Fulltextový vyhledávač

Štěpán Škrob

<stepan.skrob@firma.seznam.cz>

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

O čem bude přednáška?

1. Úvod
2. Přehled architektury vyhledávače
3. Důležité datové struktury
4. Použité algoritmy
   a) Crawlování
   b) Indexace
   c) Vyhledávání
5. Výkonové statistiky

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Úvod

- Vyhledávače jsou si prakticky velmi podobné, liší se pouze v implementačních detailech :-)

- Je to jako s auty…

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Přehled architektury vyhledávače

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Přehled architektury (1)

- Původní škálovatelnost do 100 M dokumentů
- Nyní až 1 G dokumentů
  - Rozdělení databáze do svazků
  - Zkrácení odezvy vyhledávače

- Současná databáze
  - Počet dokumentů 30 M
  - Datová velikost 100 GB

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

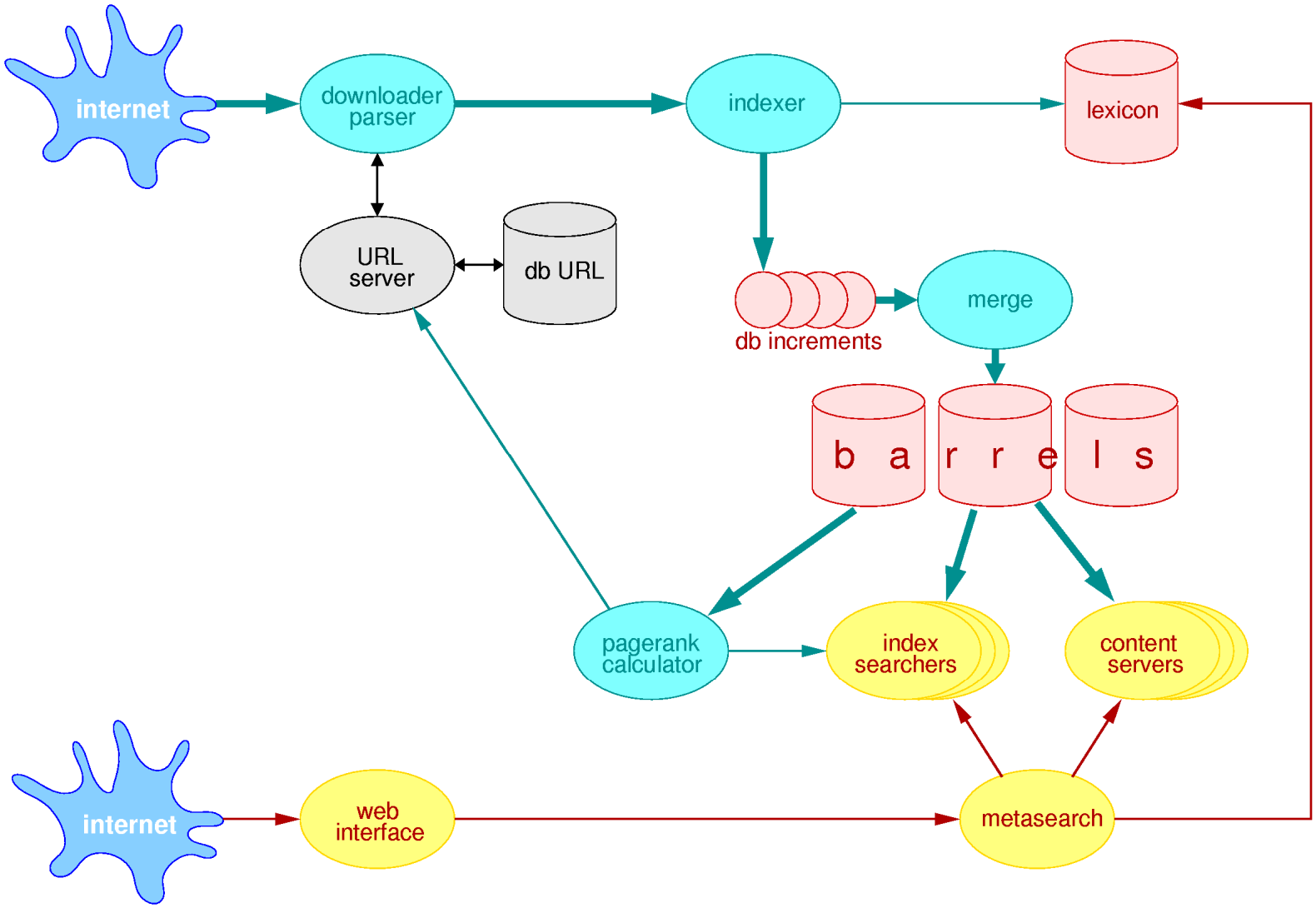
Přehled architektury (2)

- Výkon robota (crawlera)
  - 2 M dokumentů / den
- Dvě úzká hrdla
  - Výběr dokumentu k reindexaci
  - Přeložení URL na url_id
  - Obojí je náhodný přístup k datům…

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Přehled architektury (3)



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Důležité datové struktury

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

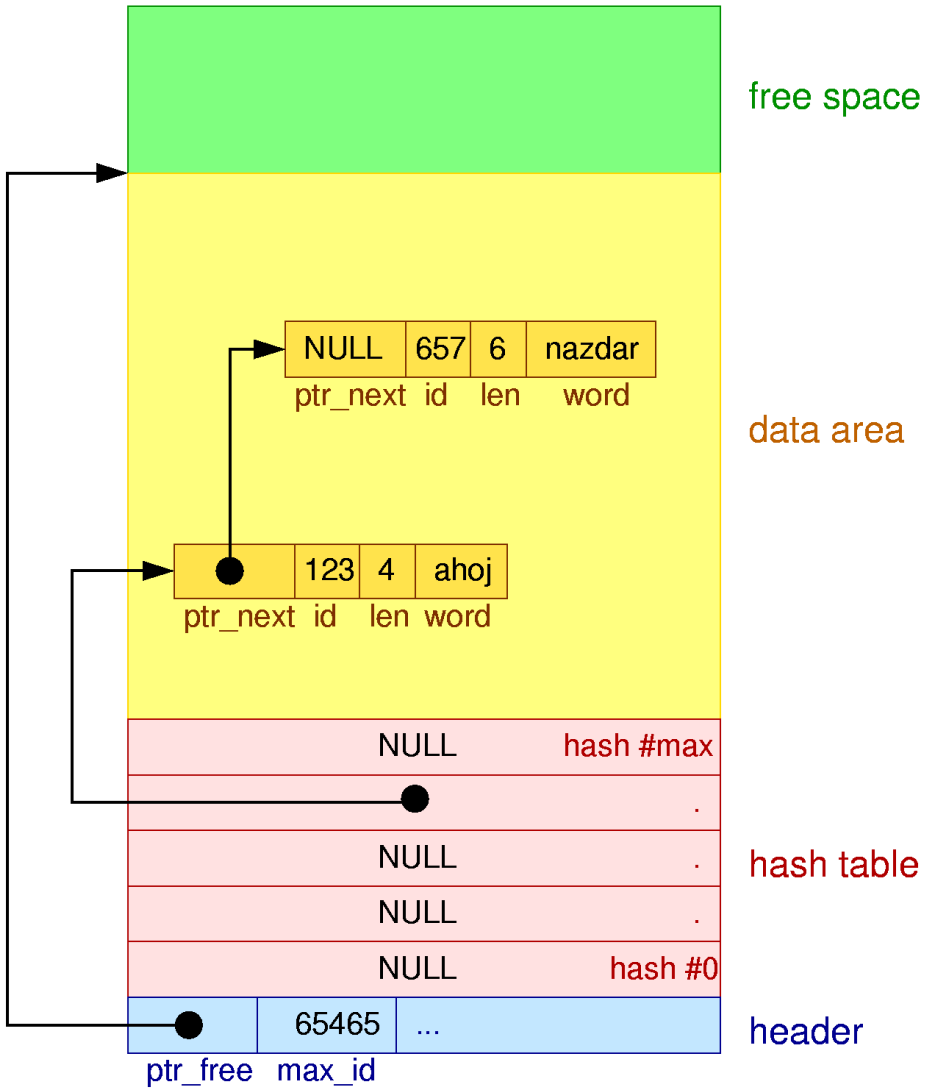
SEZNAM

Důležité datové struktury

1. Lexikon
2. Invertovaný index
3. Texty dokumentů

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Lexikon (1)



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Lexikon (2)

- Lemmatizátor je spojený s lexikonem do jedné komponenty
- Obsah lexikonu
  - Lemmatizovaná slova => lemmata
  - Nelemmatizovaná slova => termy

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

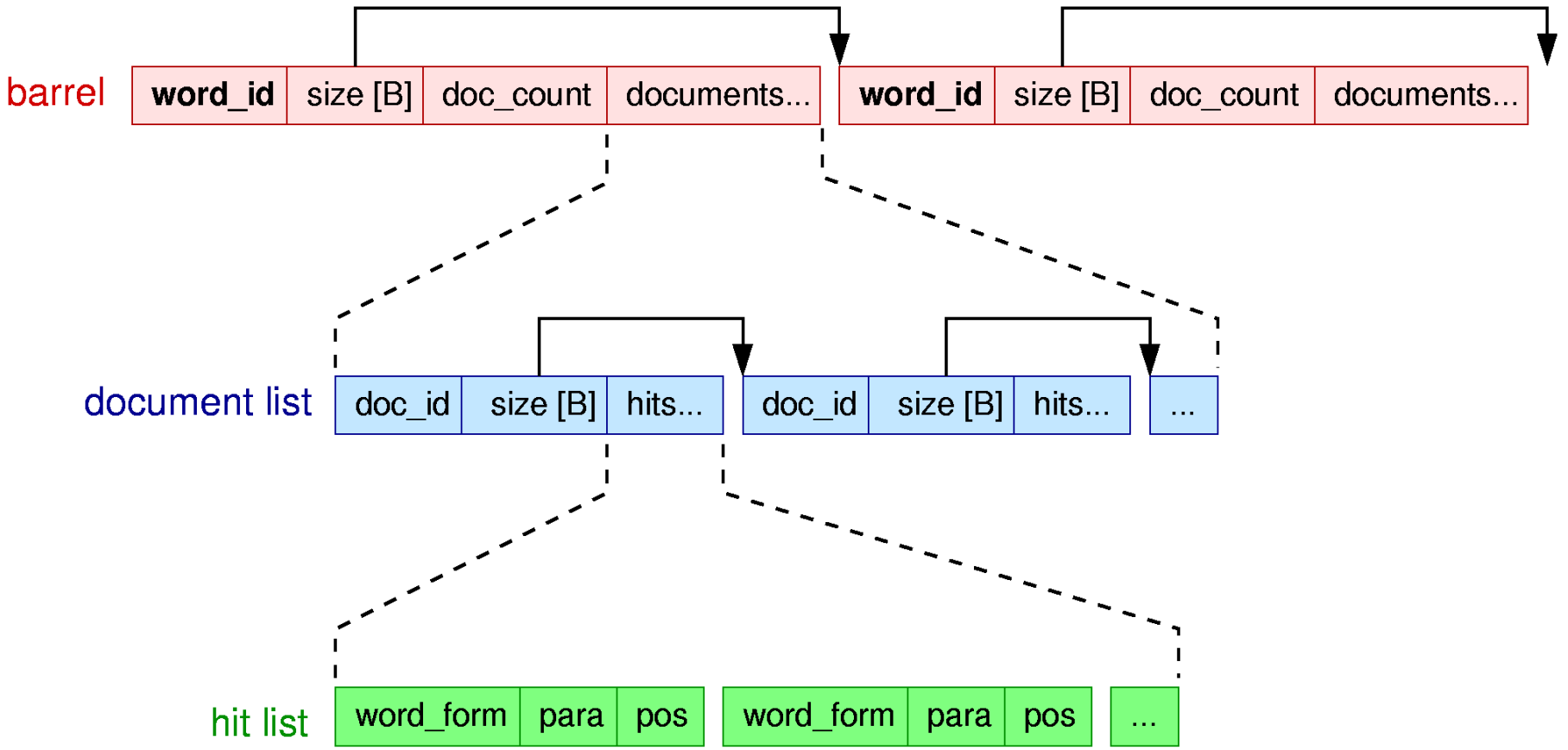
Lexikon (3)

- Statistické údaje
  - 450 MB celková velikost
  - 22 miliónů slov
  - 16 miliónů hashů (využitých 75%)
  - Největší kolize 13 slov / hash
- Výkon
  - 100.000 slov / sec

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Invertovaný index (1)



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Invertovaný index (2)

- Rozdílová + prostorová komprimace ideček
- Prostorová komprimace ostatních *int* čísel

- Uchovávané informace o slově:
  - Tvar slova (5 bitů)
  - Číslo odstavce
  - Pozice v odstavci

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

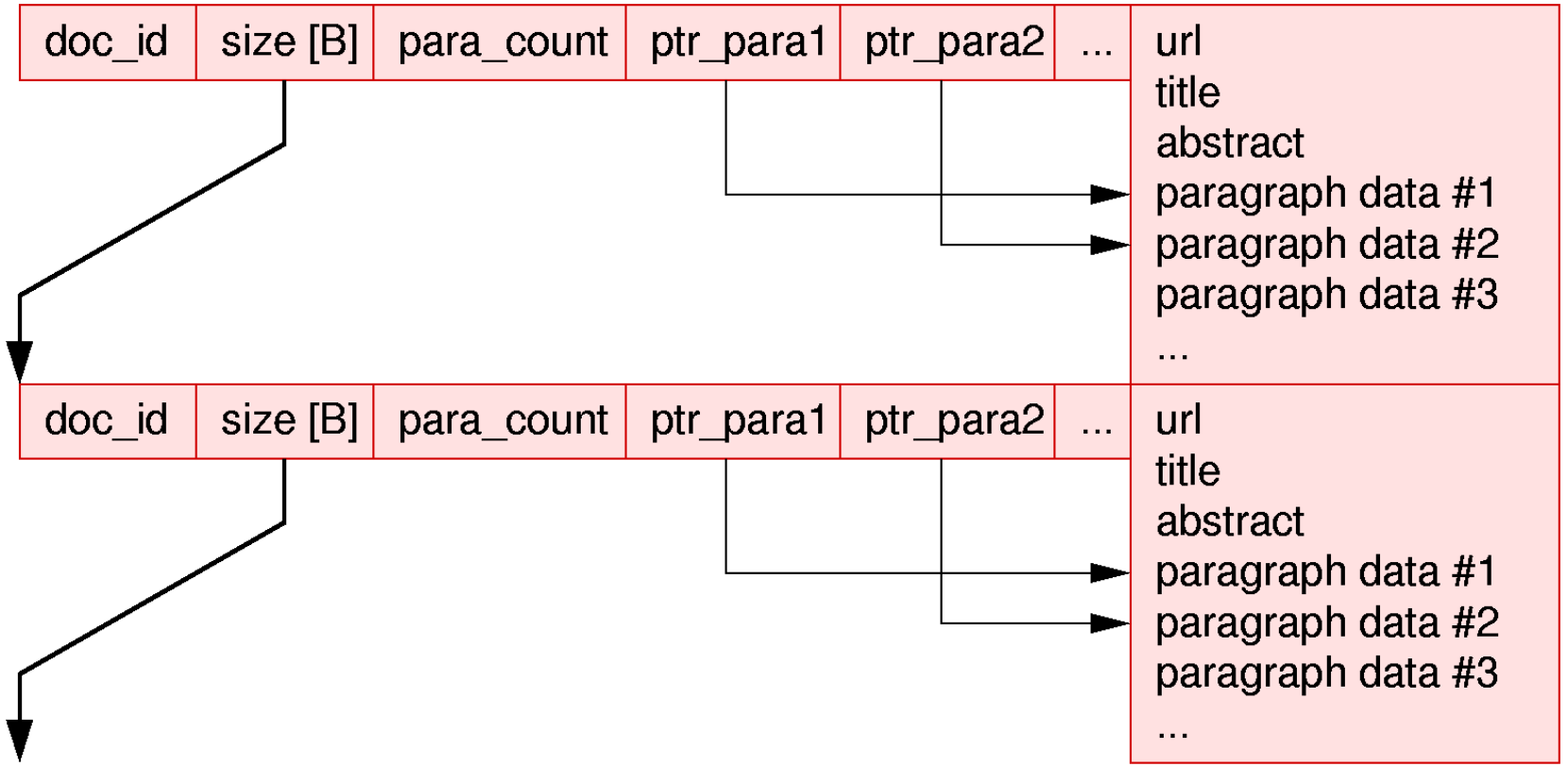
Invertovaný index (3)

- Pro nalezení záznamu slova v barrelu (invertovaného indexu) slouží další index
  - Trvale přítomný v paměti (no I/O)
  - Obsahuje informace o souboru, pozici, délce
  - Seřazený podle word_id => binární půlení
- Statistické údaje (30 M dokumentů)
  - Invertovaný index 50 GB
  - Trvale načtené struktury ~250 MB

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Texty dokumentů (1)



www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Texty dokumentů (2)

- Texty všech zaindexovaných stránek včetně balastu
  - Zachované informace o odstavcích
  - Označkované začátky a konce slov
  - Neobsahuje HTML značky
- Použití pro dynamický popisek dokumentu
- Automatický abstrakt dokumentu

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Texty dokumentů (3)

- Statistické informace
  - Celková velikost 50 GB na 30 M dokumentů
  - Průměr 200 slov na dokument
  - Průměrná velikost slova 7 znaků

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Použité algoritmy

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Použité algoritmy

1. Crawlování
   a) Filtrace dokumentů
   b) Detekce českého jazyka
2. Indexace
   a) Lemmatizace
   b) Identifikace podobných dokumentů
3. Vyhledávání
   a) Doplňování diakritiky
   b) Výběr vhodného lemmatu
   c) Hodnocení relevance

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Filtrace dokumentů (1)

- Základní filtrace textových dokumentů podle Content-Type
  - Uhádnutého z URL
  - Zjištěného z HTTP hlaviček serveru
  - Pozor na CGI scripty a URL zakončená lomítkem (http://novinky.cz/sport/)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Filtrace dokumentů (2)

- Vyřazení binárních souborů
  - Citlivost na řídící znaky ASCII < 32
  - Hledání tokenů podobných slovům (mezery a délka slova)
- Vyřazení zcela duplicitních dokumentů
  - MD5 hash z textového derivátu

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Detekce českého jazyka

- Kombinace slovníkové a statistické metody
  - Slovník ze 100.000 článků novinky.cz
  - Frekvence slov
  - Charakterističnost slov pro češtinu (žížala…)

- Vysoká citlivost na přítomnost českého jazyka
- Nutná spolehlivá funkce i pro krátké texty
- Ruční nastavení vah, neumí se učit

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Lemmatizace textu (1)

- Lemmatizátor Lingea
  - Zaměření na spisovnou češtinu
  - Více správných možností => dvě lemmata (citron a citrón)
  - Názvy, jména, příjmení…
- Lemmatizace pouze podstatných a přídavných jmen

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Lemmatizace textu (2)

- Neznámá slova
  - Indexace termu
- Nejednoznačná slova
  - Zaindexování ve prospěch všech lemmat
  - Problémy: „Jana Králová“
- Text bez diakritiky
  - Doplňuje se pokud je celý odstavec bez diakritiky
  - Výskyt hlavně v diskuzích

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Určení podobnosti dokumentů (1)

- Zkoumá se podobnost každého dokumentu se všemi ostatními dokumenty

- Pro každý dokument se spočítá *N* hashů
  - *N* hashovacích funkcí
  - Hashe se počítají pro plovoucí okno *m*-tic slov
  - Výstupem je max(hash) z každé funkce

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Určení podobnosti dokumentů (2)

- Podobné dokumenty jsou takové, kde *X* z *N* hashů je stejných

- Výstupem výpočtu je *X* souborů s identifikovanými podobnostmi
- Bohužel neplatí transitivita
  - Pokud existuje A ~ B, B ~ C
  - Pak neplatí, že A ~ C

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Určení podobnosti dokumentů (3)

- Ze souborů vytvoříme třídy podobných dokumentů jako kdyby podobnost byla transitivní

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Doplňování diakritiky (1)

- Účel
  - Stále nezanedbatelné procento dotazů bez diakritiky
  - Nesprávně psané dlouhé samohlásky (á, é, í, ó) (Slávia, ale myslí Slavia)
  - Překlepy
    (Škoda místo Škoda)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Doplňování diakritiky (2)

- Postup
  - Úplné odstranění diakritiky z dotazu
  - Doplnění diakritiky do očištěného dotazu
  - Nalezení možných lemmat pro doplněná slova
    => Zvětšení nejednoznačnosti
    => Nutný výběr

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Výběr vhodného lemmatu (1)

- Pro každé slovo dotazu existuje *N* možných lemmat
- Musíme vybrat jedno lemma pro každé slovo
  - Výkon: nelze hledat všechna slova
  - Smysl dotazu: musíme určit slovo které uživatel myslel

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Výběr vhodného lemmatu (2)

- Preference lemmat ze znalosti situace
- Typická podoba dotazů
  - Drtivá většina podstatná a přídavná jména (dovolená chorvatsko)
  - Slovesa vyjímečně (běží liška k táboru)
  - 1. pád jedn. i mn. čísla (krby)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Výběr vhodného lemmatu (3)

- Použití frekvenčního slovníku dvojic lemmat
  - Kombinace všech možností lemmatizace dotazu
  - Ohodnocení každé kombinace podle frekvence výskytu
- Výsledek – rozdílná lemmata pro:
  - Německých tancích
  - Německých lidových tancích

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Výběr vhodného lemmatu (4)

slovo (doplněná diakritika) #ID	lemma	slovní druh	word-barel		
			svazek	typ	velikost [KB]
nemeckych (německých) #1176	německý	přídavné jméno	1	velký	1 522,4
			2	velký	1 513,7
			3	velký	1 541,1
			4	velký	1 163,0
tancich (tancích) #8793	tank	podstatné jméno	1	velký	356,2
			2	velký	327,3
			3	velký	251,1
			4	velký	196,6

- Výsledek z administračního rozhraní pro „Německých tancích“

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Výběr vhodného lemmatu (5)

slovo (doplněná diakritika) #ID	lemma	slovní druh	word-barel		
			svazek	typ	velikost [KB]
nemeckych (německých) #1176	německý	přídavné jméno	1	velký	1 522,4
			3	velký	1 541,1
			4	velký	1 163,0
lidovych (lidových) #1464	lidový	přídavné jméno	1	velký	1 706,4
			3	velký	354,1
			4	velký	449,3
tancich (tancích) #22810	tanec	podstatné jméno	1	velký	686,4
			3	velký	599,4
			4	velký	542,6

- Výsledek z administračního rozhraní pro „Německých lidových tancích“

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Hodnocení relevance (1)

- Vyhodnocuje se pozice a vzdálenost slov
- Hledá se uspořádání slov co nejpodobnější dotazu

- Příklady (dotaz A B):
  - A B vzdálenost 0
  - A x B vzdálenost 1
  - A x x B vzdálenost 2
  - B A vzdálenost 1

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hodnocení relevance (2)

- V textu musejí být všechna slova
- Hodnocení slov v:
  - Titulku
  - URL
  - Textu

- Hledání složeného slova v URL (www.jakpsatweb.cz)

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Hodnocení relevance (3)

- Celková relevance dokumentu
  - Nelineární kombinace částečných relevancí všech nalezených výskytů
  - Připočítání ranku dokumentu
- Výsledek seřazen podle celkové relevance

- Ze skupiny podobných dokumentů je vybrán ten nejrelevantnější

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Výkonové statistiky

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

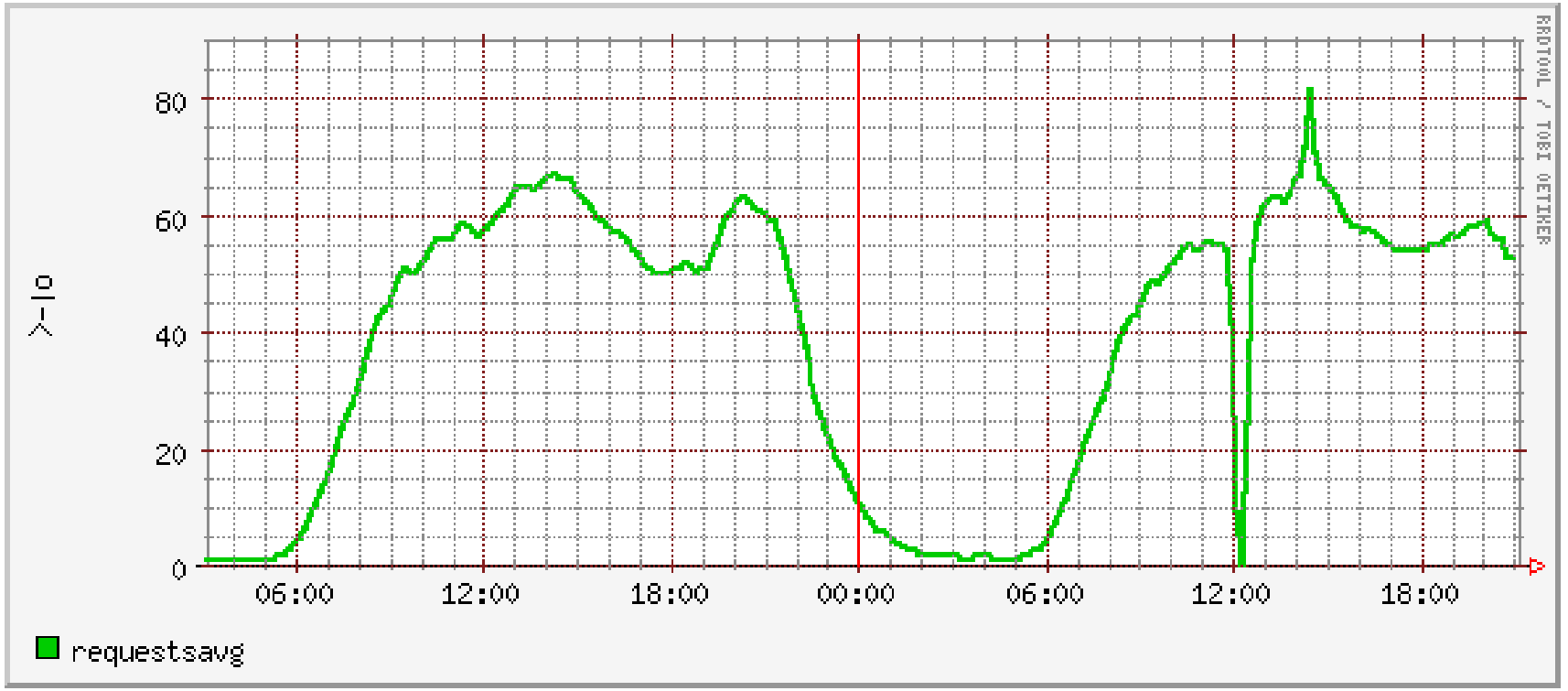
Výkonové statistiky

1. Dotazy na web interface
2. Úspěšnost query cache
3. Odezva index searcheru
4. Odezva content serveru

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

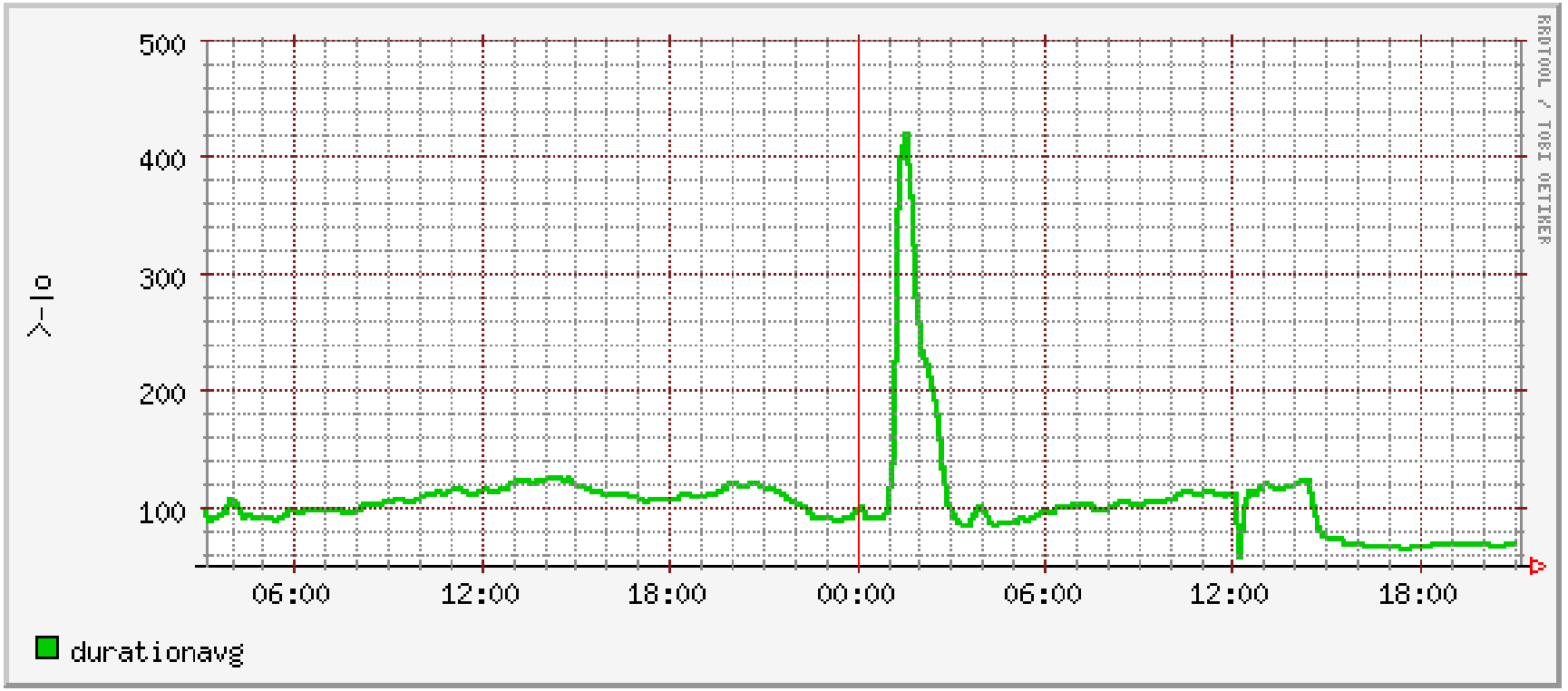
Dotazy na web interface (1)



- Počet dotazů / sec
- 1 ze 3 metasearchů

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

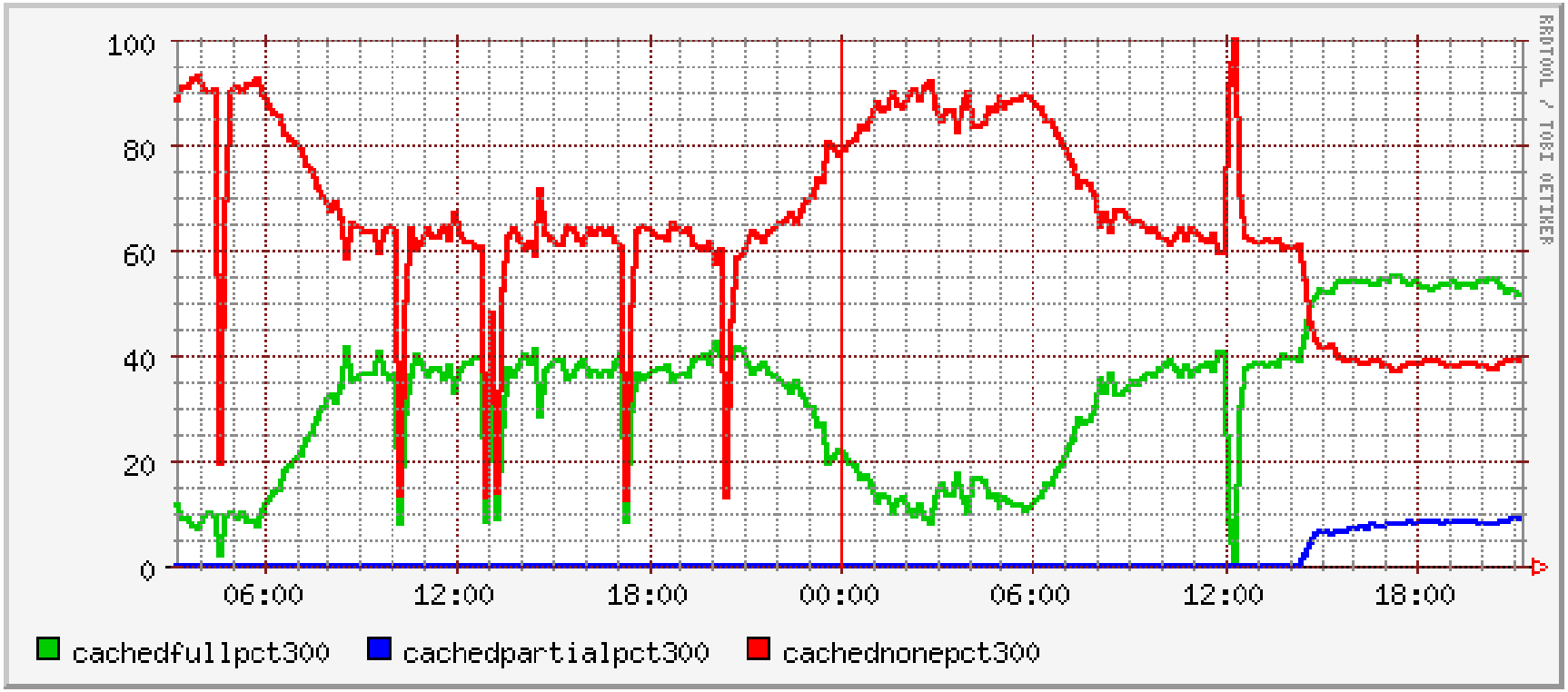
Dotazy na web interface (2)



- Průměrná latence odpovědí v msec
- Špička = update db, pokles = zapnutí cache

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

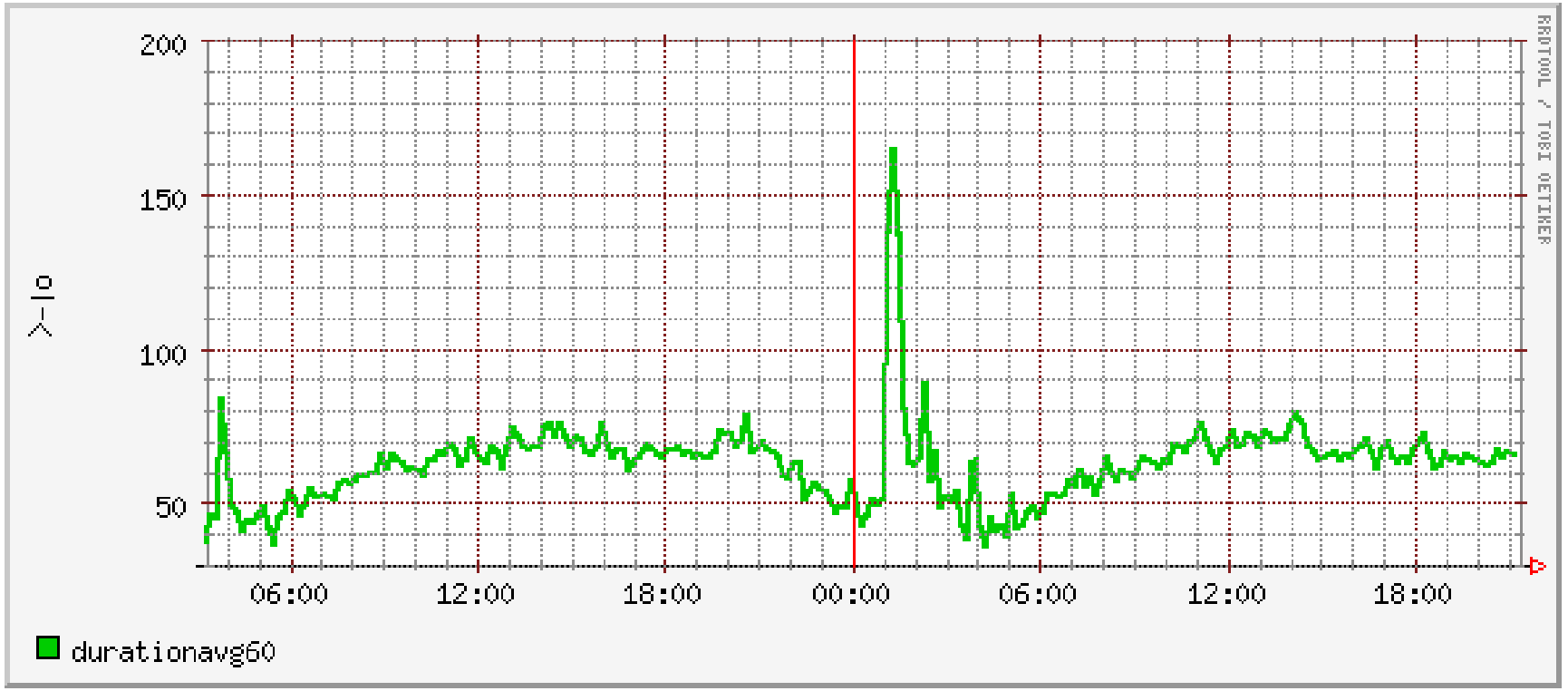
Úspěšnost query cache



- Úspěšnost cache v %
- Skok na konci = zapnutí částečného cachování

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

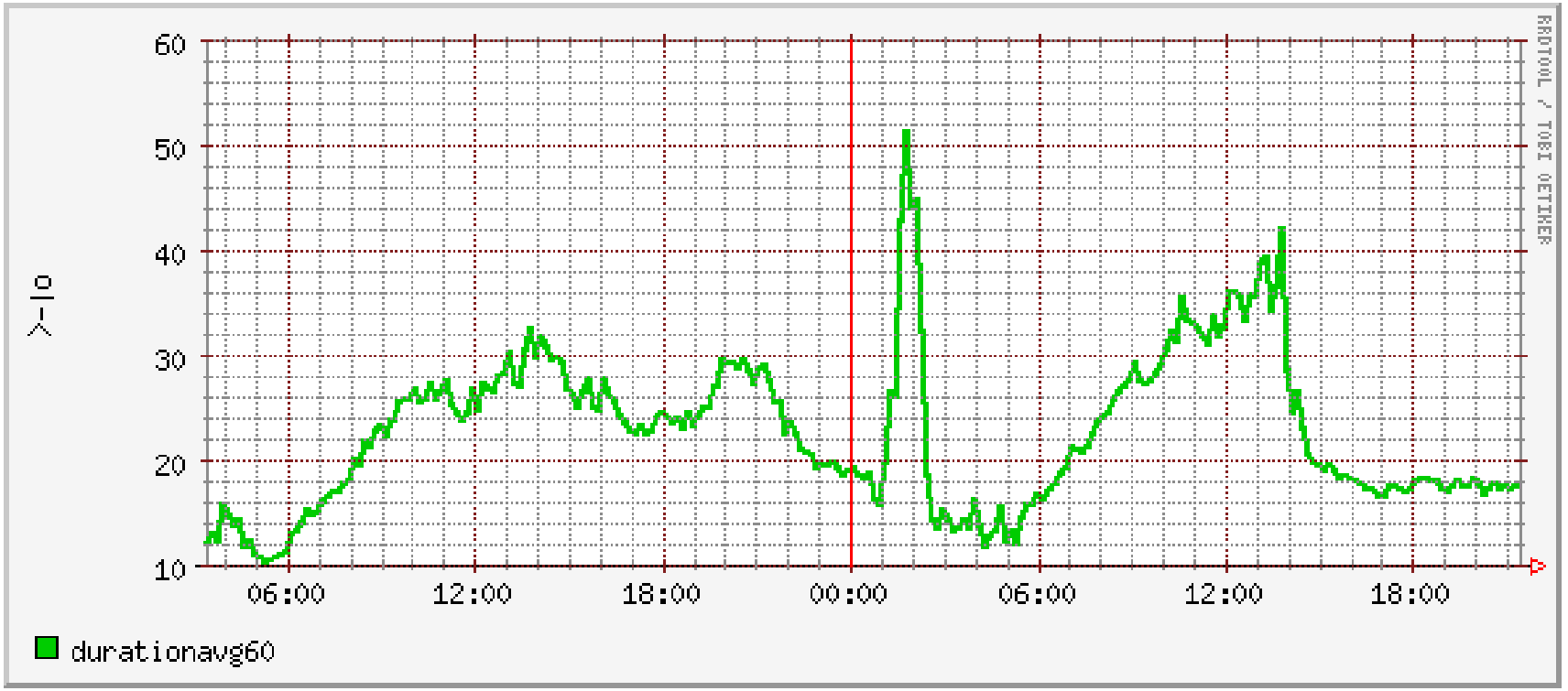
Odezva index searcheru



- Průměrná latence odpovědi v msec
- Špička = update db

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Odezva content serveru



- Průměrná latence odpovědi v msec
- Pokles na konci = zapnutí cachování

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

Další rozvoj

- Zapojení pokročilých lingvistických metod
- Zvyšování počtu dokumentů v databázi
- Zlepšování výpočtu relevance

- … Zde je šance pro Vás!

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115

SEZNAM

Konec

Děkuji za pozornost.

www.seznam.cz
Seznam.cz, a.s. I Radlická 2 I 150 00 Praha 5 I Tel.: +420 234 694 111 I Fax: +420 234 694 115