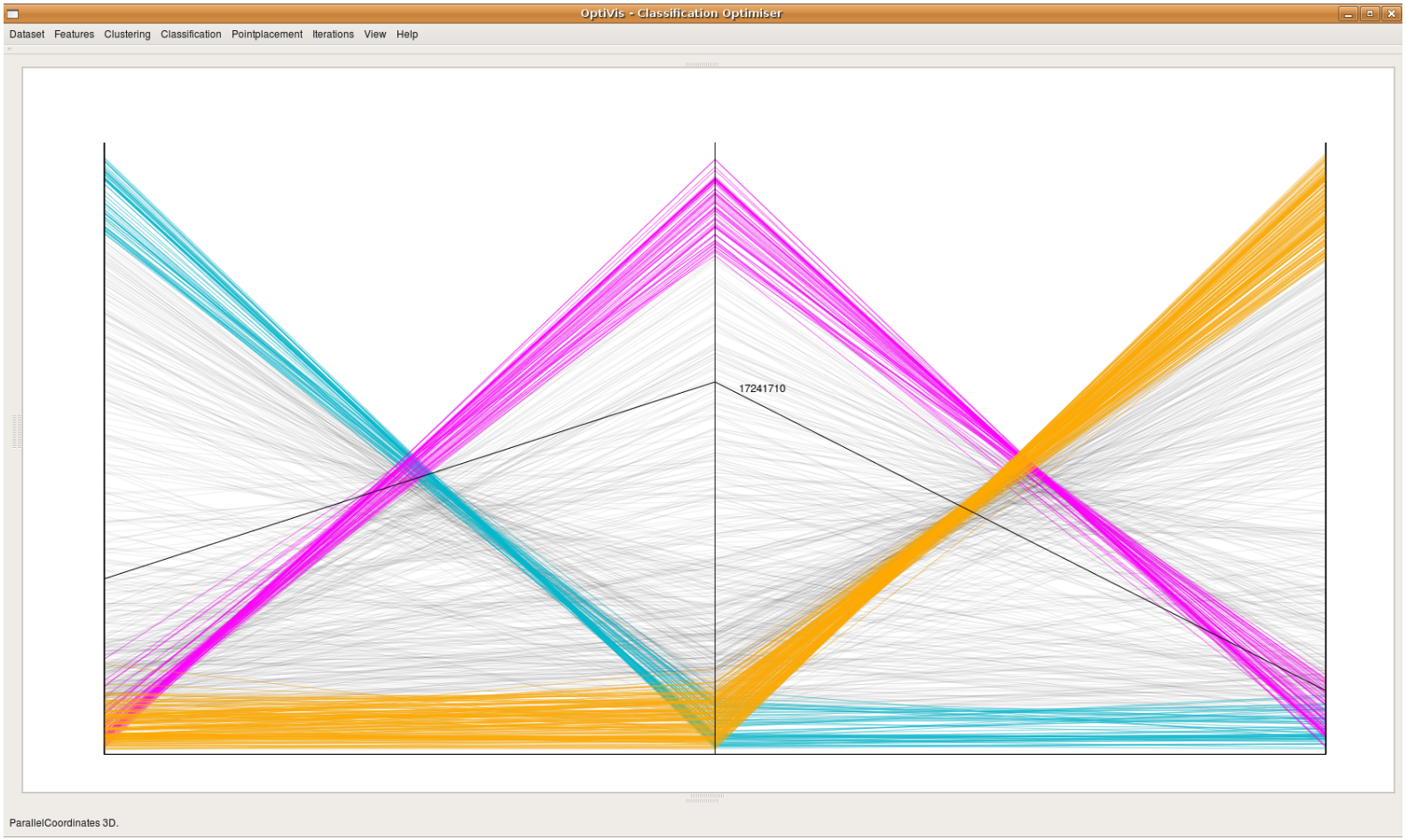


Classification and clustering of 3 classes of Merline documents

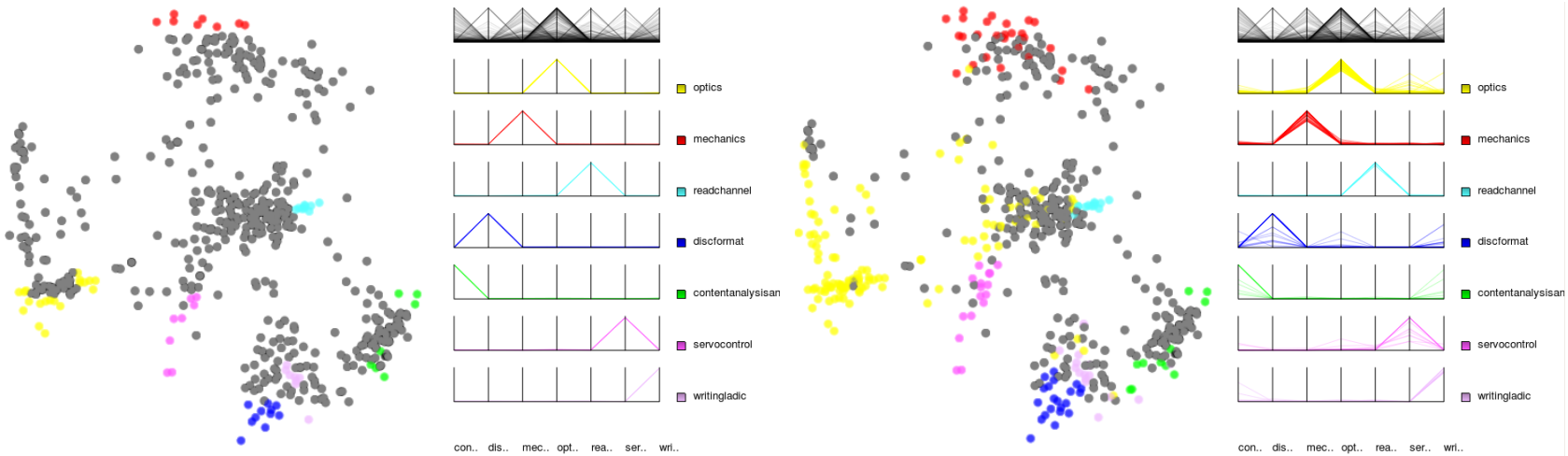
1. Select visually a statistical set of documents from the clusters to build classifiers
2. Visualize the performance of the classifier over all documents



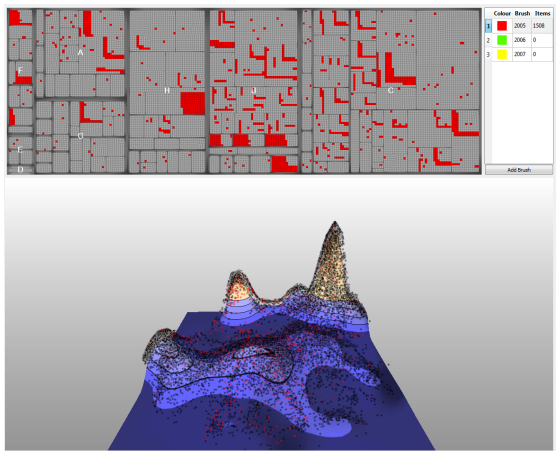
Visualization of the classifier performance

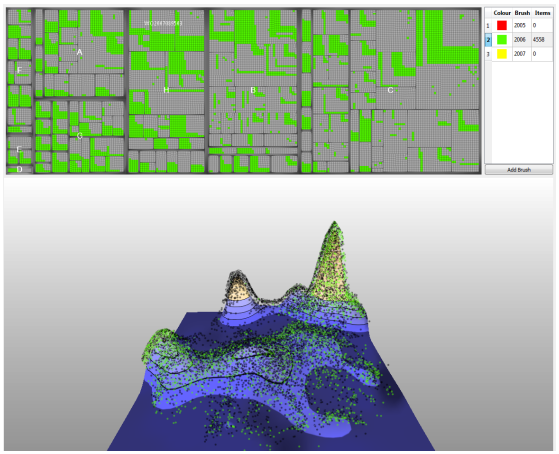


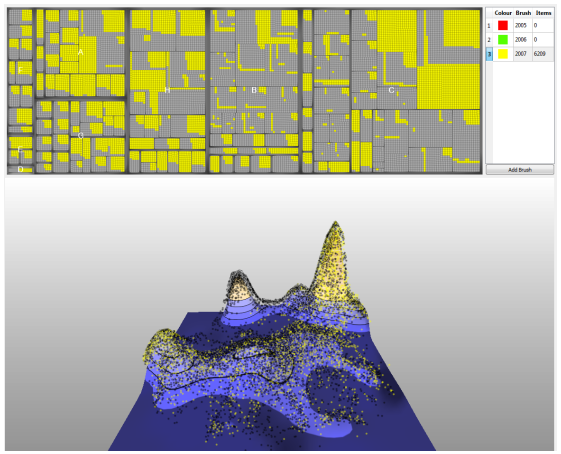
Classification and clustering of 7 classes of patent documents

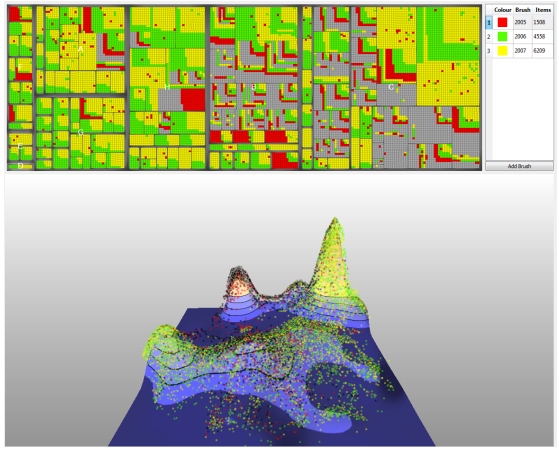


Visualization of patent trends over 2005,2006 and 2007

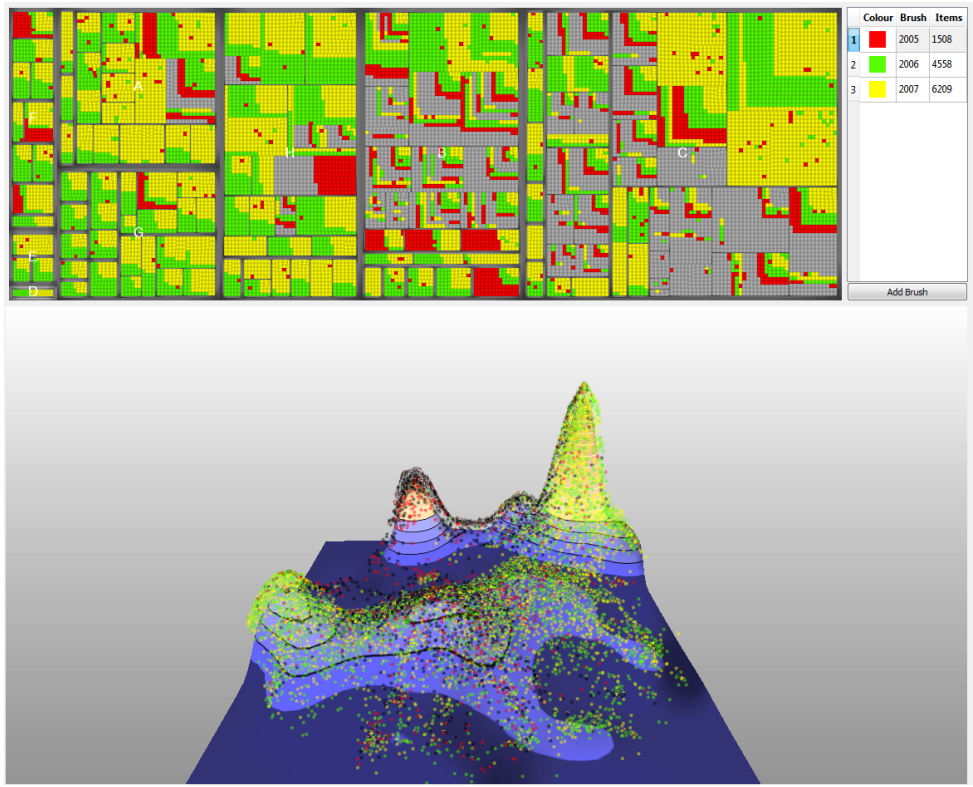






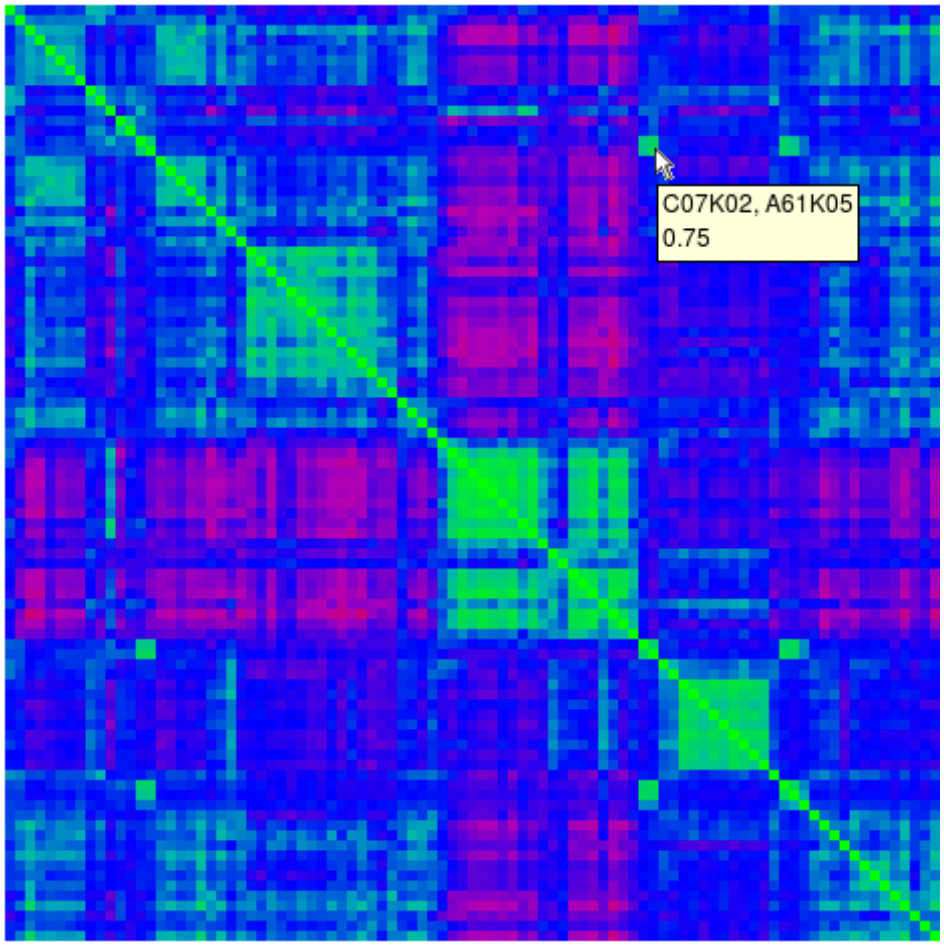


Distribution of patent clusters over time over the patent classification hierarchy

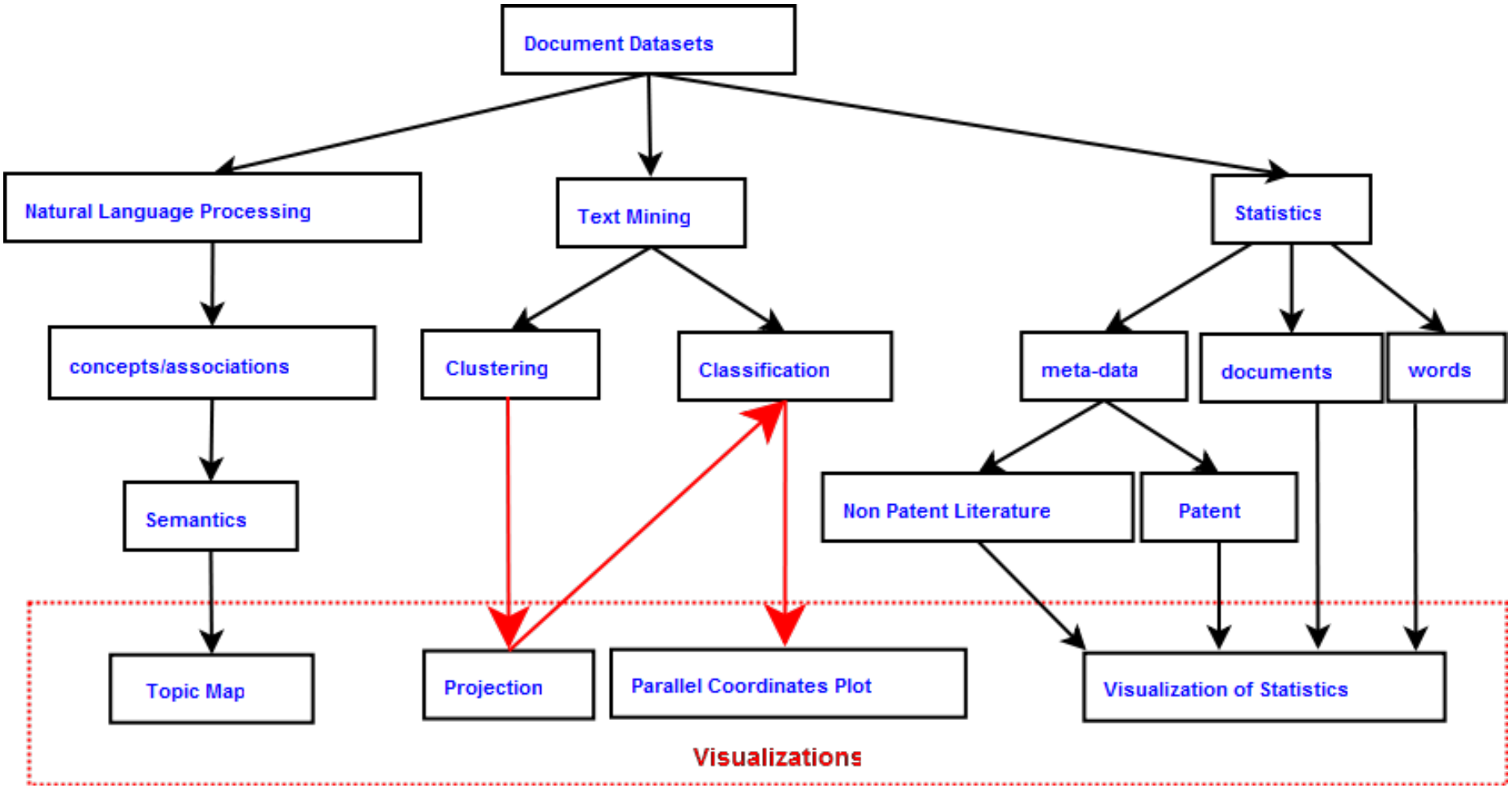


- Visualization provides insight to understand statistics better
- Statistics provides descriptive characteristics

Statistical correlation between patents



Bringing text mining, nlp and statistics to the visualization



Statistics from patent and non patent literature

- Word level : ranked word frequency distribution (Zipf's law)
- Document level : distribution of documents over topics, classes, time
- Meta data level of patents: inventor, assignee, IPC class, number of claims etc
- Meta data leven of non patent literature : MESH terms im Medline

Conclusions

- Text mining and visualization enables analysis of large document sets on patent and non patent literature
- Visualization of large document sets provides insight in patterns and trends
- Statistics provides descriptive information of large document sets
- Visualization of large document sets helps to better understand the statistics of the sets
- Many tasks in patent analysis benefit from text mining, visualization and statistical techniques

Future development

- Text analytics will include patent ranking and statistical utilization and valuation analysis
- Text mining and visualization will integrate more into the patent creation processes such as in drug discovery

Trends Patterns Relationships

http://www.treparel.com/

Enabling you to learn more and see more