Jak se vyvíjí fulltext

Jakub Černý, Ph.D.

MFF Praha, 31.3.2010

Co dnes servírujeme?

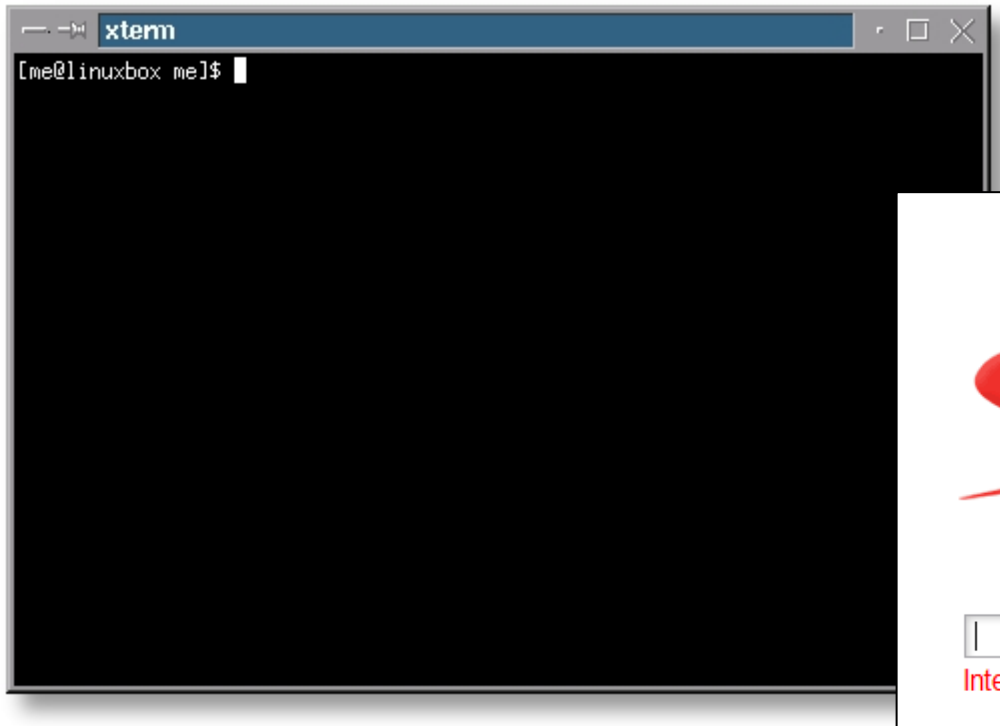
- Jak funguje fulltext?
- Jak funguje textový signál relevance?
- Jak měřit kvalitu fulltextu?
- Jak se srovnávat s konkurencí?
  Jak nastavovat parametry algoritmu hledání?
- Co se využije při vývoji fulltextu?
- Bonus (Technické parametry a statistiky)

Co byste chtěli slyšet vy?

Jak tečou uživatelé internetem?

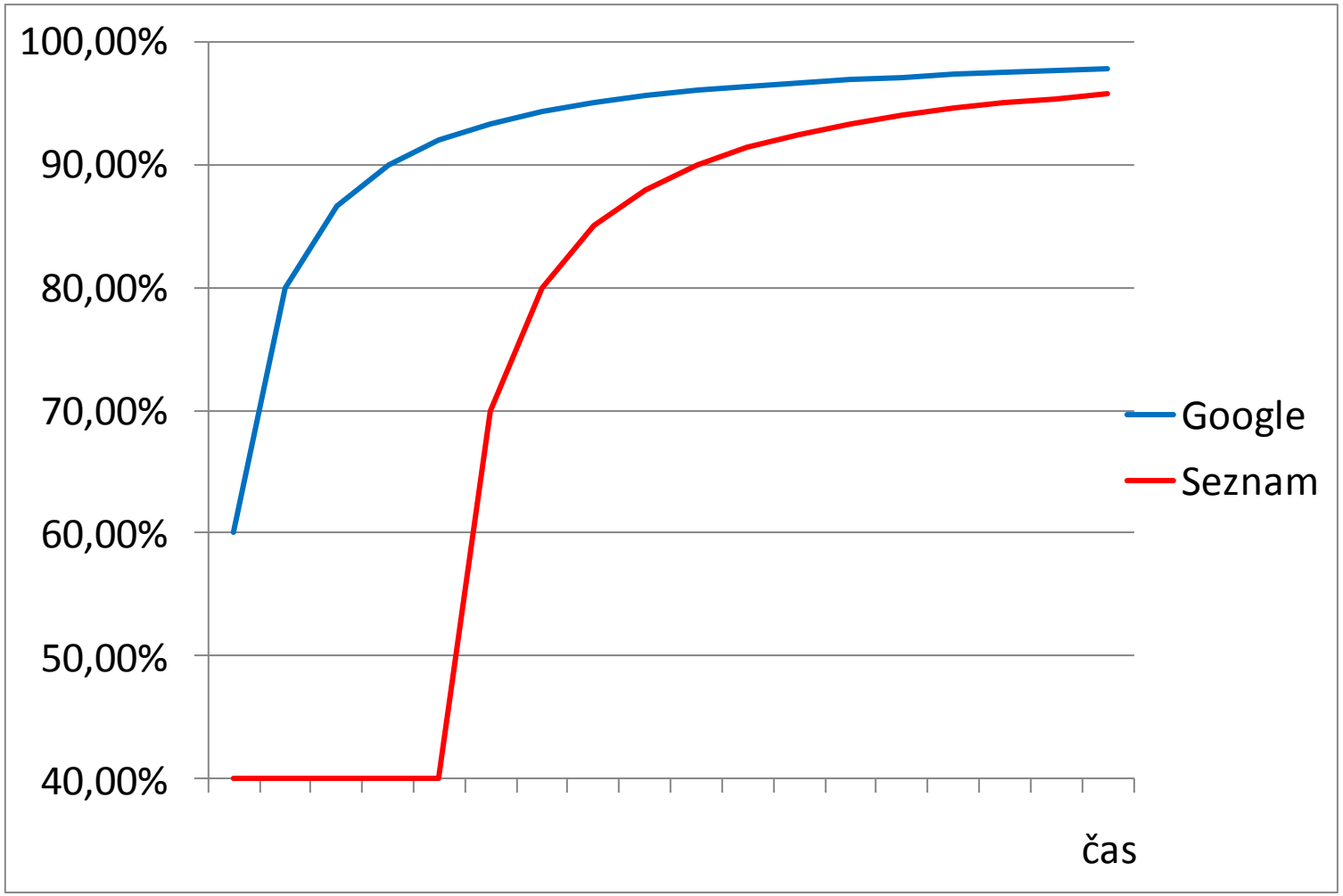
- Internet a odkazy jsou jako dálnice/sjezdovka
  - co dělá běžný uživatel z pohledu mimozemšťana?
- Kde každý začíná?
  - homepage, fulltext, znám adresu
- Máte webový portál, kde sehnat návštěvníky?
  - postavit lepší přípojku z dálnice (SEO)
  - reklama

Znovu objevení kola



Do roka to bude řádka s URL v prohlížeči.

Seznam vs. Google



Proč Seznam vydrží?

Jak funguje fulltext?

Základní myšlenka

- Analogie s knihou
- Jak zjistíte, na které stránce se nachází „fulltext“?

Inverted list (index)

Doc1

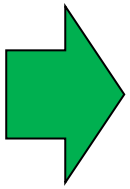
Město
Praha
bylo
založeno
….

Doc2

Každé
město
má
….

Doc3

Žiju
v
Praze.
Je to
krásné
město.



Inverted list

být	→ Doc1[3], Doc3[4]
každý	→ Doc2[1],
krásný	→ Doc3[6],
město	→ Doc1[1], Doc2[2], Doc3[7]
mít	→ Doc2[3],
Praha	→ Doc1[2], Doc3[3]
to	→ Doc3[5],
v	→ Doc3[2],
založit	→ Doc1[4],
žít	→ Doc3[1],

Hledání v indexu

Hledám dotaz „město Praha“

Inverted list

být	→ Doc1[3], Doc3[4]
každý	→ Doc2[1],
krásný	→ Doc3[6],
město	→ Doc1[1], Doc2[2], Doc3[7]
mít	→ Doc2[3],
Praha	→ Doc1[2], Doc3[3]
to	→ Doc3[5],
v	→ Doc3[2],
založit	→ Doc1[4],
žít	→ Doc3[1],

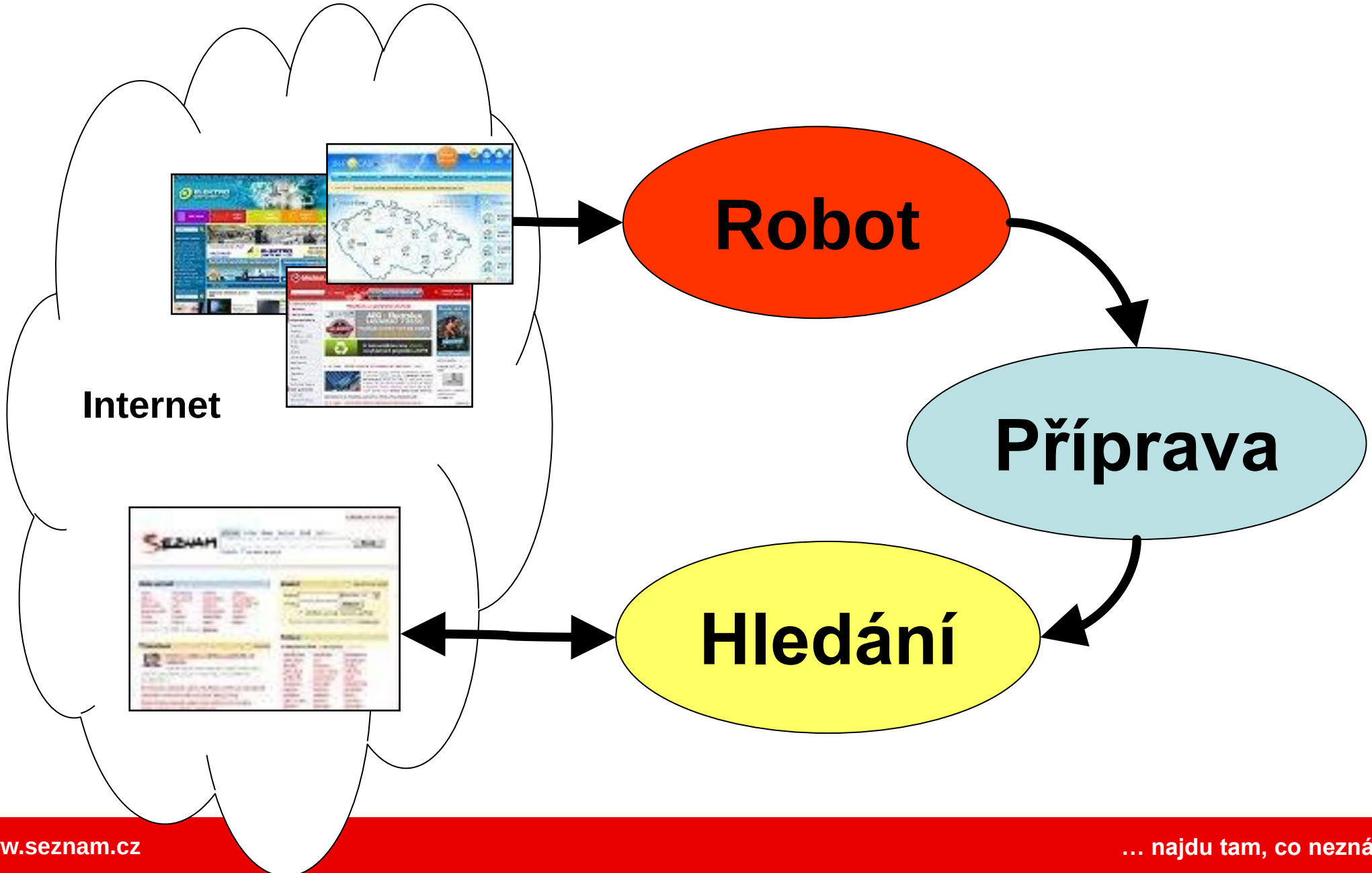
Nalezení řetízku pro slova z dotazu

a vyhodnocení ala merge.

Výsledky hledání

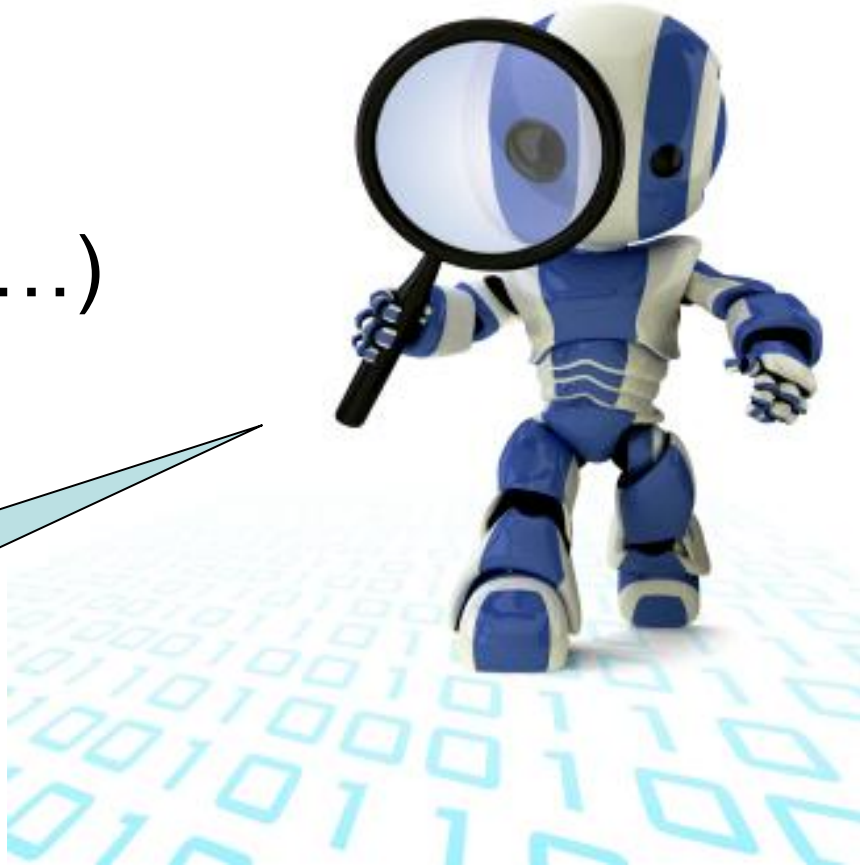
Doc1
Doc3

Jak funguje Fulltext



Robot

- **Úkol:** procházet web, hledat nové dokumenty a obnovovat současné
- Detekce jazyka
- Hledáme jen české stránky
- Další formáty (pdf, doc, rtf, ppt,…)



Jak komunikovat s robotem

- Robots.txt – standardní protokol pro zakázání přístupu robotů (www.robotstxt.org)
  http://example.com/robots.txt
- Sitemap.xml
  http://example.com/sitemap.xml

```
# comment
User-Agent: *
Disallow: /statistiky

User-Agent: Bot
Disallow: /
```

```
… <url>
        <loc>http://www.example.com/</loc>
        <lastmod>2007-10-30T16:31:04+00:00</lastmod>
        <changefreq>daily</changefreq>
        <priority>1.0</priority>
    </url> …
```

Zvládání zátěže

- Stíháme včas odpovídat 1 milionu uživatelů. Co když chceme uspokojit celou ČR? (5mil uživatelů)

- Jak zajistit dostupnost? Tj. aby nám nevadil výpadek jednoho stroje.

Zvládání zátěže

Internet

Robot

Příprava

Hledání

Hledání

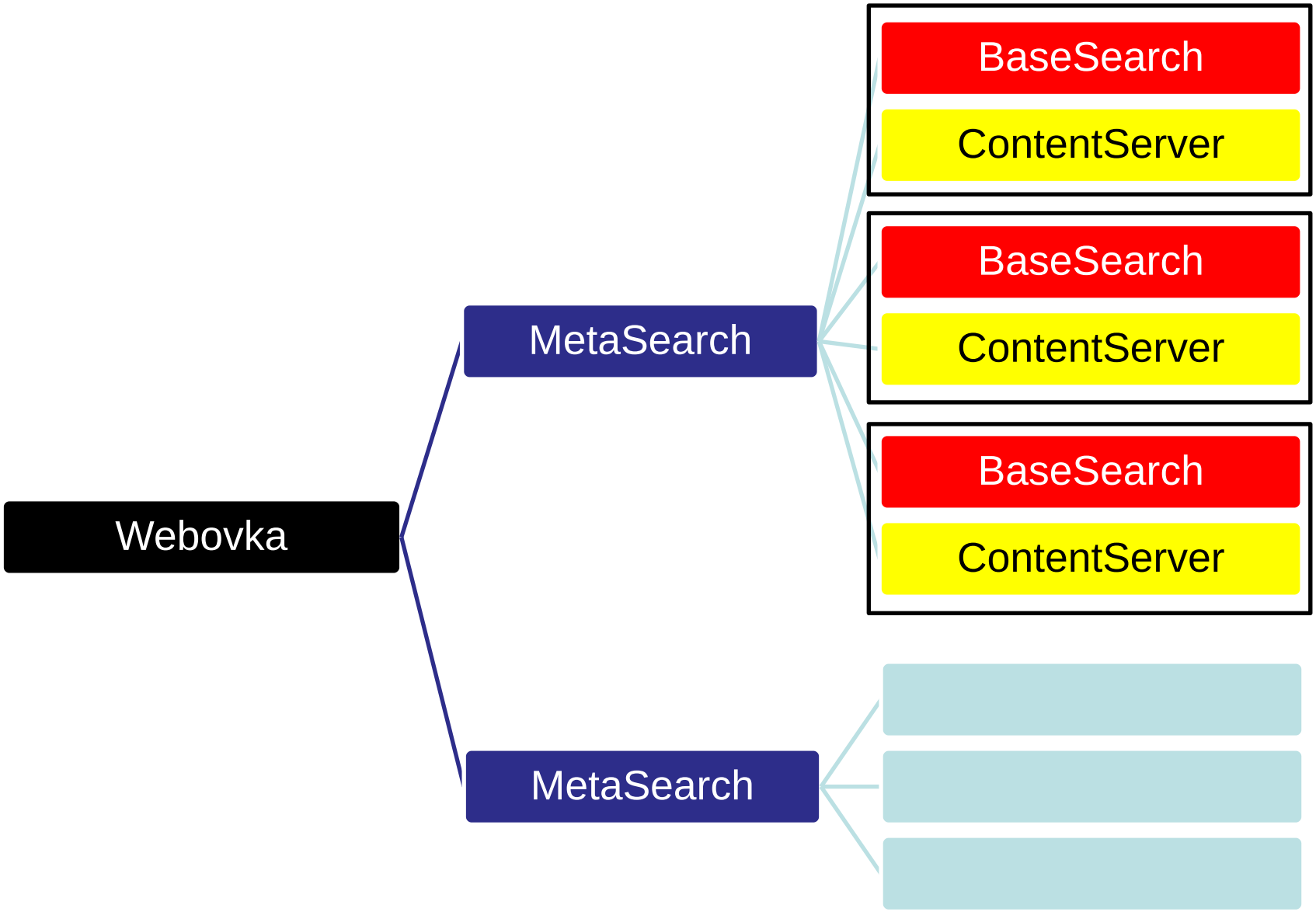
2 serverovny

LVS

Jak zrychlit výdej?

- Disky jsou pomalé. Vše musí být v cache.
- Co s tím?

Jak zrychlit výdej?



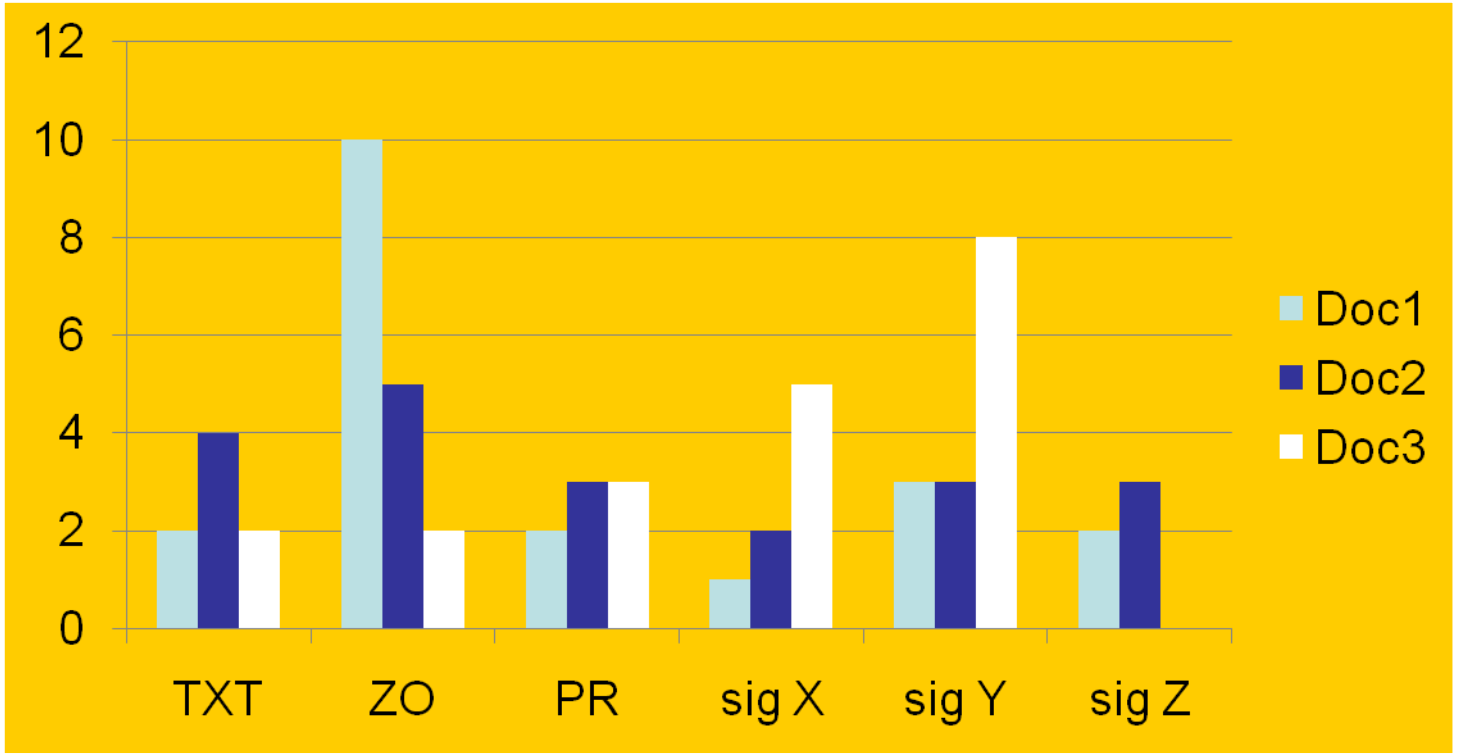
Jak se mixují signály relevance?

Signály relevance

	On page	Off page	User
obecné	Doména, historie, struktura stránky	Page Rank	???
tématické (k dotazu)	TXT	Zpětné odkazy	???

Pořadí výsledků

Mixování signálů relevance:



Kdo je lepší? Jak to míchat?

Generace mixování signálů

- 1. generace

  $$Relevance = \sum w_i \cdot S_i$$

- 2. generace

  $$Relevance = \prod (S_i + w_i)$$

- 3. generace

Tajné

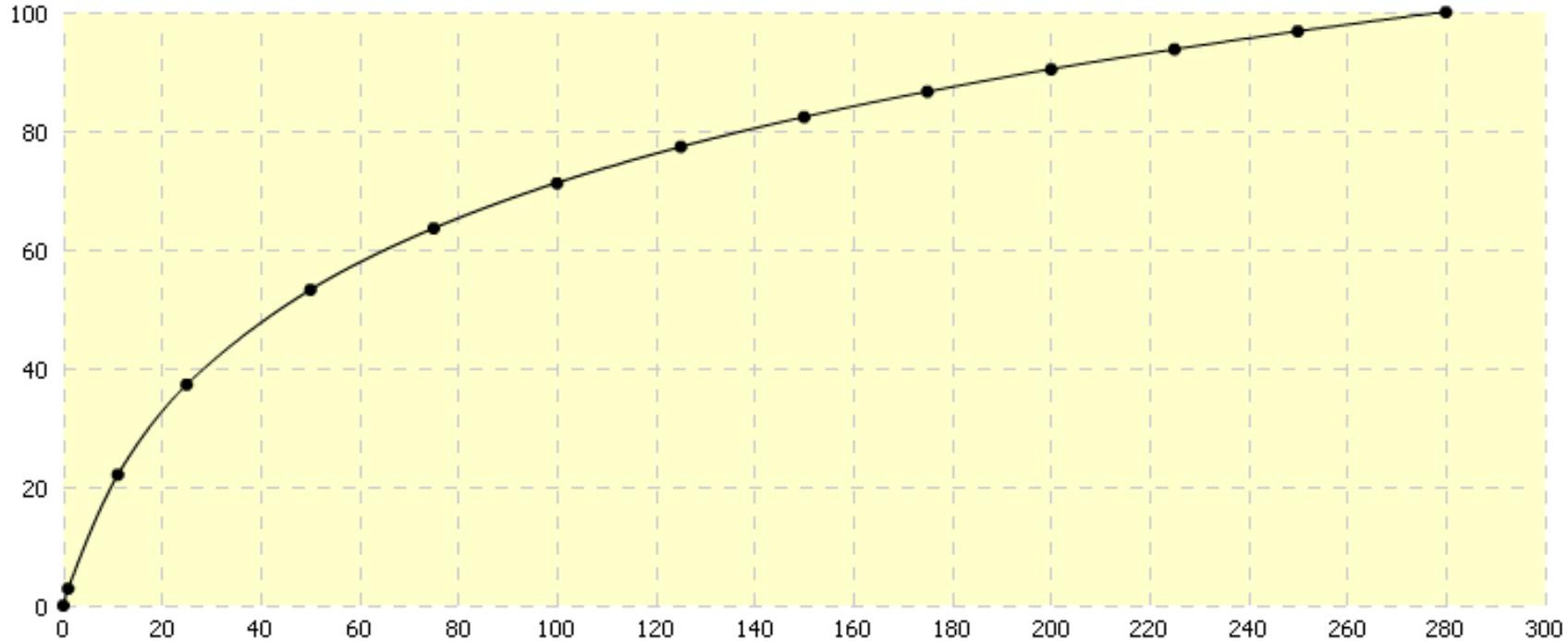
- další generace?

Ditribuční funkce relevance

Výběr dist.funkce

Typ: bsWeigh[redacted]Function

Varianta: default



0.000	25.000	50.000	75.000	100.000	125.000	150.000	175.000	200.000	225.000	250.000	280.000
0.000	37.200	53.210	63.550	71.210	77.290	82.340	86.650	90.410	93.750	96.760	100.000

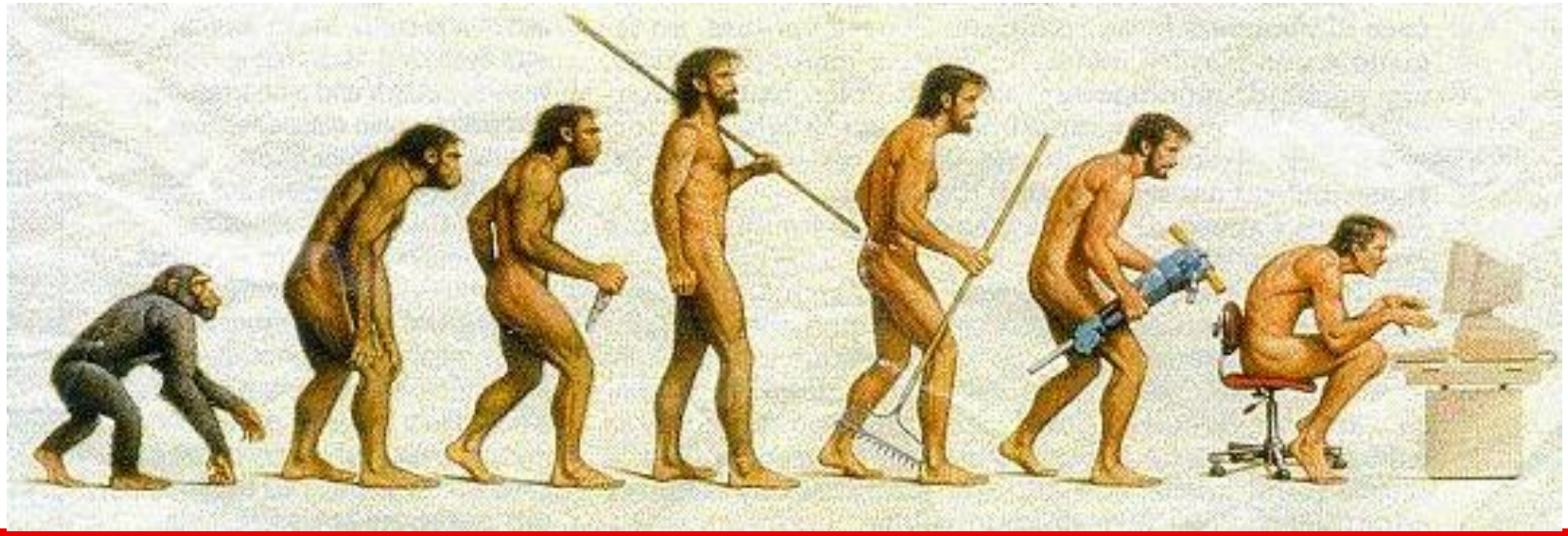
1.000
2.830
11.000
22.030

Ukazovat hodnotu | Vykreslovat tečny | default | Uložit

Textový signál relevance

Textový signál relevance

- Je to názorná ukázka evoluce 1 signálu
- …jak probíhá výzkum
- Uslyšíte, jak funguje hledání v textech
  (to můžete na vašich stránkách ovlivnit)



Vývojové generace TXT signálu

- Jen slova z dotazu, přesná shoda tvaru
  - Jen 50% relevantních dokumentů obsahuje slova z dotazu.

Příklad: Dotaz „ČNB“, ale relevantní stránka obsahuje jen „oficiální úroková míra v České národní bance“.

Vývojové generace TXT signálu

- Přidání lemmatizace slov
- Různé váhy slov podle výskytiště (H1, URL, Title, odstavec, bold, …)

- Příklady vtipné lematizace:
  - Stát, ženu, lov lína, barum, jizdní rady, dog

Vývojové generace TXT signálu

- Různé váhy slov podle jejich korpusové četnosti
  - tf x idf
  - vynechávání slov

Příklad dotazů: Petr a Pavel, Jak se odstraňuje vosí hnízdo?

Otázka pro vás:

3-slovné dotazy: Máme zvýhodňovat výsledky, kde se slova z dotazu najdou blíže u sebe? Nebo je to jedno?

Vývojové generace TXT signálu

- Proximita a pořadí slov z dotazu

- Příklady:
  - Jakub Černý x Černý Jakub
  - Václav Klaus video
  - Já do lesa nepojedu, já do lesa nepůjdu
- Kolokace
  - Velký vůz, černý Petr, Česká republika

Vývojové generace TXT signálu

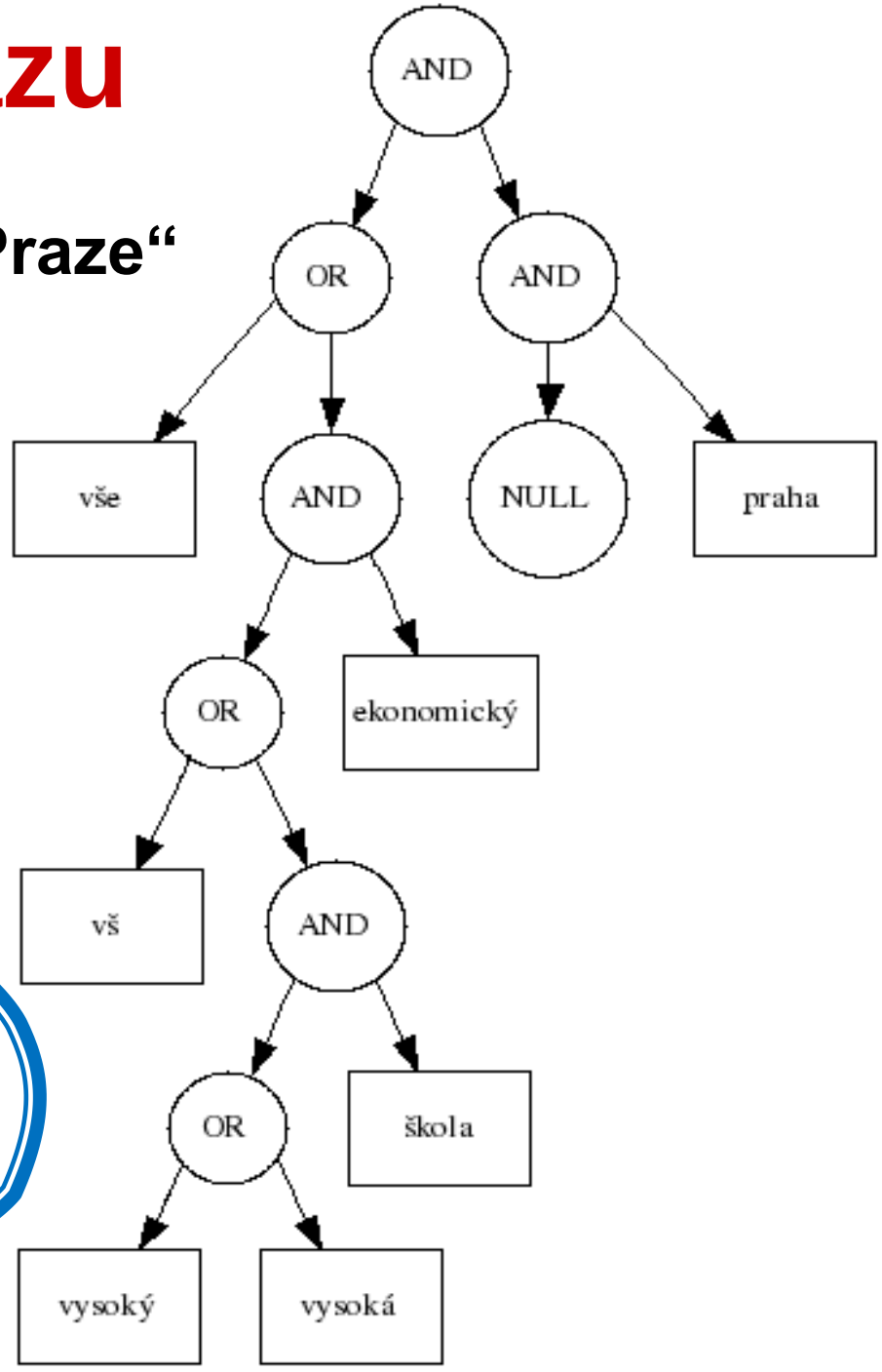
- Předzpracování dotazu
  - Poslechnu si uživatele a přeložím to do jazyka, ve kterém fulltext umí vyhledávat.
  - Nastavení proximity, …

- Příklady:
  - VŠE, MŽP, IE8 (ale i naopak)
  - Kdy vyhořelo Národní divadlo?
  - (běžné otázky jako na kamaráda)

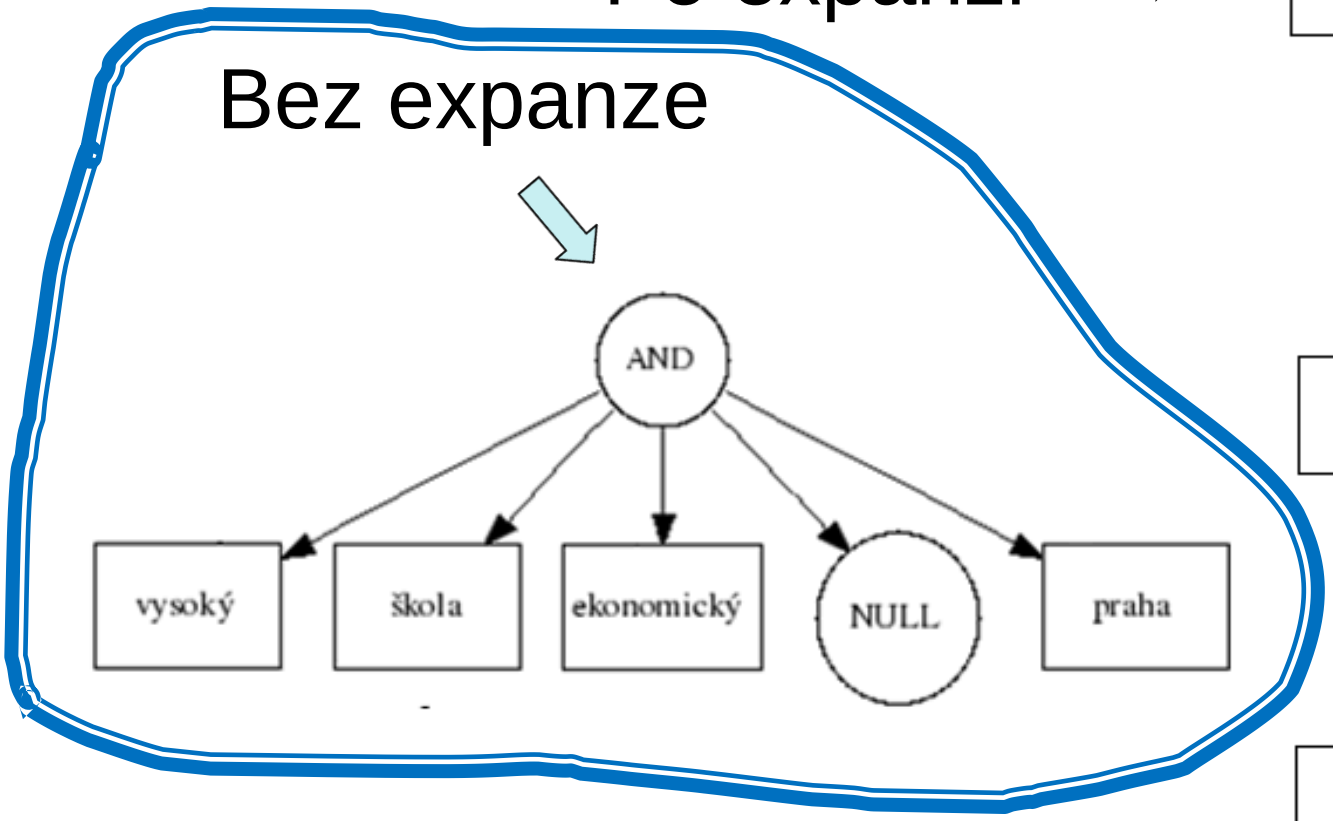
Expanze dotazu

Dotaz „Vysoká škola ekonomická v Praze“

Po expanzi



Bez expanze



Vývojové generace TXT signálu

- Doplňování slov odjinud
  - ze zpětných odkazů (bazén podolí)
  - anonymní termy
    - jméno, datum, místo, video
    - pro odpovědi na otázky: Kdo? Kdy? Kde?

- Příklady:
  - Václav Klaus video
  - Kdy vyhořelo Národní divadlo?

Další okolnosti kolem TXT signálu

- Body text extraction (BTE)
- Site-wide texty (SWT)
  - rozpoznání důležitosti slov podle vzhledu site
  - odstranění neopodstatněných nároků na důležitost
    - Všechny texty v H1 apod.

- Různé chování pro různé kategorie dotazů:
  - Navigační
  - Informační
  - Transakční

Další okolnosti kolem TXT signálu

- Desambigulace
  - Vyloučení nejednoznačnosti

  - Řekněte mi něco o německých tancích?
  - Hrách vs. (o počítačových) hrách

Jak měřit úspěch?

Proč? Co chceme?

- Měření kvality vyhledávačů
- Srovnání Seznamu s konkurencí
  - Kdo je lepší?
  - Na kterých kategoriích?
  - Na kterých dotazech?
  - Jak popsat skupinu dotazů, kde se to děje?
- Dostaneme tip, co zlepšovat
- Měřitelnost toho, jak jsme se zlepšili (SMART)

Otázka pro vás:

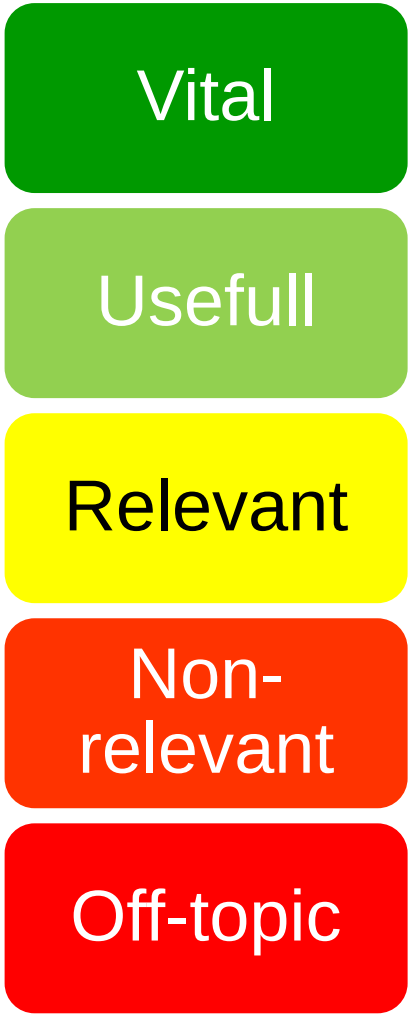
Jak měřit kvalitu výsledků fulltextového hledání?

- Čistě pořadí výsledků, ne rychlost hledání, či kvalitu webovky, snippetů

Kalibrace

- Vital
- Usefull
- Relevant
- Non-relevant
- Off-topic

Kalibrace



(navigační výsledek) Dotaz má jasnou interpretaci a stránka je oficiální stránkou (jedinečnost). q=youtube … youtube.cz

(užitečný výsledek) Stránka je hodně uspokojující, vyčerpávající výklad, vysoká kvalita,důvěryhodný zdroj. q=houby … atlashub.cz

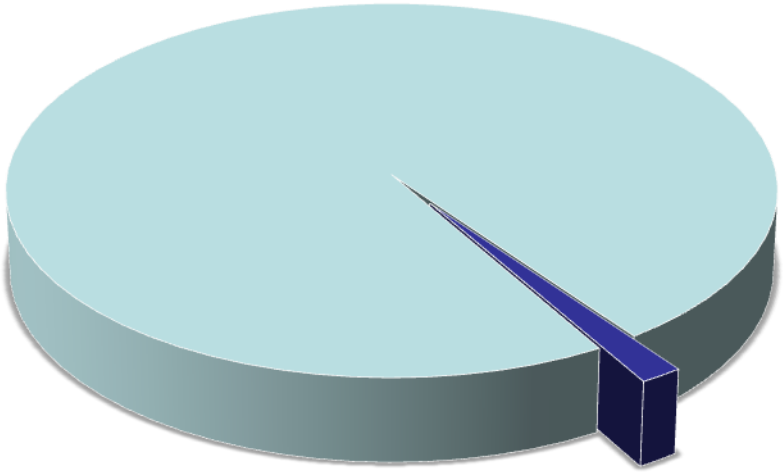
(dobrý výsledek) q=harry potter … knihy.cz/prodej/harry-potter

(blbý výsledek) Sice je to k tématu, ale není užitečné (málo informací, staré info, příliš obecné). q=praha … zoopraha.cz

(výsledek mimo mísu) Výsledek obsahuje hledaná slova, ale tématicky je mimo. q=houby … „je to na houby“

Kalibrace

Výběr dotazů



Sociodemo kalibrátorů

- Porozumění dotazu
- Kvalifikace pro zhodnocení kvality
- Muži vs. ženy (fotbal x parfémy)
- Puberťáci vs. důchodci (q=hudba)

Kvalita dotazu: bazén podolí

Tajné

Výpis výsledků

	Pořadí	URL	Dotaz	Box	Akce
☐	1	http://www.pspodoli.cz	bazén podolí	vital	
☐	2	http://www.pspodoli.cz/zarizeni.htm	bazén podolí	useful	
☐	3	http://www.bazenpodoli.cz/bazeny-podoli	bazén podolí	useful	
☐	4	http://cs.wikipedia.org/wiki/Plaveck%C3%BD_stadion_Podol%C3%AD	bazén podolí	useful	
☐	5	http://expedice.rps.cz/lokality/12388-plavecky-stadion-podoli-bazen.html	bazén podolí	relevant	
☐	6	http://zuzikwww.blog.cz/0904/jeste-krasnejsi-nez-bazen-v-praze-4-podoli	bazén podolí	relevant	
☐	7	http://www.nelso.cz/cz/place/8597	bazén podolí	relevant	
☐	8	http://www.pragueout.cz/sport/bazeny/plaveckystadionpodoli	bazén podolí	relevant	
☐	9	http://www.vitalia.cz/katalog/bazeny/plavecky-stadion-podoli-cstv	bazén podolí	relevant	
☐	10	http://naturista.cz/drupal/?q=lokality/praha_podoli	bazén podolí	relevant	
☐	11	http://www.zaket.cz/8x4p_mista.php?akce=9	bazén podolí	relevant	
☐	12	http://sechtl-vosecek.ucw.cz/en/cml/35mm/film35mm1516.html	bazén podolí	non-relevant	
☐	13	http://6rbtata.com/view/hRhkHC2Lt_0/Hu%C4%8D%C3%ADnovi_-_baz%C3	bazén podolí	non-relevant	

Srovnání výsledků fulltextů

Dotaz: avon

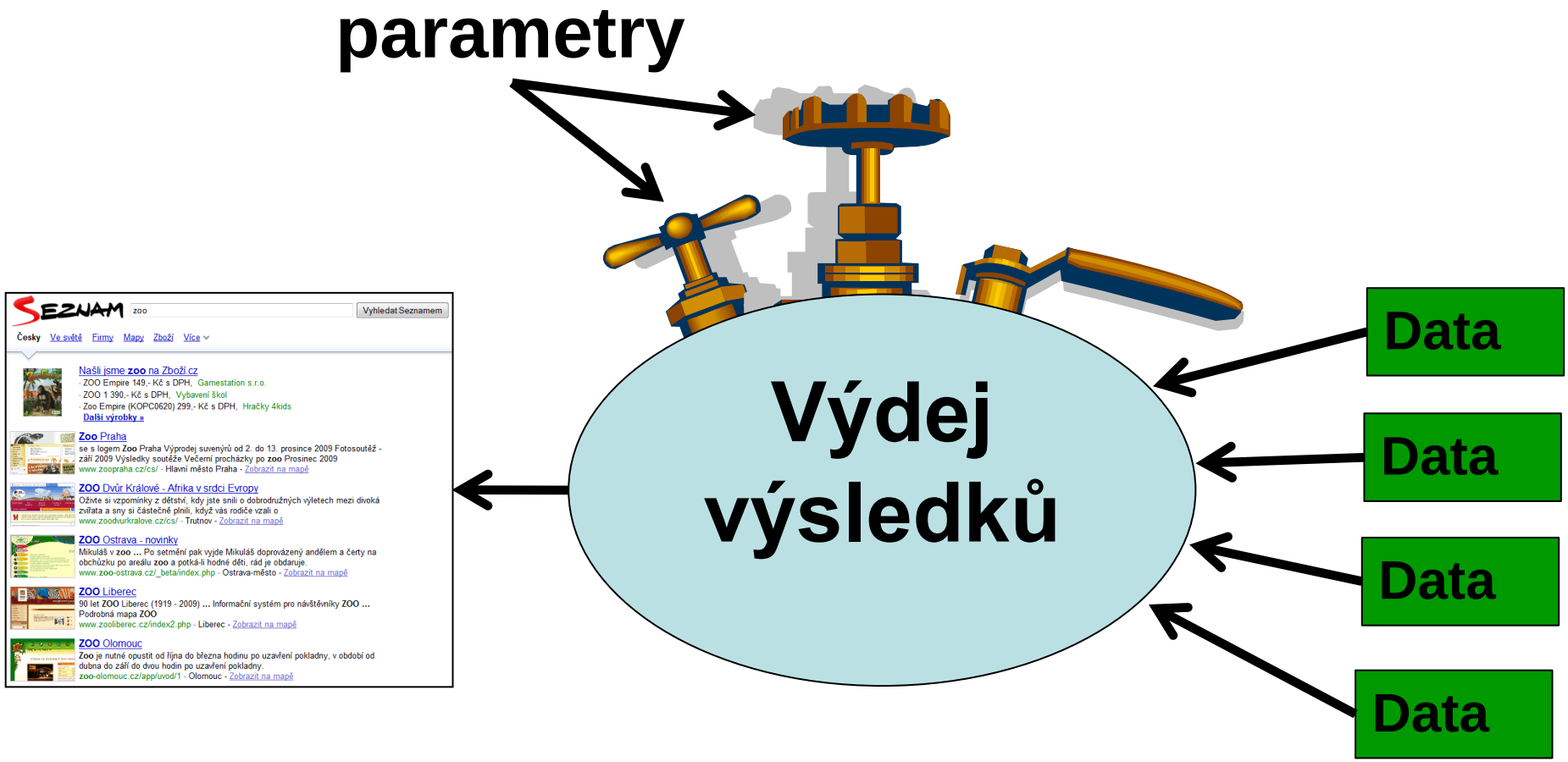
Pořadí	google	seznam	live	seznam test
Kvalita / Spolehlivost:	81.4% / 68.2%	84.3% / 91.8%	42.1% / 82.1%	84.3% / 91.8%
1	www.avoncosmetics.cz	www.avon-kosmetika.cz	www.avon.com	www.avon-kosmetika.cz
2	www.avoncosmetics.cz	www.avoncosmetics.cz	www.avoncosmetics.cz	www.avoncosmetics.cz
3	www.avon.cz	www.kosmetika-avon.cz	www.avon.com.au	www.kosmetika-avon.cz
4	www.avon-plus.cz	www.avon.cz	www.avon.ca	www.avon.cz
5	www.krasa.cz	www.avon-eshop.com	www.avon.cz	www.avon-eshop.com
6	www.krasa.cz	www.avon-kosmetika.eu	www.avon.org	www.avon-kosmetika.eu
7	www.avon-kosmetika.cz	www.avon-plus.cz	www.avon-plus.cz	www.avon-plus.cz
8	zena.centrum.cz	www.online-avon.cz	www.ar.avon.com	www.online-avon.cz
9	zena.centrum.cz	www.vuneprotebe.cz	www.pl.avon.com	www.vuneprotebe.cz
10	www.zdravaprsa.cz	www.avon-styl.cz	www.avon.ru	www.avon-styl.cz
11	www.avon-online.sk	www.avonlady-online.com	www.avon.co.nz	www.avonlady-online.com
12	vltava2000.cz	www.avon-eshop.eu	www.br.avon.com	www.avon-eshop.eu
13	www.firmy.cz	www.muj-avon.cz	www.avon.bg	www.muj-avon.cz
14	cs.wikipedia.org	www.avonland.cz	www.avon.it	www.avonland.cz
15	www.mammahelp.cz	www.kosmetika-avon.biz	www.avon.fi	www.kosmetika-avon.biz
16	www.estav.cz	www.krasa.cz	www.avon-kosmetika.cz	www.krasa.cz
17	www.gemoney.cz	www.avon-relax-centrum.com	www.avon.com.tr	www.avon-relax-centrum.com
18	tn.nova.cz	www.krasnadama.cz	www.avon.gen.tr	www.krasnadama.cz
19	avon.heureka.cz	www.avon-centrum.cz	www.avon.lt	www.avon-centrum.cz
20	www.lekarna.cz	avon-land.euweb.cz	www.cl.avon.com	avon-land.euweb.cz

Přínosy

- Možnost automatického nastavování parametrů fulltextu
- Rozhodování se na základě reálných dat
- Rychlejší vývoj a testování změn relevance fulltextu (prototypy úprav).
- Přenesení práce na externí kalibrátory
- Bonzování, co jsou nepovedené dotazy a jejich následné sledování -- víme na co se zaměřit
- Včas zjistíme, jak se zlepšila konkurence, co provedli -- můžeme je včas dohnat

Automatické ladění parametrů fulltextu

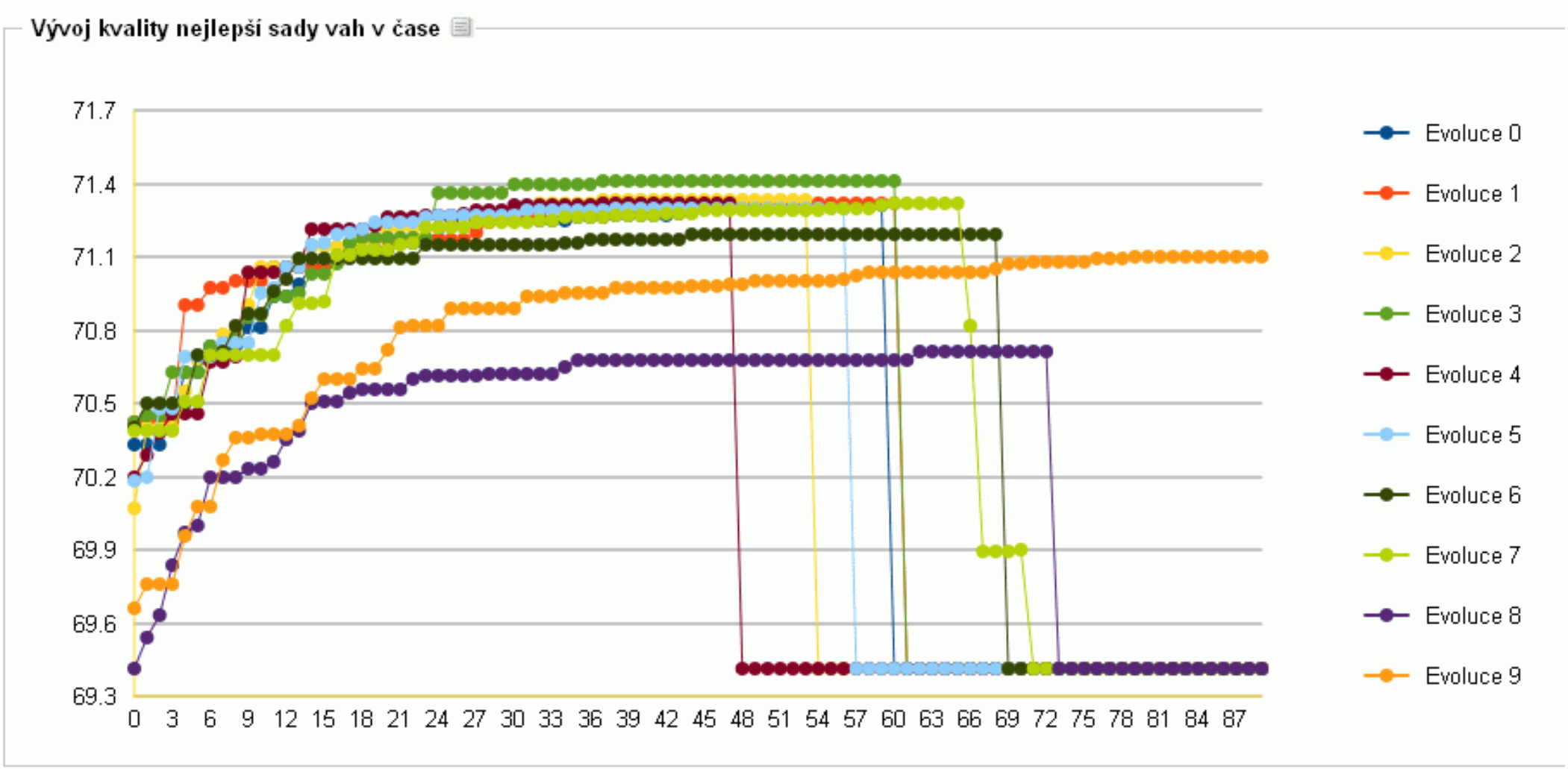
Jak nastavit parametry na optimum?



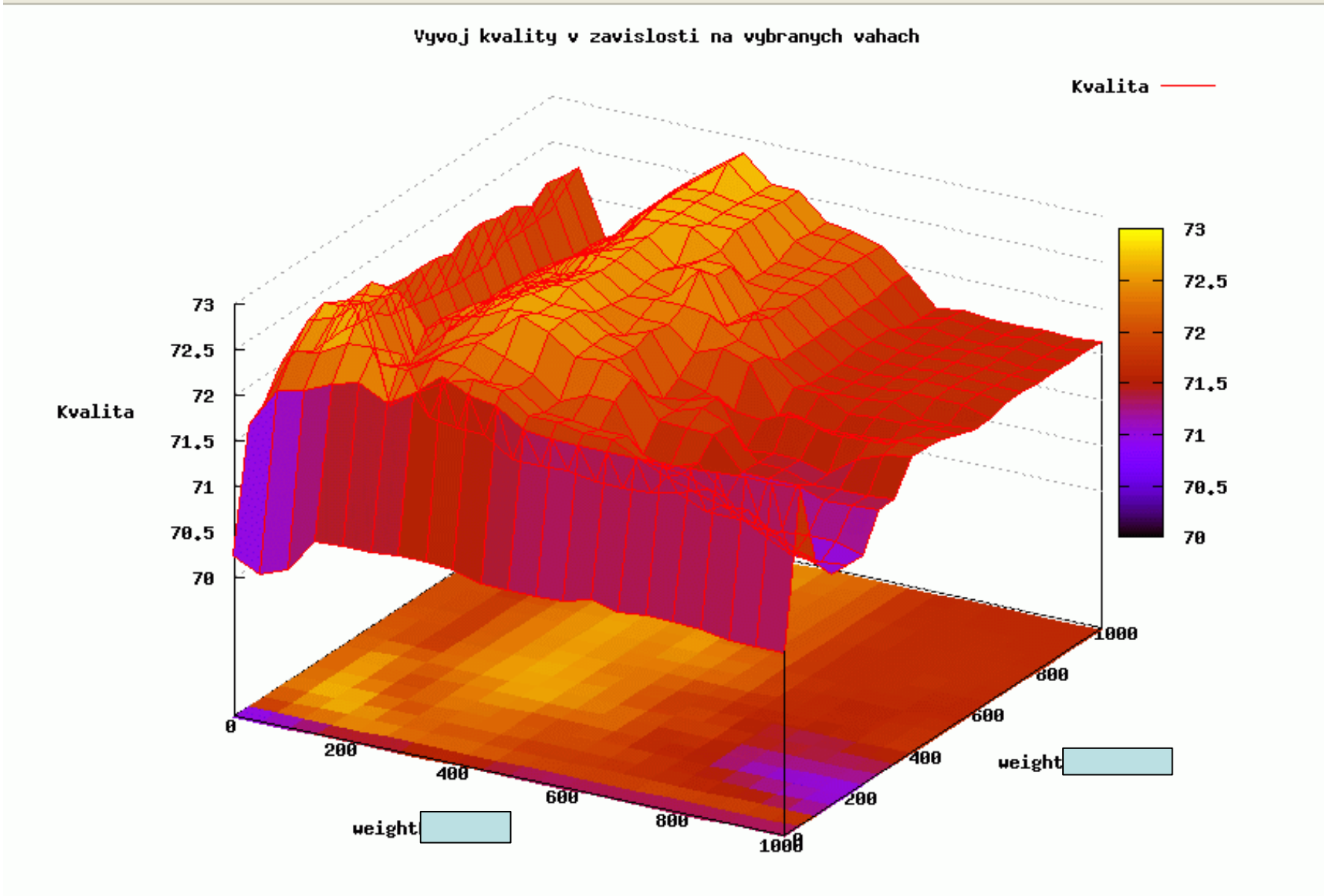
Historie ladění parametrů v Seznamu

- Od oka
  - nějak nastavit parametry a pak to nějak zkoumat
  - ve více lidech od oka, pak se hádáme
  - každý dodá dotazy, kde jsme lepší, horší, beze změny
- Využití kalibrací a měření kvality fulltextu
  - Ručně nastavovat, ale hned vidím kvalitu (i dotazy, na kterých to drhne)
- Automatické nastavování vah

Nastavovače vah



Nastavovače vah

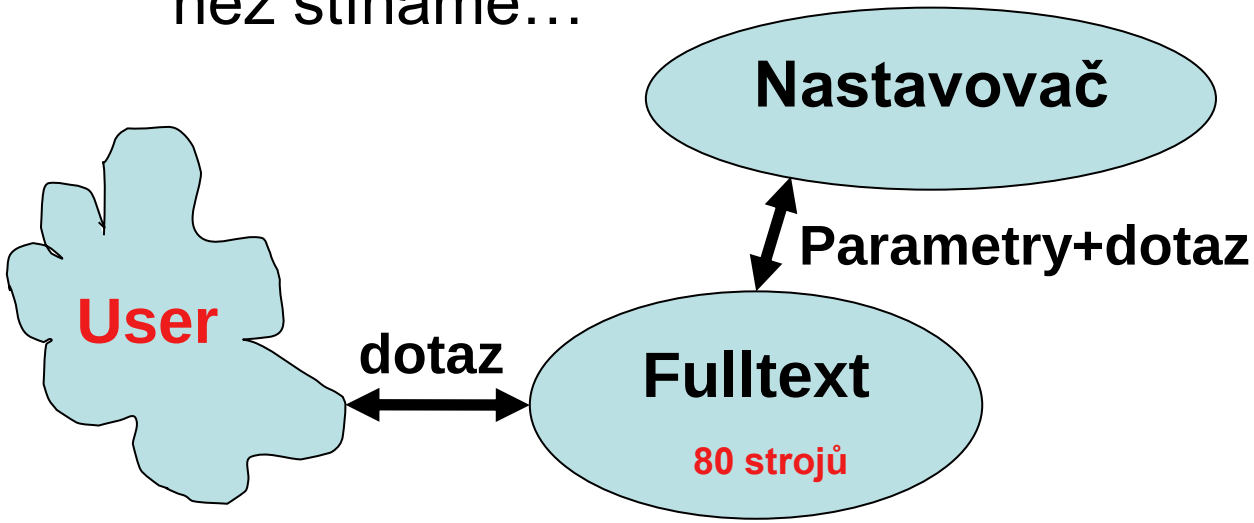


Otázka pro vás:

Jak odstranit bottle neck?

Když změníme parametry, tak se musíme pro všechny nakalibrované dotazy zeptat fulltextu na nové pořadí výsledků. Podle toho poznáme, jestli jsme si pomohli…

Potřebujeme se ptát mnohem více než stíháme…



Co vše se při návrhu fulltextu využije?

Pestrý tým vývojářů

- Softwarové inženýrství
  - Práce s velkým objem dat
  - Poskytování online výsledků
  - Databáze a distribuované systémy
  - Optimalizace na výkon, paralelizace
- Strojové učení
  - Klasifikátory dotazů
  - Klasifikátory stránek (např. citlivý obsah)
- Statistika, datamining

Pestrý tým vývojářů

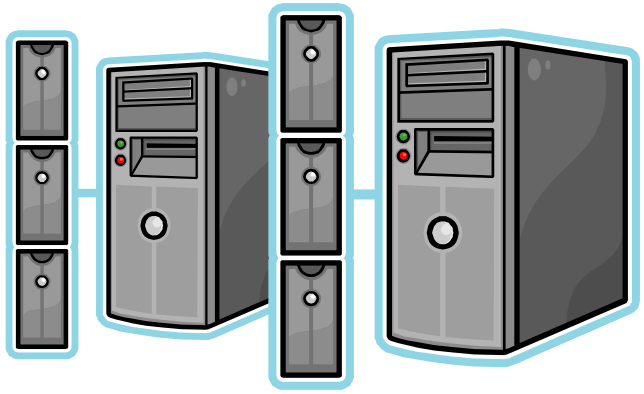
- Lingvistika
  - Lemmatizace, syntaxe věty
  - Pochopení dotazu
  - Zkratky
  - Oprava překlepů (např. fonetický přep
  - Kolokačnost slov
  - Desambigulace
- Grafové algoritmy
  - Odkazová síť, graf internetu
- Další chytré přístupy

Děkuji za pozornost.

Technické parametry a statistiky

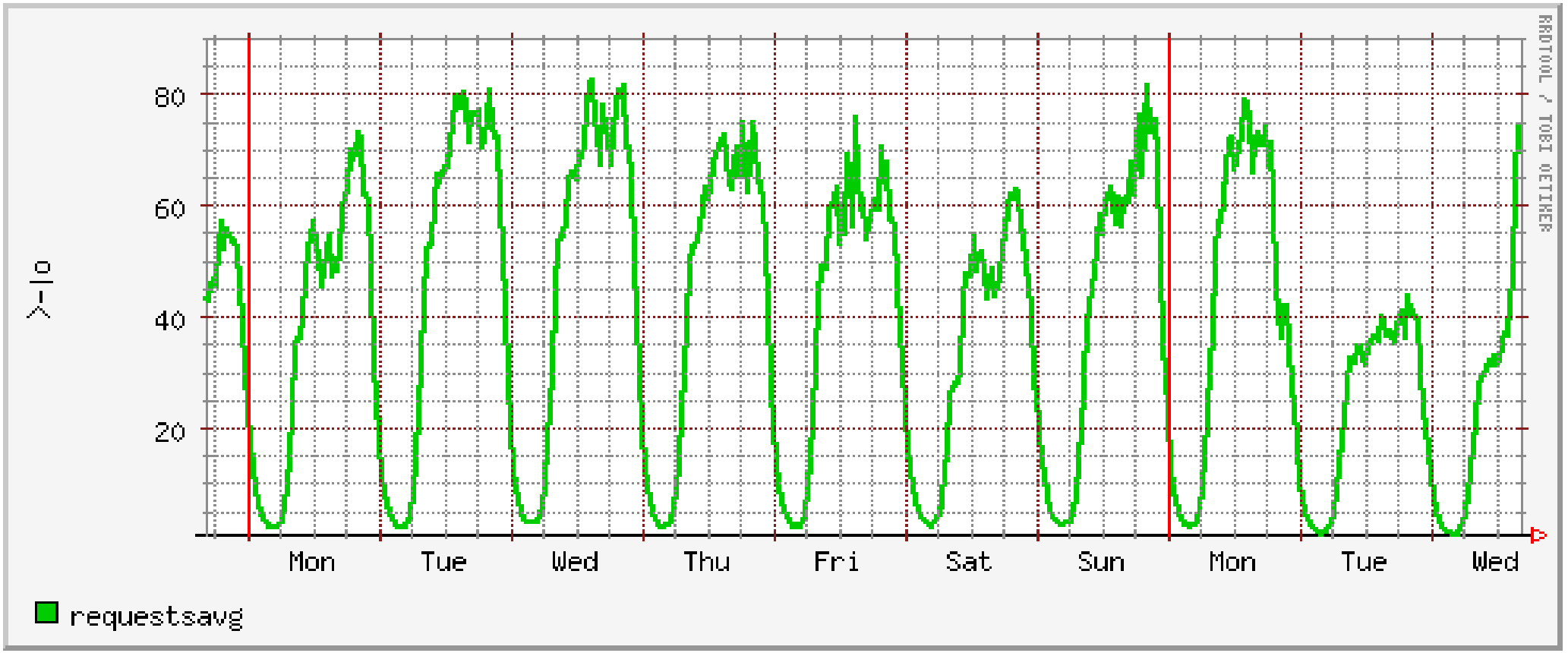
Hardware v provozu

- 4x webovka, metasearch, lemmatizace
  - Quad-Core Xeon X3550, 2x2Ghz
  - Disky: 2x70G
  - Paměť: 3G
- 72x basesearch, content server
  - Quad-Core Xeon X3650, 2x2Ghz
  - Disky: 6x160G
  - Paměť: 16G
- 24x strojů pro Robota a Indexaci



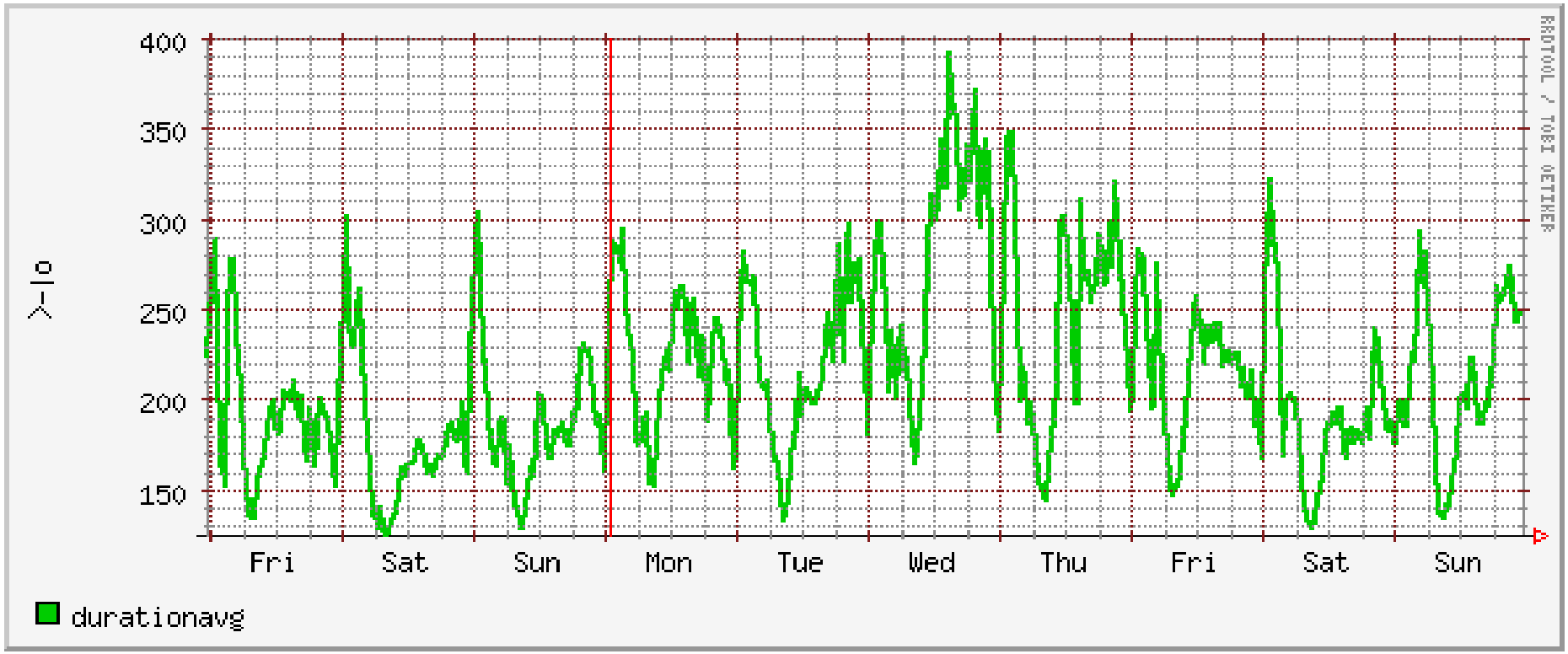
Stroj pro výpočet PageRanku má 64G RAM.

Zátěž během týdne



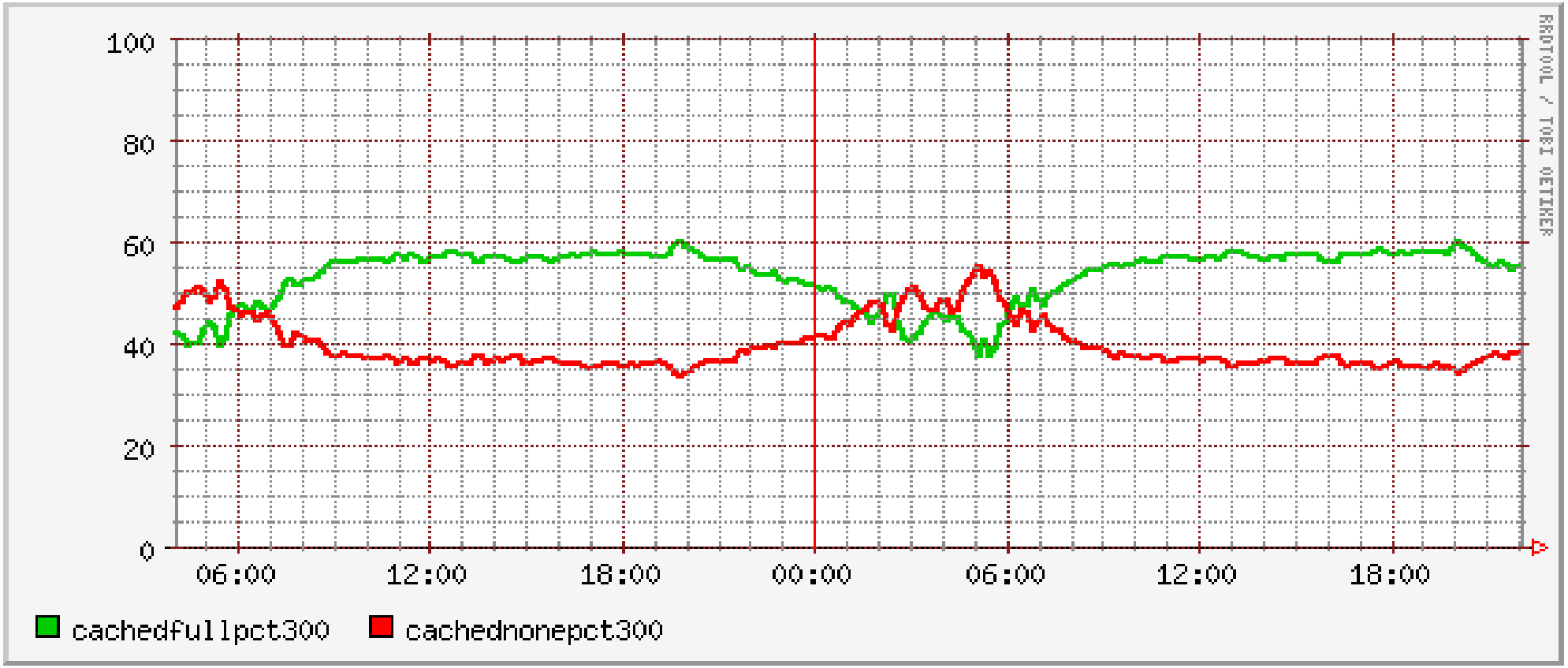
- 1/4 zátěže
- až 400 dotazů/s

Doba odezvy během týdne



- Doba odezvy v msec

Úspěšnost query cache



- Úspěšnost cache v %

Výkon robota

Rychlost stahování	> 450 stránek / sec
Průměrná stránka	~11 kB (zdojový kód)
Denní objem	~40 miliónů dokumentů cca 410 GB dat