Seznam.cz Fulltext Architecture

Vladimír Kadlec
vladimir.kadlec@firma.seznam.cz

Seznam.cz a.s.

December 2, 2017

Daily web search stats

Stats from 22.2.2016 (Monday):

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

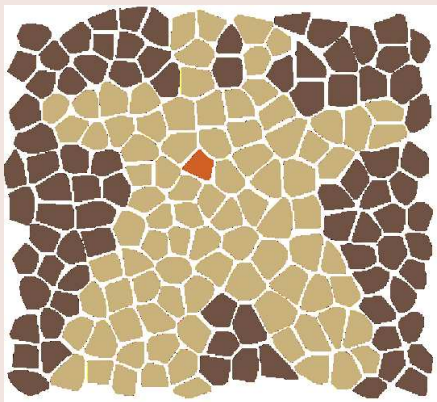
Web search

Web search overview

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

Downloader/Spider/Crawler/Robot

Technologies



APACHE GIRAPH

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	22G
Downloaded	2.4G
Parsed	1.4G
Indexed	0.7G
Error	2.1G
Redirect	0.35G

Stats from 22.2.2016

Seznam.cz crawler stats (cont.)

Top level domains, url stats

	all	.COM	.CZ	.MUNI.CZ
Known	22G	12G	5.3G	4748K
Downloaded	2.4G	1.1G	0.7G	474K
Parsed	1.4G	0.6G	0.5G	320K
Indexed	0.7G	0.2G	0.3G	200K

Seznam.cz crawler, languages

Language	Number of documents
Czech	585M
English	673M
Slovak	77M
German	30M
Other	65M
Total	1 430M

Seznam.cz corpora, languages

Data from 2012:

Langguage	#docs $\times 10^6$	#tokens $\times 10^9$
English	105	59
Czech	94	24
German	6.7	2.6
Slovak	6.1	2.1
French	2.5	1.4

Stats from the corpus after deduplication/cleaning.

Data from 2013: all – 208 $\times 10^9$ tokens (all), czech – 26 $\times 10^9$ tokens.

Indexer

Indexer

Overview

- Tokenization – simple $\times$ complex
- Dictionary – 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
  - $A, B, C, D, \ldots \rightarrow$
    $A, (B - A), (C - B), (D - C), \ldots$
  - postings, document identifiers
- Space compression (numbers)
  - small numbers $\rightarrow$ small number of bits

Seznam.cz indexer hardware

Finder, Title server

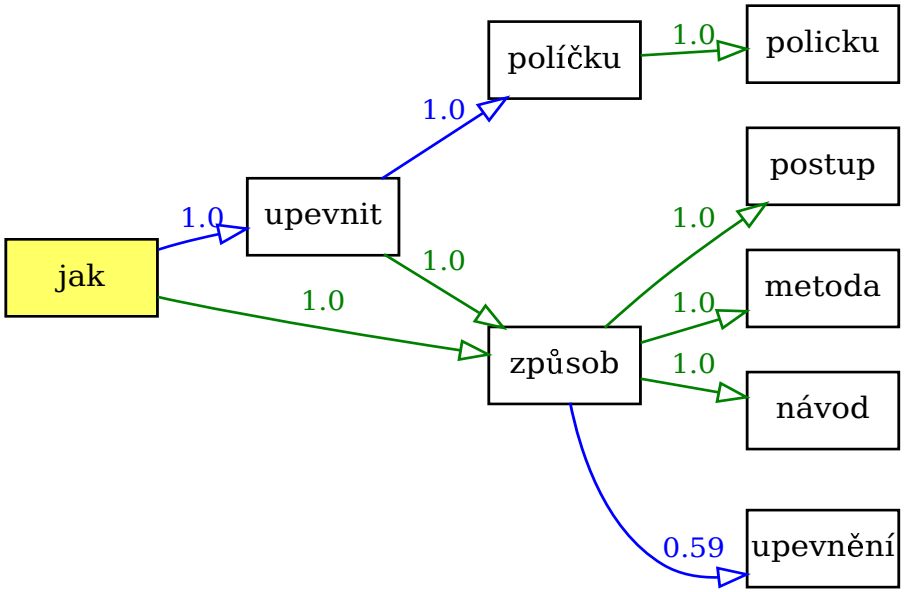
- 168 machines
- 32G RAM, 12 CPU cores (24 with hyperthreading), 300G HDD – why so small capacity?

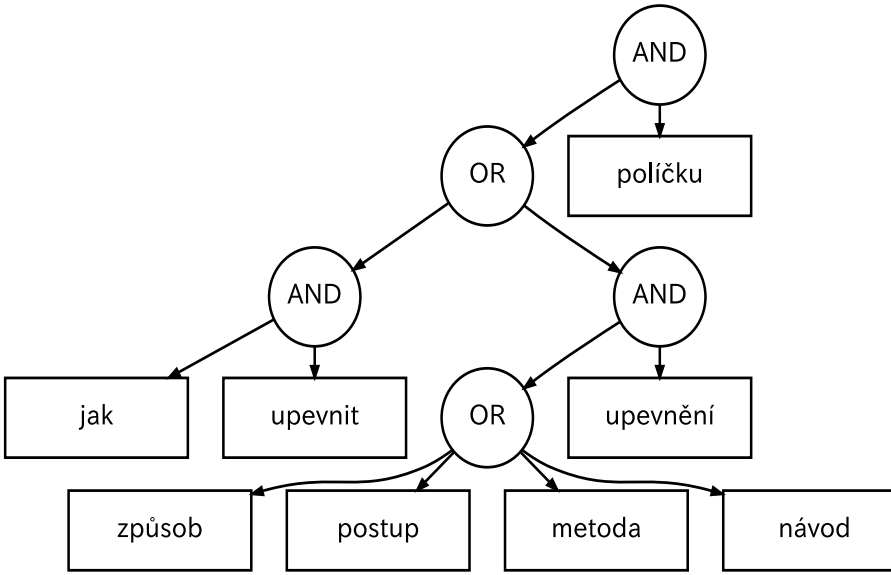
Query Reformulation

- Expand user query to find more relevant results.

Seznam.cz search graph

jak upevnit policku





Query Language Detection

فيينا

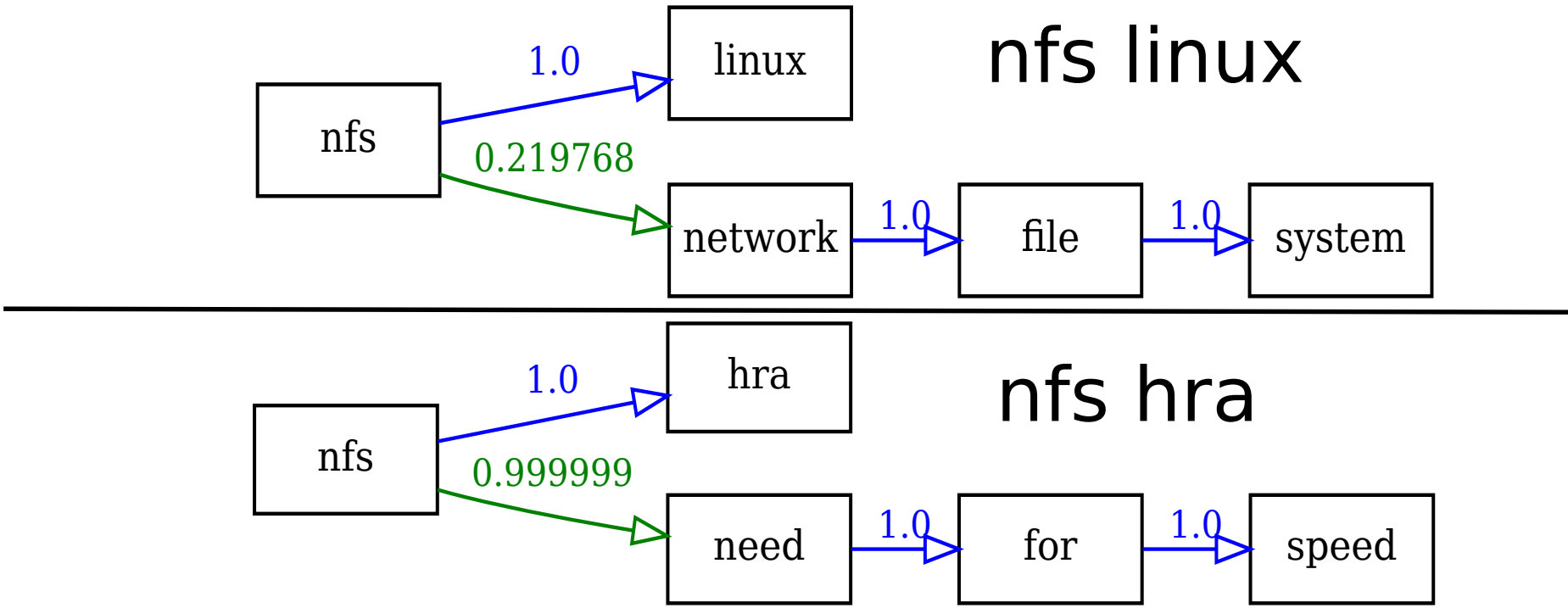
Wien

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

- Boosted Random Forest
- Oblivious Decision Trees

Evaluation

- Manual annotations
- A/B testing

User experience

Image/Video search

Image search

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

Video search

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

Image/Video search

Text based search

- Anchors
- Titles (page, image, video)
- Image content analysis

Image classification

Machine learning

- Car classification for sAuto.cz
- Photo analysis for sReality.cz
- Porn classification

Recap

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

Conclusion

Fulltext blog

- `http://fulltext.sblog.cz`

We're hiring

- `http://seznam.sprace.cz`
- Výzkumník ve vyhledávání

Contact

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