# Research Project

## Search Strategy: Automatic Document Tagging

Master Computer Science

Marie Drieghe
Daan Spitael
Diogo Vaz Nunes

23 oktober 2015

# 1 Graph-based

All the papers were found using *Web of Science*
Keywords used:

- graph based document tagging (42 results)
  - An approach to graph-based analysis of textual documents, Antoon Bronselaer (graph-method only withi document,)
    Papers found with backtracking
    - Text classification using graph mining-based feature extraction
    - GRAPH THEORETIC FOUNDATIONS OF PATHFINDER NETWORKS
    - Graph-based text representation model and its realization
    - A Folksonomy Ranking Framework-A Semantic Graph-based Approach (not good, only abstract in English)
- automat* doc* tag* graph* (29 results)
  - Personalized Tag Recommendation Using Graph-based Ranking on Multi-type Interrelated Objects
  - Automatic tag recommendation algorithms for social recommender systems
  - Automatic Structuring of Radiology Free-Text Reports

# 2 Structure-based

All the papers were found using *Web of Science*
First I looked up our reference article and checked the citing articles (22) and the related records (45979). In the related records I only went through the first couple of pages and found:

- A tutorial on support vector machine-based methods for classification problems in chemometrics: interesting because it might offer a way to build a model to classify document types

In the cited articles I found:

- Multi-label learning: a review of the state of the art and ongoing research: might be interesting, but not accesible

Keyword learning to rank (6694) in title (168):

- Joint Structural Learning to Rank with Deep Linear Feature Learning: based on structure of multimedia document but may be generalisable

Keyword learning to rank structure-based (22):

- nothing

Keyword structure-based clustering (1238):

- A weighted common structure based clustering technique for XML documents: focused on XML, may be generalisable
- A Graph-Structure-Based Method for Chinese Document Representation towards ClusteringApplication: focused on not interpretable text, so might work on scientific documents

Keyword structure-based clustering in title (44):

- A Structure-Based Clustering on LDAP Directory Information: interesting because it focusses on documents in directories
- Clustering XML documents by structure based on common neighbor

Keyword structure-based document classification (21):

- Structure-sensitive learning of text types
- Sequential pattern mining for structure-based XML document classification
- Improving recognition accuracy on structured documents by learning structural patterns
- First order Gaussian graphs for efficient structure classification

Keyword document structure classification in title (22):

- Feature Vector Construction Combining Structure and Content for Document Classification
- A simple, structure-sensitive approach for web document classification: works with DOM structure

Keyword machine learning document structure in title (6):

- A machine-learning approach for analyzing document layout structures with two reading orders: considers document layout
- Analyzing document logic structure by machine learning: using logical structure to categorize

# 3 Multi-Label Learning

All the papers were found using *Web of Science*
The topic of Multi-Label Learning was found by looking at citations of the original paper. Keywords used:

- multi label learning text doc* (52 results)

  - ML-KNN A lazy learning approach to multi-label learning

  - Multi-label learning based on iterative label propagation over graph
  - A Tutorial on Multilabel Learning
  - Introduction to the special issue on learning

## 4 Machine Learning in Document Tagging

All the papers were found using *Web of Science*
Keywords used:

- machine learning document* text* (942 results) (Order by most cited)
  - Machine Learning in Automated Text Categorization