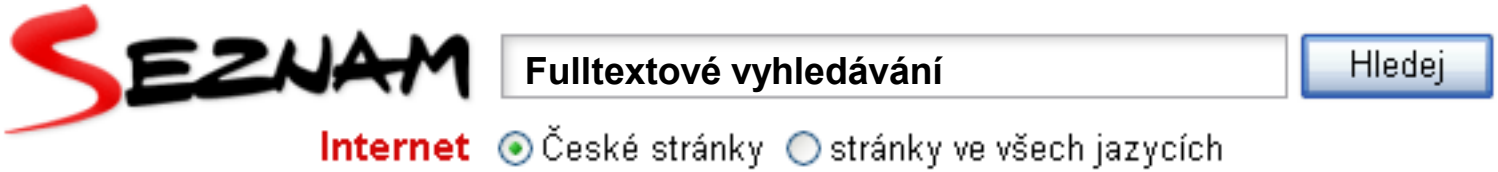


Petr Nevrlý <petr.nevrly@firma.seznam.cz>

Obsah přednášky

- Vyhledávání
  - Cíl vyhledávání
  - Architektura ve zkratce
  - Vyhledávání
  - Robot
  - Údaje z provozu
- Novinky ve fulltext (2009)
  - Screenshot generátor
  - Rozpoznání citlivého obsahu
  - Populární odkazy
  - Oprava překlepů
  - „Miniaplikace“
  - Podpora GEO-mikroformátu
  - Nová verze vyhledávání

Cíl fulltextového vyhledávání

- Poskytnutí odpovědi na dotaz uživatele

Cíl fulltextového vyhledávání

- Poskytnutí odpovědi na dotaz uživatele
  - Shromažďování
    - Rychlý robot
    - Spolehlivá indexace
    - Zakládání „správných“ dokumentů
  - Zpracování
    - Vhodná struktura DB
  - Vydání (řazení)
    - Výkon (rychlost)
    - Dostupnost
    - Konzistence
    - Kvalita

Typy fulltextů

- Vyhledávače jsou si velmi podobné, liší se jen v detailech

- Jako…

Část 1 – Architektura ve zkratce

1. Hlavní části
2. Redundance v provozu
3. Blokové schéma

Hlavní části

stranky na internetu

stahovani stranek

robot

indexovane stranky

priprava databaze

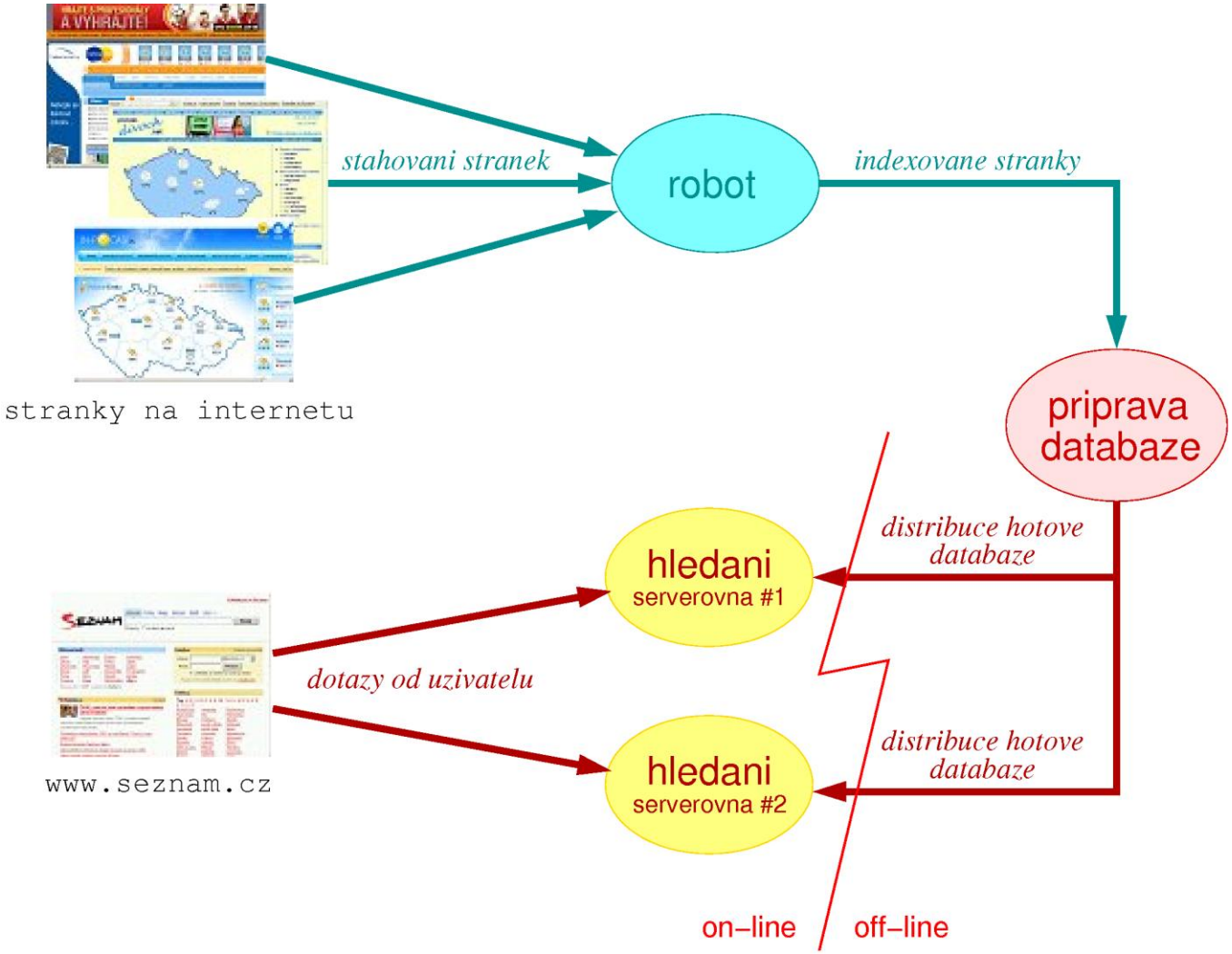
distribuce hotove databaze

hledani

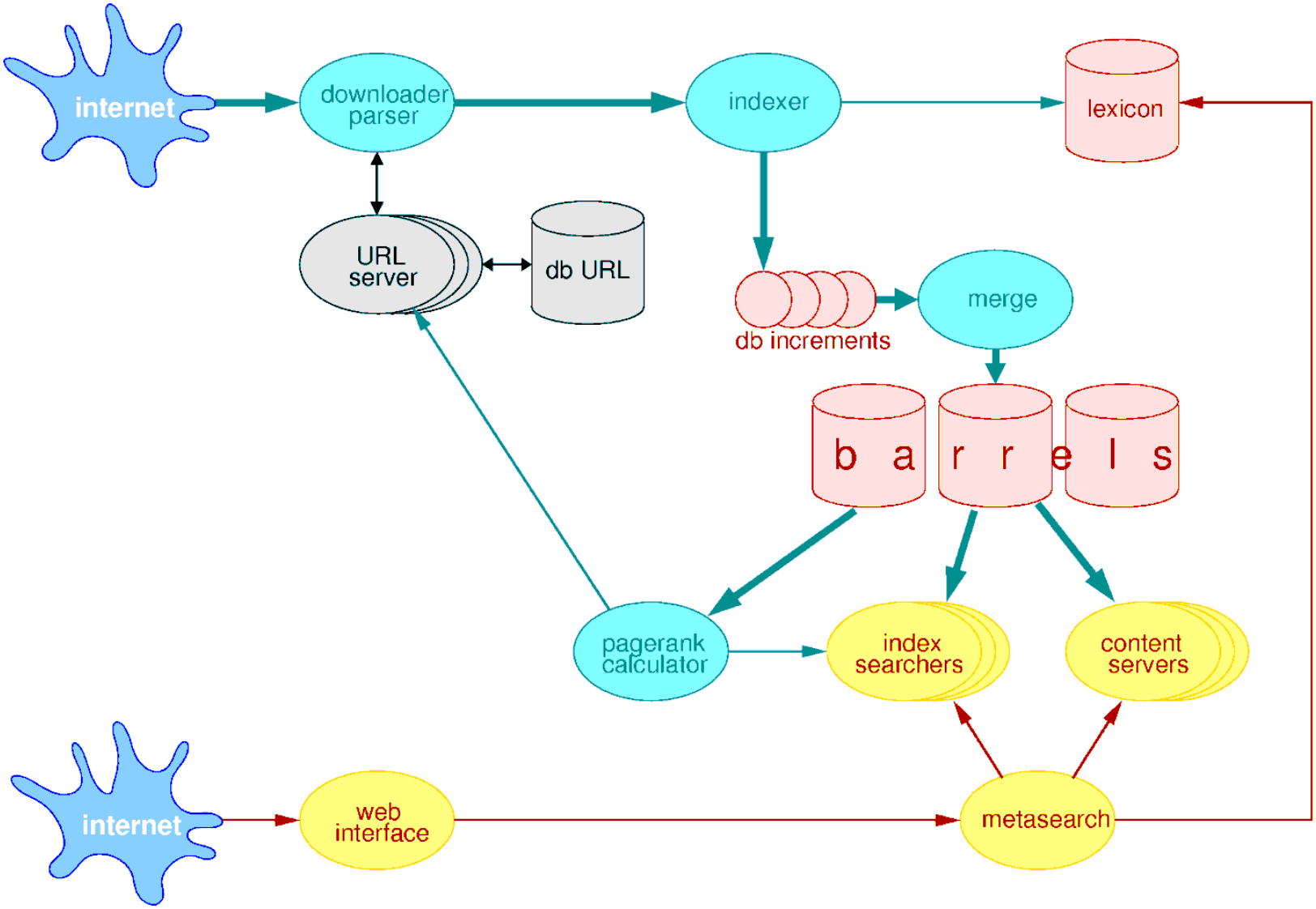
dotazy od uzivatelu

www.seznam.cz

Hlavní části – Redundance v provozu



Blokové schéma



Část 2 – Vyhledávání

1. Zadávané dotazy
2. Lemmatizace
3. Hodnocení stránek

Zadávané dotazy (1)

- 10 náhodných dotazů
  - posilovna
  - plné hry ke stažení zdarma
  - plemena koní
  - planovac tras
  - petra němcová fotky
  - paragrafy a zákony
  - papírové vystřihovánky
  - panenka chou chou
  - paintball bazar
  - oplocení

Zadávané dotazy (2)

- Forma dotazů:

  Nejedná se přímo o otázky

  – přídavná a podstatná jména
  – 1. pád
  – jednotné i množné číslo
  – občas bez diakritiky

Lemmatizace

- Lemma = základní tvar slova
- Věta:
  „Jeden z nejlepších zdrojů o německých tancích.“
- Lemmatizováno:
  Jedna/Jíst z dobrý zdroj o německý tank/tanec.
- Disambiguace = vyloučení nejednoznačnosti

Hodnocení stránek (1)

Hlavní město **Praha** - Informační server pražské radnice
Praha udělá maximum, aby dopravní investice město bolely co nejméně. Informační server Hlavního Města **Prahy**
magistrat.**praha**-mesto.cz/ - Hlavní město Praha

- **Titulek !!**
- Obsah stránky
- URL

Hodnocení stránek (2)

Citační analýza pro dotaz „Ostrava“

Hodnocení stránek (3)

- Pagerank = statická „důležitost“ stránky založená na citační analýze
- Předpoklad: statisticky náhodné chování
- SPAM - blackSEO

Část 3 – Robot

1. Hledání nových stránek
2. Reindexace stránek
3. Ne-HTML formáty

Hledání nových stránek (1)

- Před 5 lety start
- Procházení nalezených odkazů
  - Domény .cz, .sk, .com, .org, .net, .info, …
- Hledá stránky v českém jazyce
- Alternativní zdroje: RSS a sitemap

Hledání nových stránek (2)

- Robots.txt – standardní protokol pro zakázání přístupu robotů (www.robotstxt.org)

- Textový soubor http://example.com/robots.txt

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

User-Agent: Bot
Disallow: /
```

Reindexace stránek (1)

- Každý den se vybere množina stránek pro reindexaci

- Při výběru se hodnotí
  - Datum poslední návštěvy
  - Rank (Srank)
  - Frekvence změn

Reindexace stránek (2)

- Přetěžování webserverů
  - Shapování podle IP adresy
  - Omezení max počet URL / sec

Ne-HTML formáty

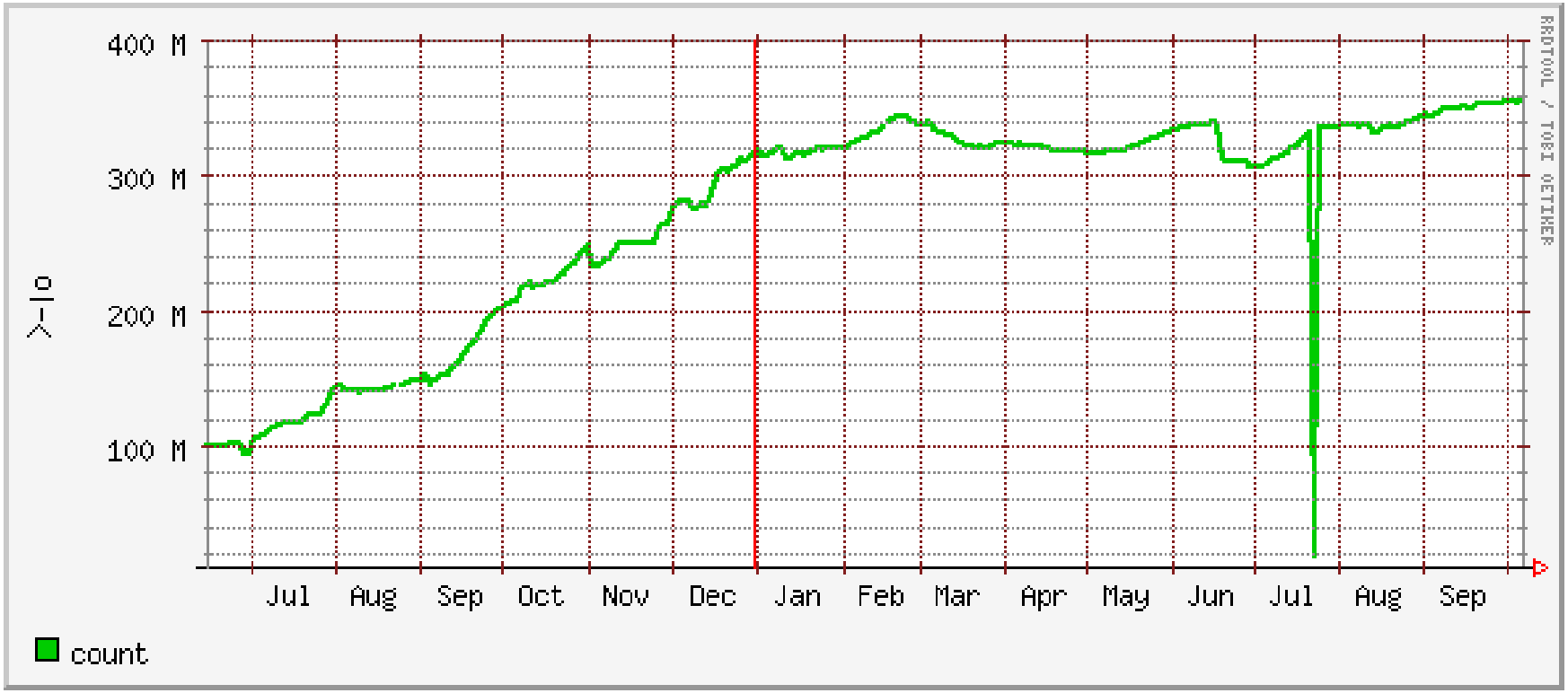
- PDF
- DOC (MS Word)
- RTF
- **PPT** (v roce 2009)

- Operátor filetype:

query filetype:pdf,html,ppt | Vyhledat Seznamem

Část 4 – Aktuální údaje z provozu

Velikost databáze (1)

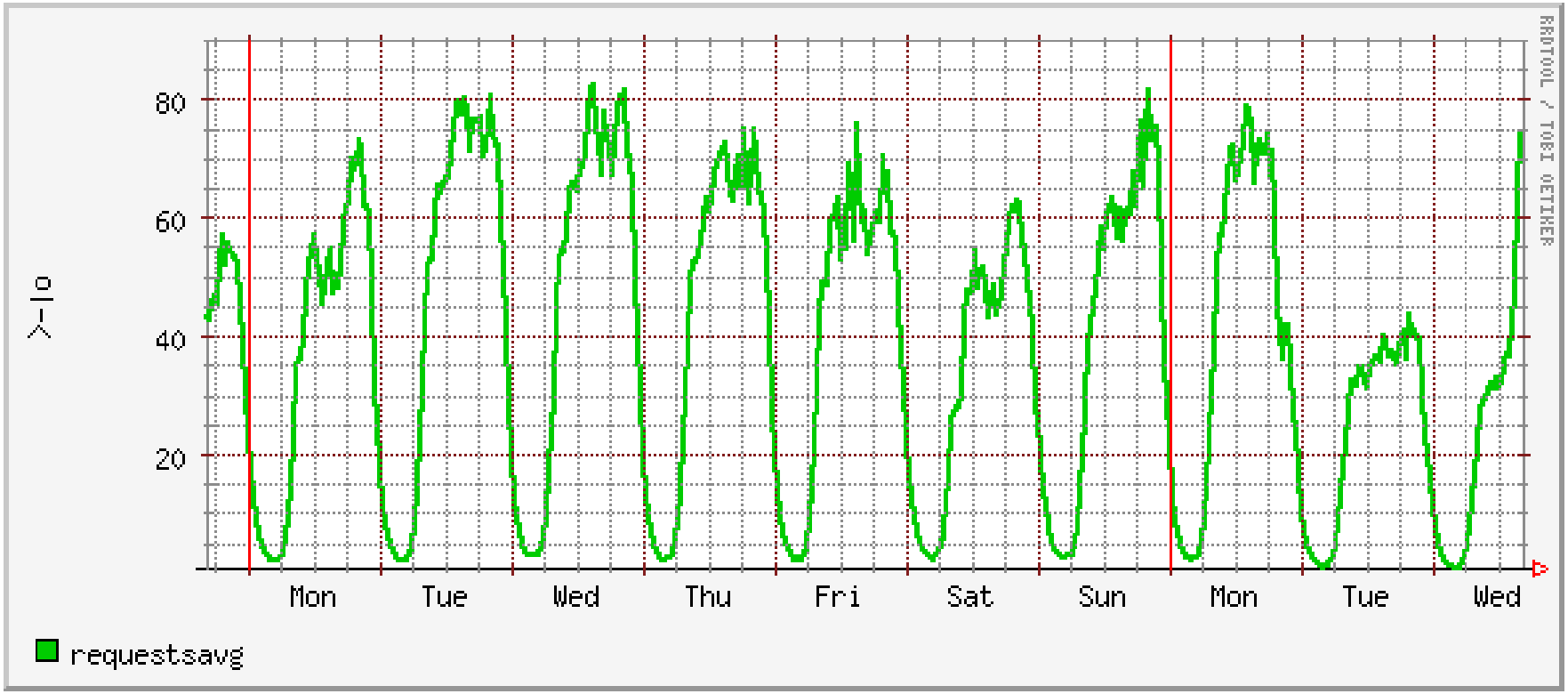


- Počet dokumentů

Velikost databáze (2)

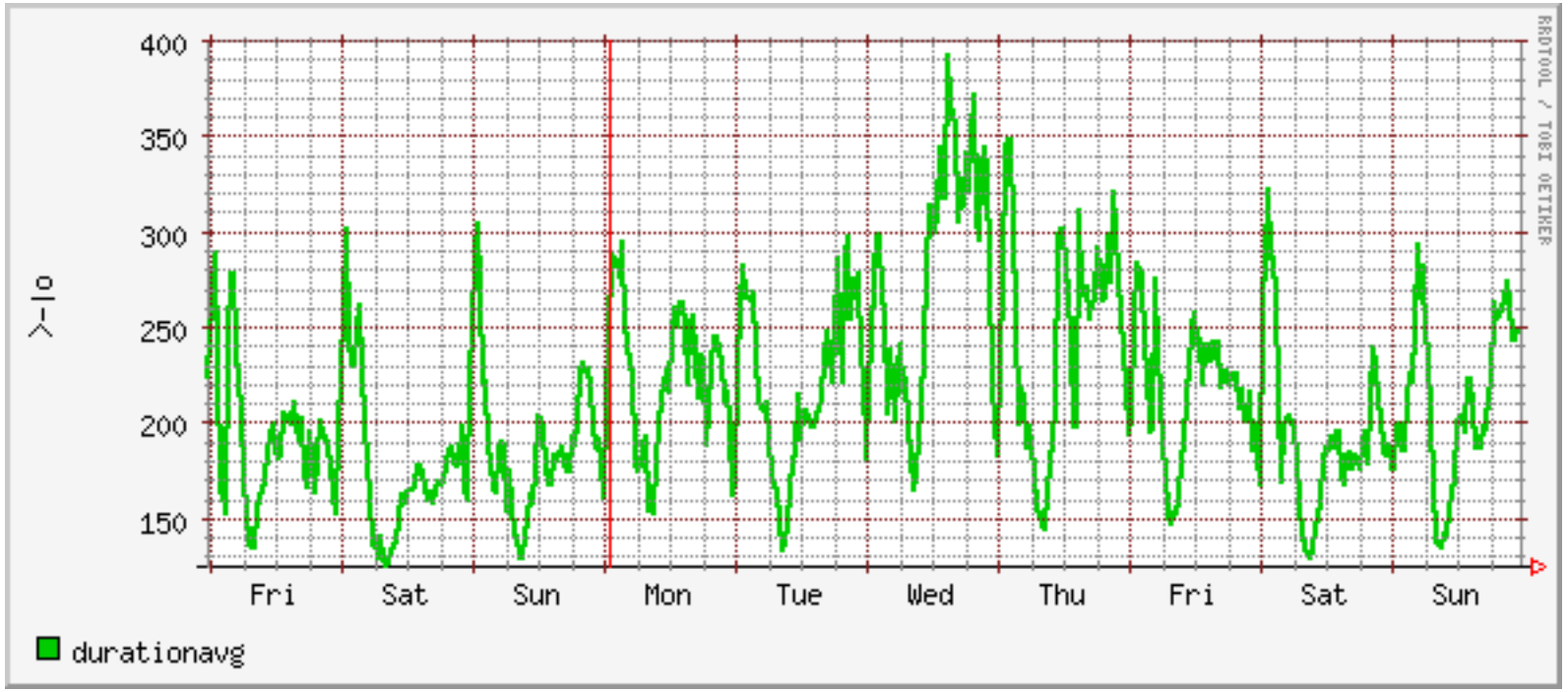
Počet dokumentů	355 miliónů
Indexy	1,8 TB
Obsah dokumentů (texty)	1,4 TB
Průměrný text	6 kB / dokument

Zátěž během týdne



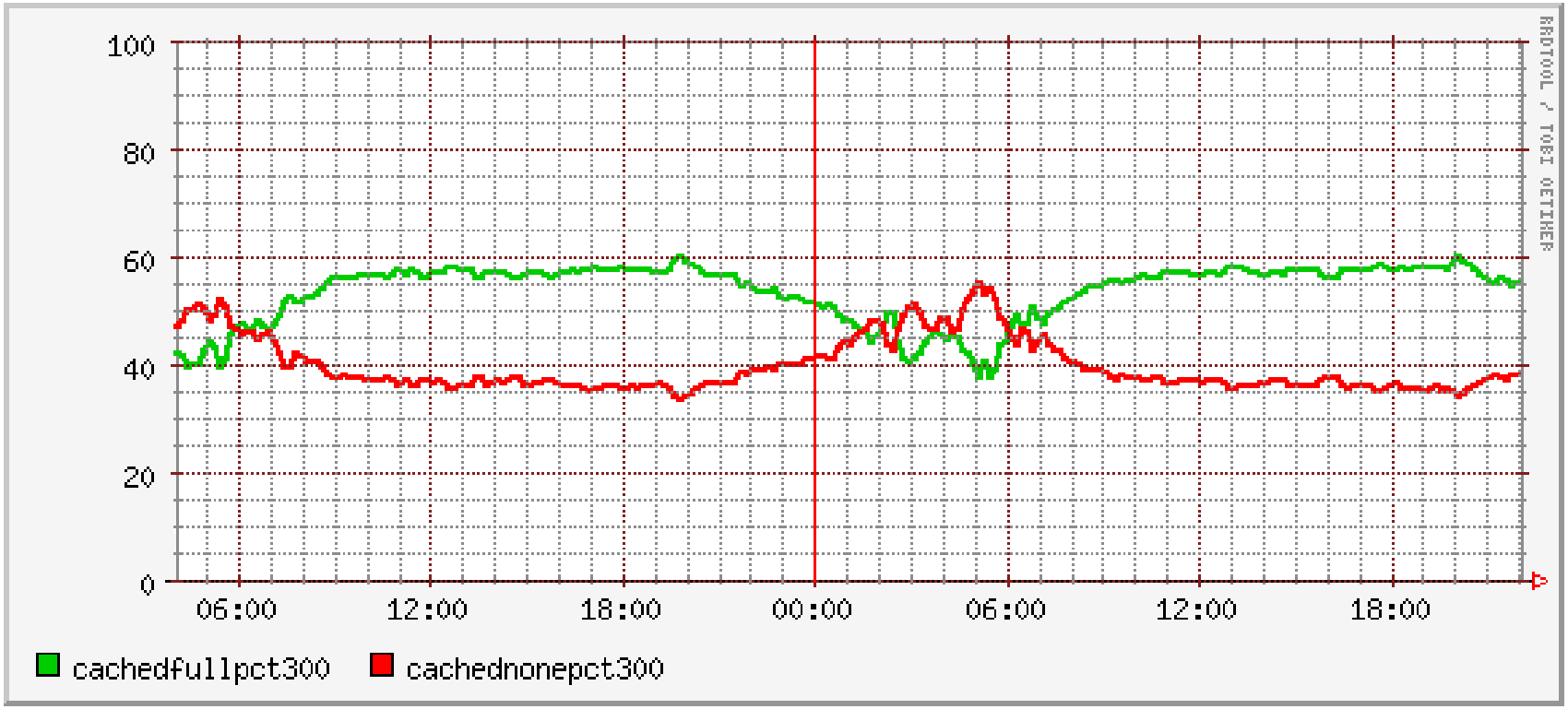
- 1/4 zátěže resp. 1/6
- až 400 dotazů/s

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

Stáří dokumentů ve dnech

Minimální	1
Maximální	135
Průměr	6,9
Nejčastěji	1,2 – 9,5

Novinky v roce 2009

- Screenshot generátor
- Rozpoznání citlivého obsahu
- Populární odkazy
- Podpora GEO-mikroformátu
- Nová verze vyhledávání

Screenshot generátor - snímání

- 10 URL/sec (1M URL/den)
  - Max >20 url/sec
- 6 GB dat/den

- Rozlišení 700x525 px
- Barevná hloubka 5 bitů
- Formát PNG

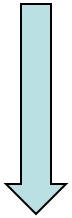
Screenshot generátor - storage

- 660M obrázků
  - 150M unikátních dokumentů
- Data cca 1,6TB
- PNG v speciální data storage
- 2,2kB avg img

Screenshot generátor - výdej

- >1 800 img/sec
- NoImage ~0,75%
- http://fimg.seznam.cz/?spec=ft100x75&url=http%3A//search.seznam.cz/

- Zkracování cesty



http://www.vse.cz/vedeni/hindls.php

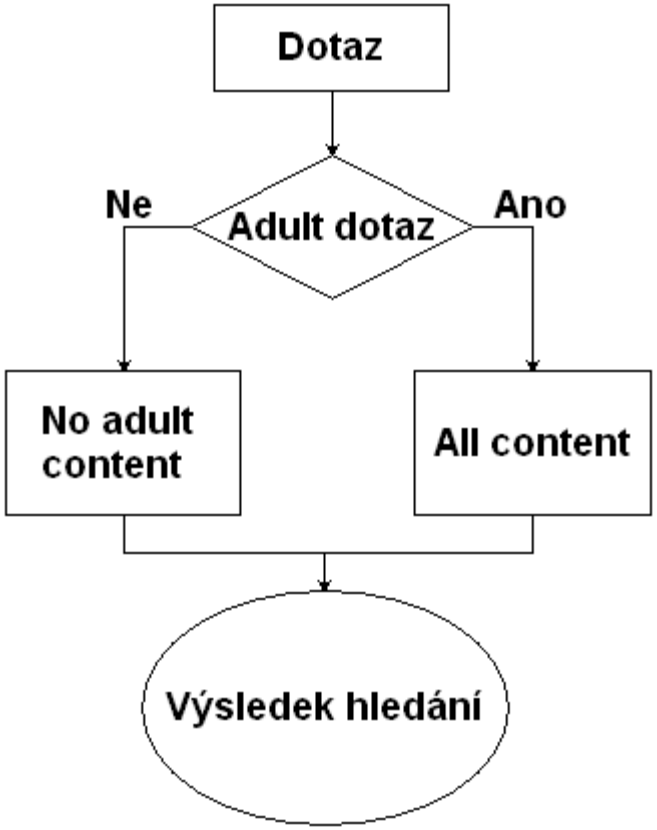
http://www.vse.cz/vedeni/

http://www.vse.cz/

Screenshot generátor - HW

- Výdej
  - 2 x 8 serverů
  - 2 x QuadCore
  - 8GB RAM
- Generátor
  - 1 x 4 servery x 4 vituály x 10 Mozilla
  - 2 x QuadCore
- Repository
  - 1 x 1 server
  - 16 x 1TB SATA

Rozpoznání citlivého obsahu



Rozpoznání citlivého obsahu

1. Detekce adult dotazů
2. Detekce adult dokumentů

- http://search.seznam.cz/?q=**pupendo+fotky**
  - filtr funguje automaticky, stejné jako s parametrem **&safe=auto**
- http://search.seznam.cz/?q=**pupendo+fotky**&safe=no
  - filtr je vypnutý bez ohledu na vyhodnocení dotazu
- http://search.seznam.cz/?q=**pupendo+fotky**&safe=yes
  - filtr je zapnutý a do SERP nejsou zařazeny nevhodné stránky bez ohledu na zadaný dotaz

Populární odkazy

Novinky.cz

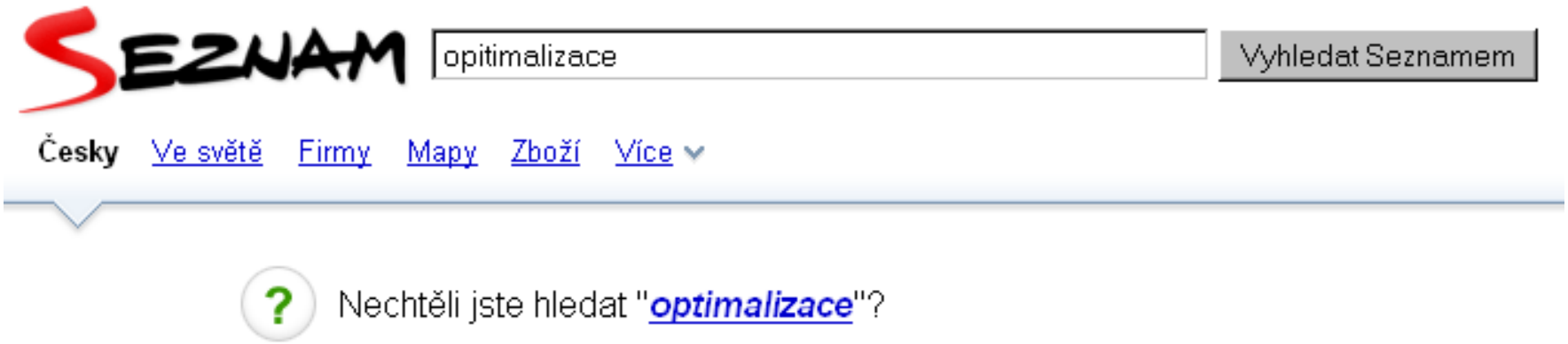
Novinky.cz. Klávesové zkratky na tomto webu Na obsah stránky. Redakci pište na email redakce@**novinky**.cz Deník Právo oslovíte přes adresu redakce@pravo.cz Pokud chcete informaci ...

- Denní tisk
- Zahraniční
- Domácí
- Stalo se
- Krimi
- Koktejl

www.**novinky**.cz/

- Text odkazu z textu odkazu na stránce
- Jen u prvního výsledku
- Podstránky webu
- Statistické zpracování

Oprava překlepů



„Miniaplikace“

Podpora GEO-mikroformátu

Botanická zahrada, Praha

Botanická zahrada, Praha (c) 2006 Dan Meszaros dna@cdi.cz. Posezení pro návštevníky pražské **botanické zahrady**. Nebyl sice zrovna ideální sluneční den, ale tém pár **...**

dna.cdi.cz/fotky/zelen/**botanicka-zahrada-...** - Zobrazit na mapě

- http://microformats.org/wiki/geo

```
<cokoliv class="geo">
    <cokoliv class="latitude">50.071583</cokoliv >
    <cokoliv class="longitude">14.400785</cokoliv >
</cokoliv>
```

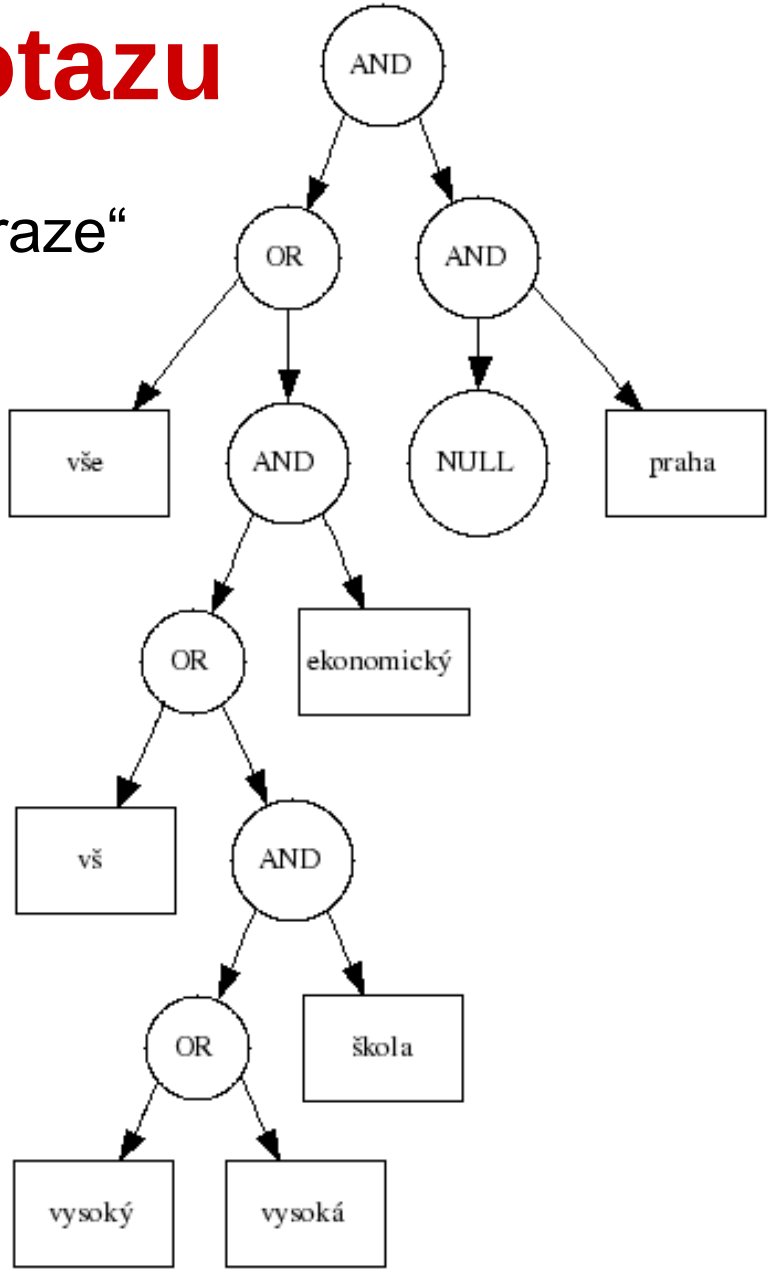
Nová verze vyhledávání

- Hlavní změny
  - OR + expanze dotazu
  - Nová lemmatizace
  - Lepší „oháčkování“
  - Kolokace
  - Využití „Admintools“

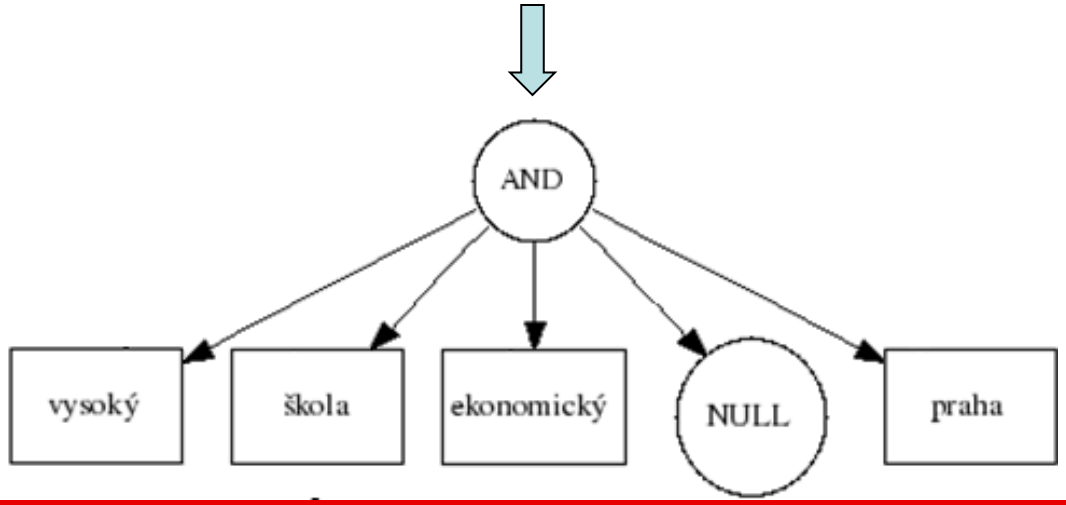
OR, expanze dotazu

Query: „Vysoká škola ekonomická v Praze“

Nové hledání



Staré hledání



Nová lemmatizace + Lepší „oháčkování“

- Umí i „nová“ a převzatí slova

- Staré hledání
  - „barum“ → bar
  - „barům“ → bar

- Nové hledání
  - „barum“ → barum
  - „barům“ → bar

Kolokace

- Význam spojitosti dvou sousedních slov
- Zohledňění ve výpočtu vzdál. slov na stránce

  – Dotaz „plzeňské pivo“ → kolokace=0,9
  – Dotaz „jiří topolánek“ → kolokace=0,4
  – Dotaz „vše uk“ → kolokace=0,1

AdminTools

- Porovnání vybraných vyhledávačů
- Ověřování dopadů změn v hledání
- „Automatické“ nastavení vah pro hledání

- Externí kalibrátoři hodnotí řádově stovky dotazů a desetitisíce dokumentů (počet se neustále navyšuje)

- Víc informací o AdminTools na další přednášce

Konec

Děkuji za pozornost

http://fulltext.sblog.cz

„Bonusy“

1. TOP 10 dotazů
2. SEO

Top 10 dotazů

r. 2008	r. 2009
1. ""	1. ""
2. youtube	2. youtube.com
3. libimseti.cz	3. libimseti.cz
4. superhry	4. superhry
5. freefoto	5. o2
6. freevideo	6. freevideo
7. redtube.com	7. facebook
8. sms zdarma	8. aukro.cz
9. google	9. google
10. porno	10. porno

SEO

(search engine optimalization)

1. URL
2. Obsah stránky
3. JavaScript a Flash

URL

- Vhodně zvolená doména
  - www.csas.cz
  - www.ceskasporitelna.cz
- Optimalizované URL a rewrite
  - super.cz/index.php?clid=18656
  - novinky.cz/vladni-spis-jak-zabranit-uniku-informaci-na-internet-unikl-na-internet
- Minimalizovat duplicity!!

Obsah stránky

- Titulek
  - Důležitá součást stránky
  - Unikátní na každé stránce

- Text
  - Správně používat sémantické značky
  - Nepoužívat text jen na obrázku

JavaScript a Flash

- Robot neumí procházet přes:
  - formuláře
  - JavaScript navigaci
  - Flash presentace
  - JavaScript přesměrování

- Textová alternativa k dynamické navigaci

Konec (2)