INTRAFIND

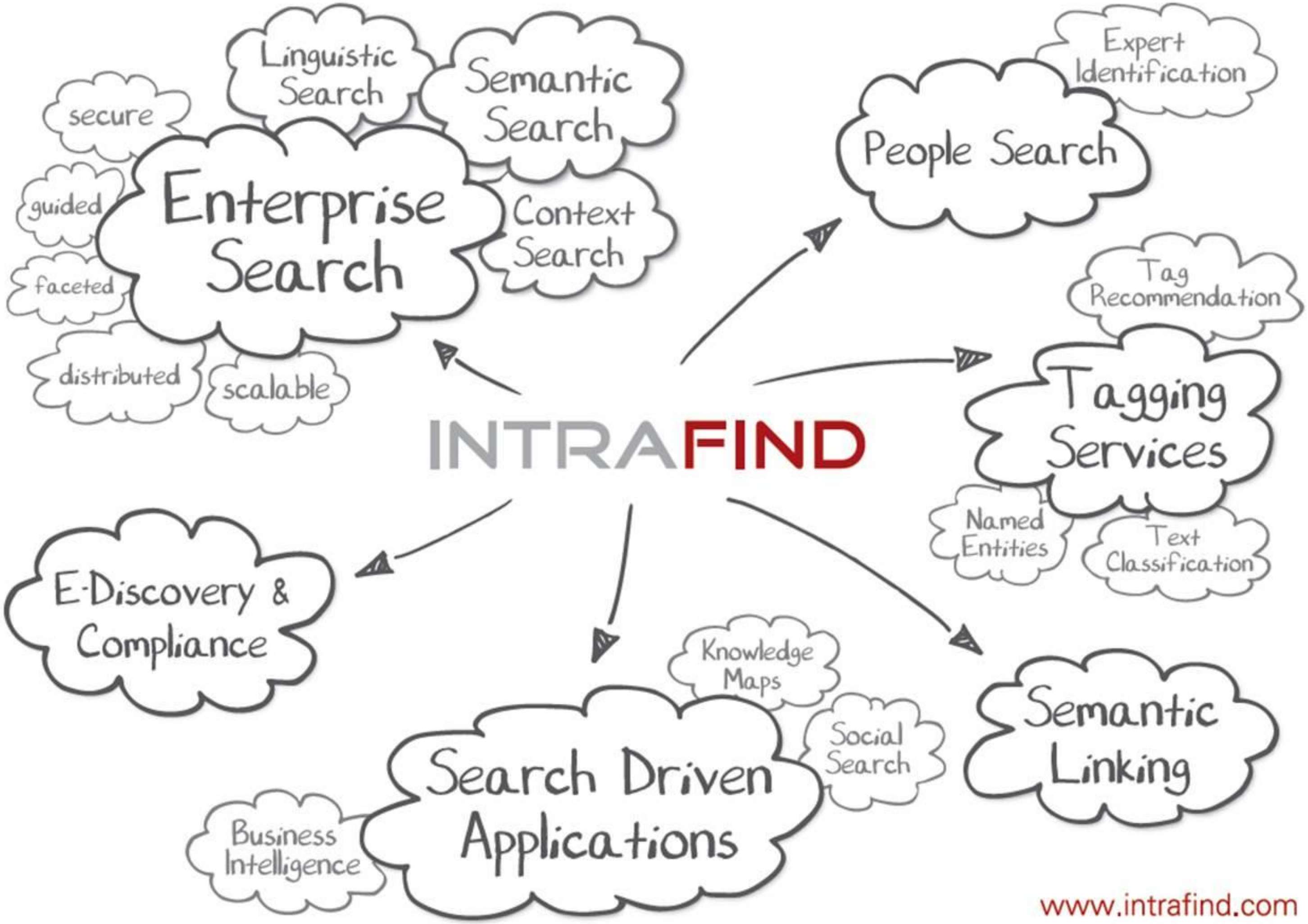
Text Classification based on Lucene and LibSVM / LibLinear

Berlin Buzzwords, June 4th, 2012,
Dr. Christoph Goller, IntraFind Software AG

Outline

- IntraFind Software AG
- Introduction to Text Classification
  - What is it?
  - Applications
  - Lessons Learned
  - Required Features
- Implementation Details
  - Lucene, LibSVM / LibLinear
  - Feature Selection & Training
  - Production Phase: HyperplaneQuery

IntraFind Software AG



IntraFind Software AG

- Founding of the company: October 2000
- More than 700 customers mainly in Germany, Austria, and Switzerland
- Partner Network (> 30 VAR & embedding partners)
- Employees: 30
- Lucene Committers: B. Messer, C. Goller

Our Open Source Search Business:

- **Product Company**: iFinder, **Topic Finder**, Knowledge Map, Tagging Service, …
- Products are a combination of Open Source Components and in-house Development
- Support (up to 7x24), Services, Training, Stable API
- **Automatic Generation of Semantics**
  - Linguistic Analyzers for most European Languages
  - Semantic Search
  - Named Entity Recognition
  - Text Classification
  - Clustering

www.intrafind.de/jobs

Introduction to Text Classification

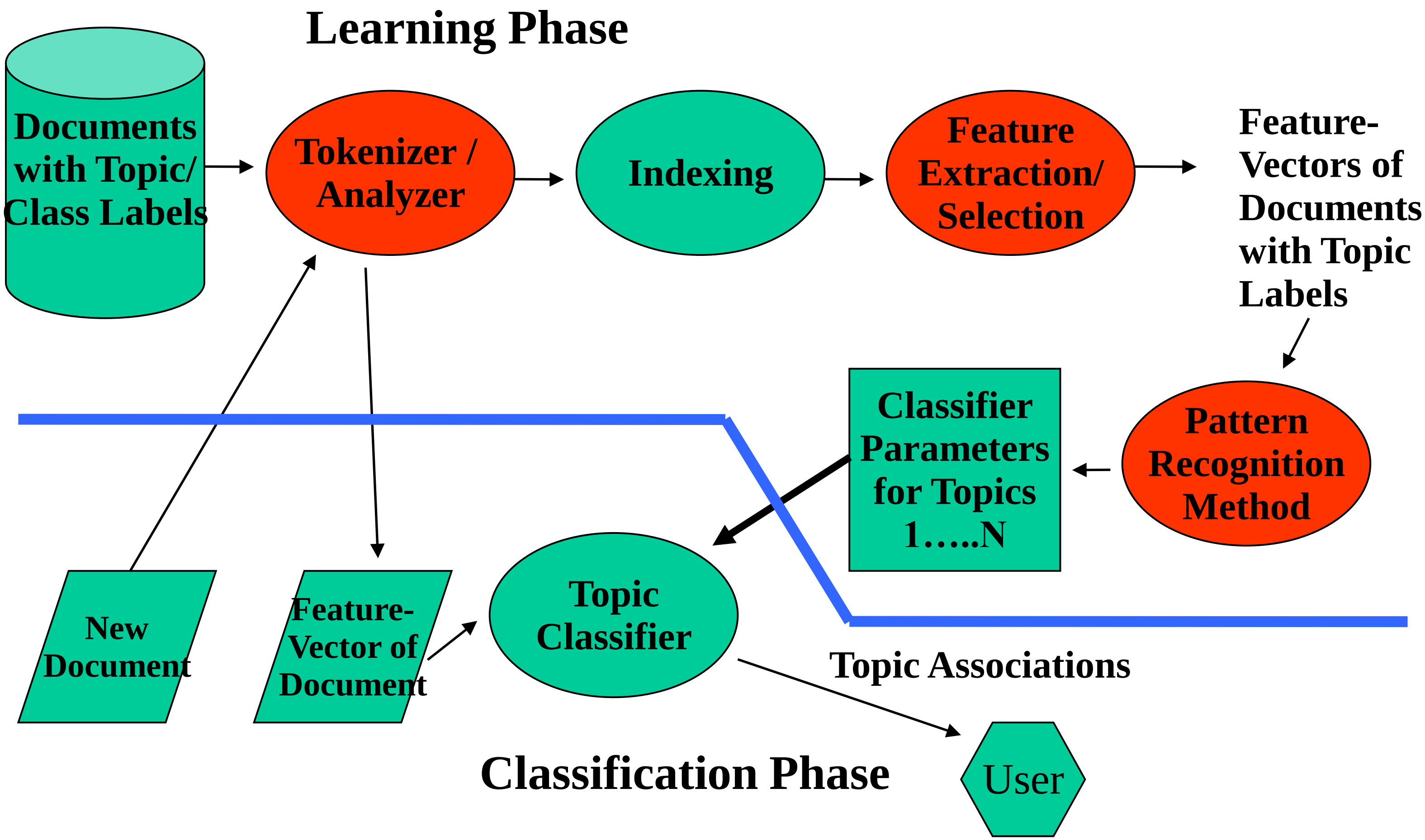
Goal:

- Automatically assign documents to topics based on their content.
- Topics are defined by example documents.

Applications:

- News: Newsletter-Management System
- Spam-Filtering; Mail / Email Classification
- Product Classification (Online Shops), ECLASS /UNSPSC
- Subject Area Assignment for Libraries & Publishing Companies
- Opinion Mining / Sentiment Detection
- Part of our Tagging Services

Text Classification Workflow


Learning Phase
Documents with Topic/ Class Labels
Tokenizer / Analyzer
Indexing
Feature Extraction/ Selection
Feature-Vectors of Documents with Topic Labels
Classifier Parameters for Topics 1…..N
Pattern Recognition Method
New Document
Feature-Vector of Document
Topic Classifier
Topic Associations
Classification Phase
User

Lessons Learned

- Analysis / Tokenization:
  - Normalization (e.g. Morphological Analyzers) and Stopwords improve classification
- Feature Selection:
  - TF*IDF, Mutual Information, Covariance / Chi Square, ...
  - Multiword Phrases, positive & negative correlation
- Machine Learning:
  - Goal: Good Generalization
  - Avoid Overfitting: „entia non sunt multiplicanda praeter necessitatem“ (Occam´s Razor)
  - SVM: linear is enough
- **Don’t trust blindly in**
  - **Manual Classification by Experts**
  - **Statistics / Machine Learning Results: Test !**

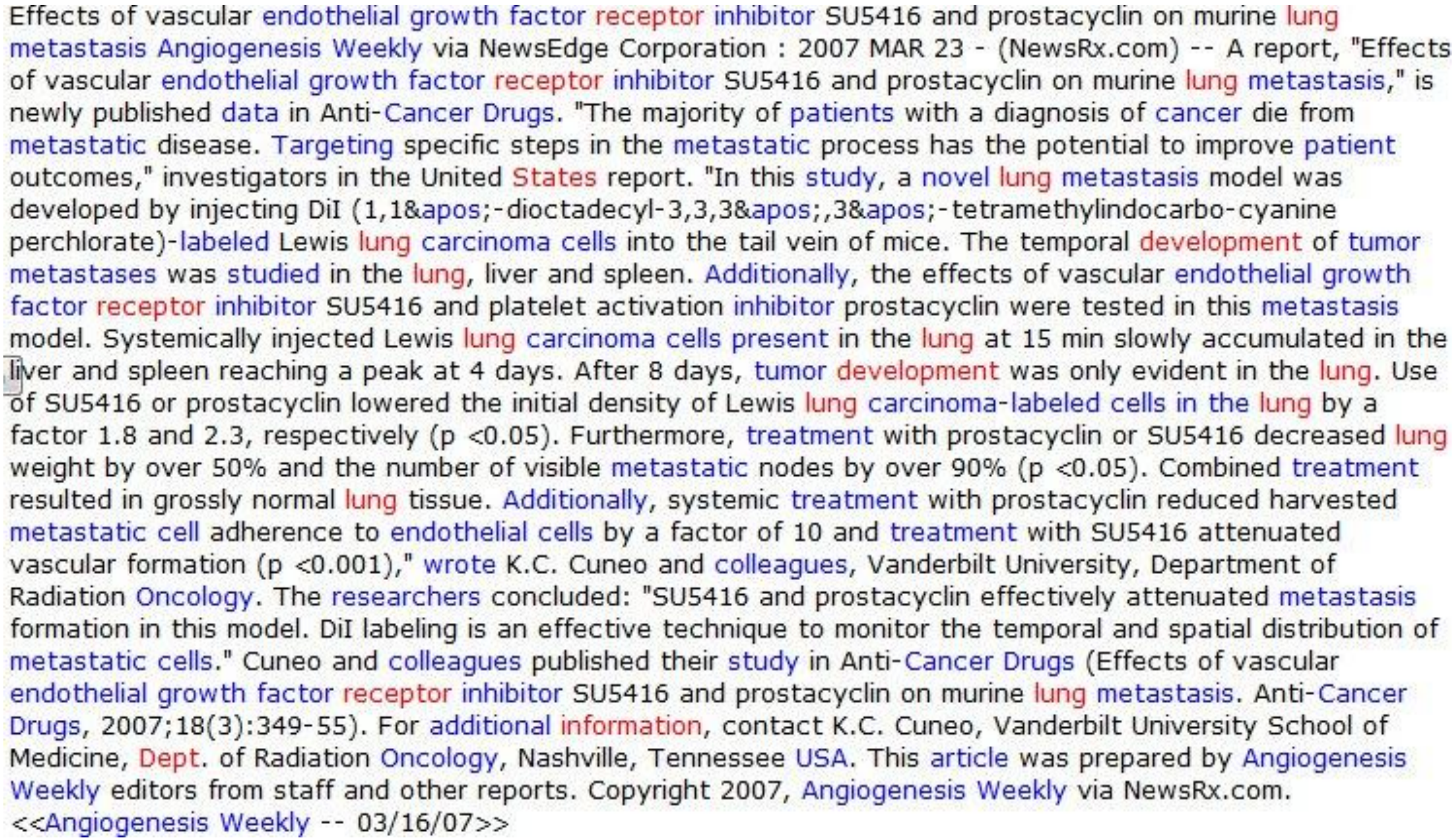
Required Features

- Training & Test GUI needed
- Automatically identify inconsistencies in training & test data
  - Duplicates detection
  - Similarity Search (More Like This)
- Automatic Testing: Cross-Validation (Multi-Threaded!)
- **Classification Rules have to be readable**
- **False Positive and (False Negative) Analysis,**
  - **Iterative Training**
  - **Clustering of False Positive / False Negative**

Product Classification: Example Rules

- **Server:**
  einbauschächte^24.7 | speicherspezifikation^22.1 | tastatur^-0.7 | monitortyp^21.5 | socket^-9.2 -1.15
- **Workstation:**
  monitortyp^28.8 | arbeitsstation^38.8 | cpu^0.1 | tower^8.9 | barebone^35.8 | audio^3.7 | eingang^5.2 | out^6.5 | core^9.0 | agp^5.2 -2.1
- **PC:**
  kleinbetrieb^7.9 | personal^18.3 | db-25^2.2 | technology^5.6 | cache^10.0 | arbeitsstation^-28.1 | dynamic^7.4 | bereitgestelltes^25.7 | dmi^5.5 | ata-100^13.7 | socket^6.2 | wireless^2.5 | 16x^10.0 | 1/2h^13.1 | nvidia^1.0 | din^4.6 | tasten^13.4 | international^7.2 | 802.1p^8.1 | level^-4.4 -1.5
- **Notebook:**
  eingabeperipheriegeräte^64.0 – 1.3
- **Tablet PC:**
  tc4200^16.4 | tablet^6.9 | konvertibel^10.6 | multibay^4.6 | itu^3.3 | abb^2.7 | digitalstift^8.5 | flugzeug^1.8 – 1.75
- **Handheld:**
  bildschirmauflösung^39.8 | smartphone^8.1 | ram^0.29 | speicherkarten^0.53 | telefon^0.35 - 1.4

Pharmaceutical Newsletter: Highlighting Example



Effects of vascular endothelial growth factor receptor inhibitor SU5416 and prostacyclin on murine lung metastasis Angiogenesis Weekly via NewsEdge Corporation : 2007 MAR 23 - (NewsRx.com) -- A report, "Effects of vascular endothelial growth factor receptor inhibitor SU5416 and prostacyclin on murine lung metastasis," is newly published data in Anti-Cancer Drugs. "The majority of patients with a diagnosis of cancer die from metastatic disease. Targeting specific steps in the metastatic process has the potential to improve patient outcomes," investigators in the United States report. "In this study, a novel lung metastasis model was developed by injecting DiI (1,1&apos;-dioctadecyl-3,3,3&apos;,3&apos;-tetramethylindocarbo-cyanine perchlorate)-labeled Lewis lung carcinoma cells into the tail vein of mice. The temporal development of tumor metastases was studied in the lung, liver and spleen. Additionally, the effects of vascular endothelial growth factor receptor inhibitor SU5416 and platelet activation inhibitor prostacyclin were tested in this metastasis model. Systemically injected Lewis lung carcinoma cells present in the lung at 15 min slowly accumulated in the liver and spleen reaching a peak at 4 days. After 8 days, tumor development was only evident in the lung. Use of SU5416 or prostacyclin lowered the initial density of Lewis lung carcinoma-labeled cells in the lung by a factor 1.8 and 2.3, respectively (p <0.05). Furthermore, treatment with prostacyclin or SU5416 decreased lung weight by over 50% and the number of visible metastatic nodes by over 90% (p <0.05). Combined treatment resulted in grossly normal lung tissue. Additionally, systemic treatment with prostacyclin reduced harvested metastatic cell adherence to endothelial cells by a factor of 10 and treatment with SU5416 attenuated vascular formation (p <0.001)," wrote K.C. Cuneo and colleagues, Vanderbilt University, Department of Radiation Oncology. The researchers concluded: "SU5416 and prostacyclin effectively attenuated metastasis formation in this model. DiI labeling is an effective technique to monitor the temporal and spatial distribution of metastatic cells." Cuneo and colleagues published their study in Anti-Cancer Drugs (Effects of vascular endothelial growth factor receptor inhibitor SU5416 and prostacyclin on murine lung metastasis. Anti-Cancer Drugs, 2007;18(3):349-55). For additional information, contact K.C. Cuneo, Vanderbilt University School of Medicine, Dept. of Radiation Oncology, Nashville, Tennessee USA. This article was prepared by Angiogenesis Weekly editors from staff and other reports. Copyright 2007, Angiogenesis Weekly via NewsRx.com.
<<Angiogenesis Weekly -- 03/16/07>>

Lucene, LibSVM & Liblinear

- Apache Lucene (http://lucene.apache.org/):
  - Built in late 90’s by Doug Cutting…. Apache release 2001
  - State of the art Java library for indexing and ranking
  - Wide acceptance by 2005
- LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)
  - Authors: Chih-Chung Chang and Chih-Jen Lin
  - NIPS 2003 feature selection challenge (third place) ….
  - Full SVM implementation in C++ and Java
  - License similar to the Apache License
- LibLinear (http://www.csie.ntu.edu.tw/~cjlin/liblinear/):
  - Machine Learning Group at National Taiwan University
  - Optimized for the linear case (hyperplanes)
  - Same License as LibSVM

Feature Selection and Training

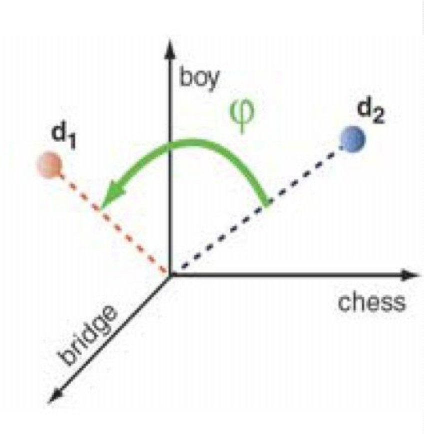
- Training- and Test Documents are stored in a Lucene Index
- Information about topics is stored in a separate untokenized field
- Feature Selection simply consists of comparing posting lists of topics and terms form the text-content
- Consistency of manual topic-assignement can be checked by
  - using MD5-Keys for duplicates checks
  - Lucene's Similarity Search for checking for near duplicates
- Feature vectors are generated from Lucene posting lists
- Training is completely done by LibSVM / LibLinear
- Instead of storing support vectors, hyperplanes are stored directly

Vektor-Space Model for Documents and Queries

Vektor-Space Model:

- Dokument 1: „The boy on the bridge“
- Dokument 2: „The boy plays chess“
- Term / Dokument Matrix:

	Boy	Bridge	Chess	the	on	plays
Document 1	1	1	0	2	1	0
Document 2	1	0	1	2	0	1



Cosinus Similarity:

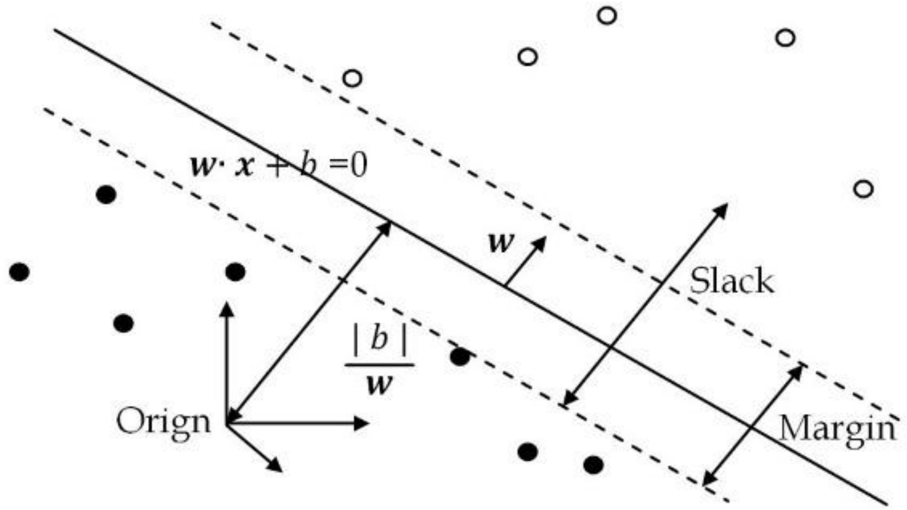
$$\mathrm{Sim}(A, B) = \text{cosine } \theta = \frac{A \bullet B}{|A||B|} = \frac{x1^{*}x2 + y1^{*}y2}{(x1^2 + y1^2)^{1/2} \ (x2^2 + y2^2)^{1/2}}$$

Queries treated as simply very short documents

Fulltext-Search : direct product of query vector with all document vectors

Document-Score: Cosinus-Similarity

Hyperplane Query



Hyperplane Equation: direct product of two vectors minus bias

HyperplaneQuery:

generalized BooleanQuery

no coord, no idf, no queryNorm

- A complete index may be classified by one simple search
- Classifying one document:
  - build a 1-document index
  - apply Classification Queries
- Many topics:
  - Store Queries in Index (Term Boosts as Payloads)
  - Apply Documents as Queries

Questions?

INTRAFIND

Dr. Christoph Goller
Director Research

Phone: +49 89 3090446-0
Fax: +49 89 3090446-29
Email: christoph.goller@intrafind.de
Web: www.intrafind.de

IntraFindSoftware AG
Landsberger Straße 368
80687 München
Germany

www.intrafind.de/jobs