# Seznam.cz Fulltext Architecture

Vladimír Kadlec
vladimir.kadlec@firma.seznam.cz

Seznam.cz a.s.

April 4, 2018

# Seznam.cz, history of a web search

## Directory

- 1996, pages organized in a link directory

## Fulltext

- Kompas
- 2003 – 2005, outsourcing (Empyreum, Google, Jyxo)
- in-house, since 2005 (non-czech web search since 2012)

# Daily average web search stats

| | |
|---|---|
| Users | 2.5 mil. |
| Queries | 14 mil. |
| Queries by humans | 8 mil. |

# Web search

# Web Search Overview

- Downloader/Spider/Crawler/Robot
- Indexer
- Query Processor
- Sorter: Query-Document Relevance

# Downloader/Spider/Crawler/Robot

## Technologies



# Exploration vs Exploitation

# Exploration vs Exploitation

- Prioritize links to be visited
- Download/Parse/Index “useful” pages only

# Problems

## Websites

- Redirects – HTTP 3XX, cycles
- Errors – re-visit again?
- Duplicities – url parameters

## Spider traps

- Dynamic pages: `http://e.com/bar/foo/bar/foo/...`
- Dynamic domains: `aaa.e.com,bbb.e.com,ccc.e.com,...`

# Seznam.cz Crawler Stats

| Number of urls: | |
|---|---|
| Known | 41G |
| Downloaded | 2.4G |
| Parsed (text) | 2.1G |
| Indexed (text) | 1.1G |
| Indexed (img) | 0.7G |
| Indexed (video) | 0.2G |
| Error | 2.1G |
| Redirect | 0.46G |

Stats from 3.4.2018

# Seznam.cz crawler stats (cont.)

## Top level domains, url stats

| | all | .COM | .CZ | .MUNI.CZ |
|---|---|---|---|---|
| Known | 41G | 21G | 10G | 14M |
| Downloaded | 2.4G | 1.0G | 0.9G | 1.4M |
| Parsed | 2.1G | 0.9G | 0.5G | 1.3M |
| Indexed (text) | 1.1G | 0.4G | 0.6G | 0.8M |

# Seznam.cz Crawler, Languages

| Language | Number of documents |
|---|---|
| Czech | 1 044M |
| English | 987M |
| Slovak | 90M |
| German | 43M |
| Other | 266M |
| Total | 2 430M |

## Content predictions

- Primary language
- Images – is the **content type** an image?
- Video – does the page contain a video?
- Porn
- Spam

# Indexer

## Overview

- Tokenization – simple $\times$ complex
- Inverted index – Token $\rightarrow$ Document
- Postings – token positions within a document

# Tokenization

## Simple

- Whitespaces
- Punctuation
- Non-alphanumeric characters

## Complex

- Email, phone number, date, address, url, ...

# Index compression

- Delta compression – sorted lists
  - Original: $1466, 1467, 1469, 1470, \ldots$
  - Compressed $1466, 1, 2, 1, \ldots$
  - Original: $A, B, C, D, \ldots$
  - Compressed:
    $A, (B - A), (C - B), (D - C), \ldots$
  - postings, document identifiers
- Space compression (numbers)
  - small numbers $\rightarrow$ small number of bits

# Seznam.cz indexer hardware

## Finder, Title server

- 211 machines
- 64G RAM, 12 CPU cores (24 with hyperthreading), 300G HDD – why so small capacity?
- two datacenters (Kokura – Seznam, Nagano – O2)

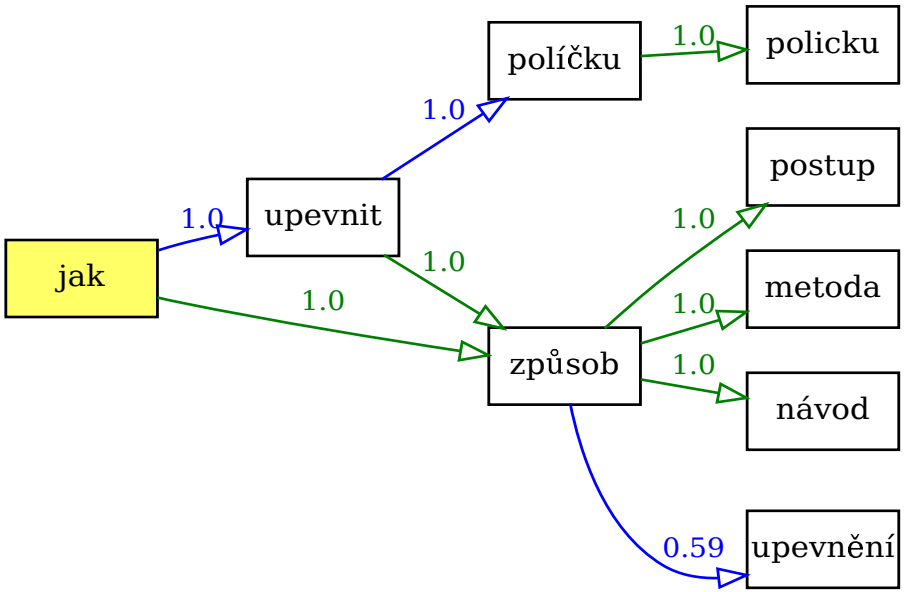
# Query Reformulation

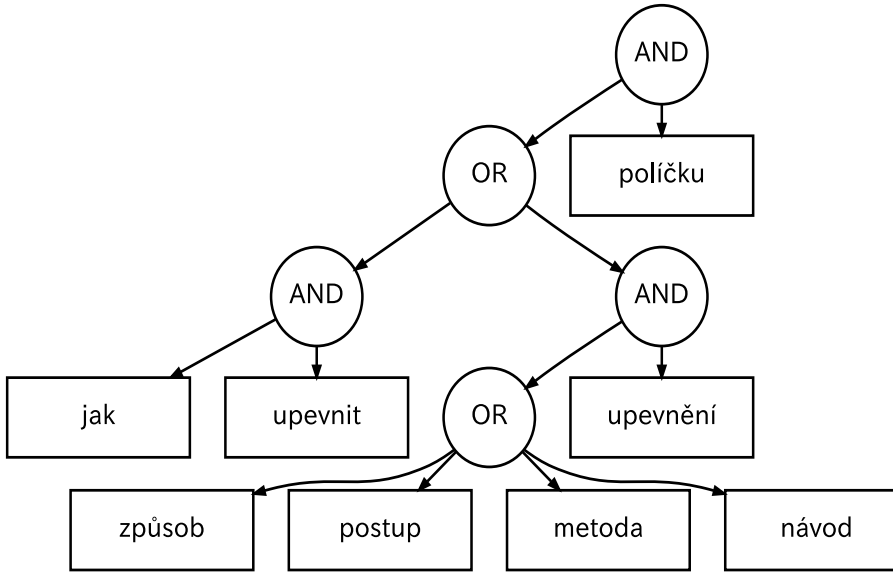
# Query Reformulation

- Expand user query to find more relevant results.

# Seznam.cz search graph

## jak upevnit policku



# Query Language Detection

فيينا

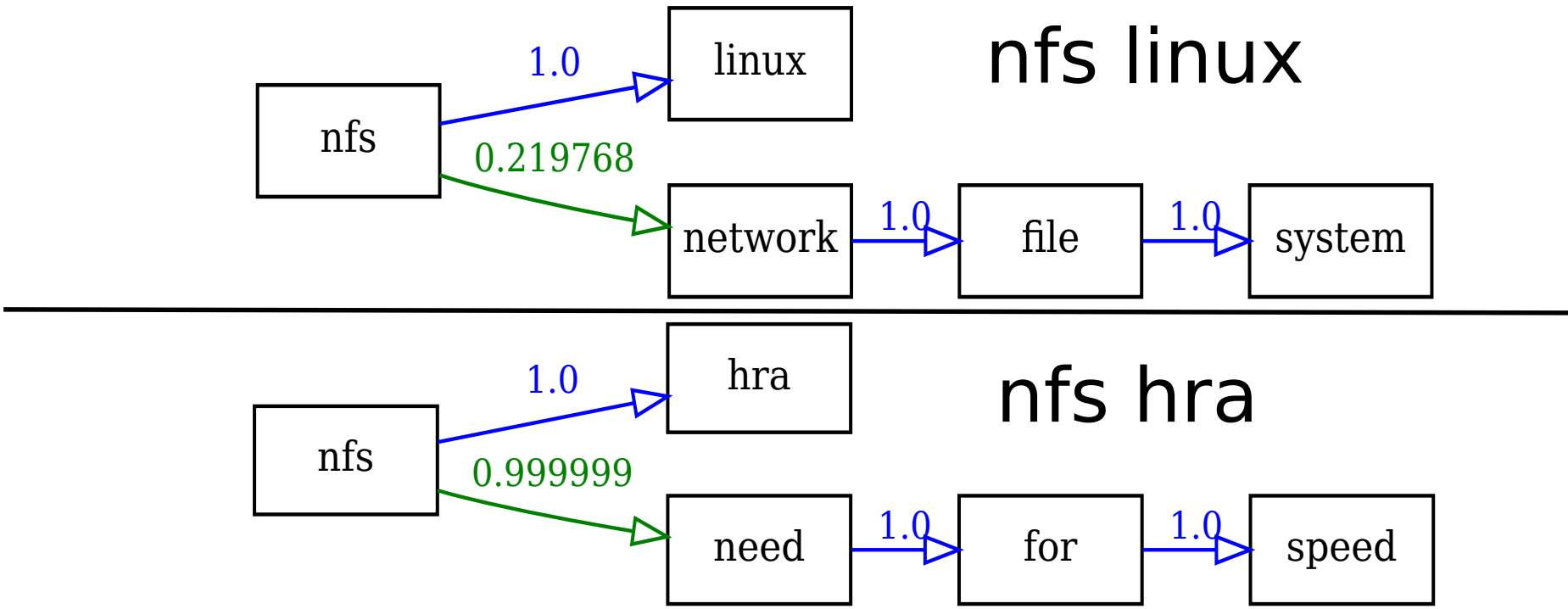
Wien

Vienna

Vídeň

Вена

# Acronym Expansion



# Query Reformulation

- Diacritics reconstruction
- Stopwords
- Entities identification
- Other terms related to the query
- Porn detection

# Query-Document Relevance

## Machine learning

- Learning to rank
- Boosted Random Forest
- Oblivious Decision Trees

## Features

- Off-page – PageRank, popularity, backlinks, ...
- Text relevance scoring – Tf-idf, ...

# Query-Document Relevance (cont.)

## Evaluation

- Manual annotations, discounted cumulative gain
- A/B testing

# User Interface

# Image/Video Search

## Image Search

- History: outsourcing, picSearch
- in-house since October 2015

## Video Search

- History: outsourcing, Yandex
- in-house since October 2015

# Image/Video Search

## Text Based Search

- Anchors
- Titles (page, image, video)
- Image content analysis (ResNet50 + Czech word2vec)

# Image Classification

## Technologies

- Transfer learning
- Caffe, ResNet50, InceptionV3, ...
- Keras, TensorFlow

## Applications

- Car classification for Sauto.cz
- Photo analysis for Sreality.cz
- Porn classification

# Sauto.cz, Photo Analysis






vs

- Better look & feel
- Classification (22 categories): exterior_front_left, exterior_wheels, interior_dashboard, ...

# Recap

- Downloader/Spider/Crawler/Robot
- Indexer
- Query processor
- Sorter: Query-Document relevance
- Image/Video search

# Conclusion

## Meet Me

- Machine Learning Meetups (Brno, Praha), `http://www.mlmu.cz`
- DataPivo Brno, `https://www.meetup.com/dataheads`
- Seznam Research, `https://kariera.seznam.cz/`

## Contact


- `vladimir.kadlec@firma.seznam.cz`