Třídění webových dokumentů v reálném čase

Autor: Jan Hošek
Školitel: RNDr. Radim Řehůřek

Fakulta jaderná a fyzikálně inženýrzká
České vysoké učení technické v Praze

25. 5. 2009

Osnova

Motivace

Problém

- Vyhledávače vrací výsledky jako dlouhý seznam seřazený podle relevance
- Uživatelé často kladou příliš obecné dotazy
- První subjektivně vyhovující dokument může být v seznamu velmi daleko

Řešení

- Nalézt v seznamu výsledků skupiny podobných dokumentů a spojit je do shluků
- Uživatel nemusí procházet celý seznam, ale je navigován do skupiny dokumentů, která ho zajíma
- Seznam dokumentů lze přeskupit (proložit) tak, aby na začátku byli zástupci ze všech skupin.

Ukázka technologie Vivisimo

Ukázka technologie Vivisimo

Top **181** results of at least **216,900** retrieved for the query **Czech Technical University** (details)

Search Results

1. **Czech Technical University** in Prague
   Contact information, basic facts, management and campus.
   www.cvut.cz - [cache] - Live, Open Directory, Ask, Gigablast

2. **Czech Technical University** Prague Center For Machine Perception
   Center for Machine Perception. The Center for Machine Perception (CMP) is a research unit focussing on computer vision, pattern recognition, and mathematics.
   cmp.felk.cvut.cz - [cache] - Live, Open Directory, Ask, Gigablast

3. Department of Cybernetics, **Czech Technical University** in Prague
   Research is conducted in two centers: (a) Gerstner Laboratory for Intelligent Decision Making Knowledge Based Systems, Multi-Agent Systems, Machine Learning, Mobile Robotics and (b ...
   cyber.felk.cvut.cz - [cache] - Live, Ask, Open Directory, Gigablast

4. **Czech Technical University**
   Department of Mathematics. Includes information about the faculty, staff list, courses and graduate study.
   math.feld.cvut.cz - [cache] - Live, Open Directory, Ask, Gigablast

5. **Czech Technical University**, Prague
   Department of Computer Science and Engineering, Faculty of Electrical Engineering.
   cs.felk.cvut.cz - [cache] - Open Directory, Gigablast

Ukázka technologie Vivisimo

Ukázka technologie Vivisimo

clusters sources sites

remix

All Results (185)

- Czech Technical University in Prague (70)
- Faculty of Electrical Engineering (21)
- Czech Technical University, Prague (21)
- Conference (13)
- Faculty of Civil Engineering (13)
- Laboratory, Cadence (9)
- ČVUT (7)
- Mathematics, Department (5)
- Liberec (6)
- Experimental (4)

more | all clusters

find in clusters: Find

Top **181** results of at least **216,900** retrieved for the query **Czech Technical University** (details)

Search Results

1. **Czech Technical University** in Prague
   Contact information, basic facts, management and campus.
   www.cvut.cz - [cache] - Live, Open Directory, Ask, Gigablast
2. **Czech Technical University** Prague Center For Machine Perception
   Center for Machine Perception. The Center for Machine Perception (CMP) is a research unit focussing on computer vision, pattern recognition, and mathematics.
   cmp.felk.cvut.cz - [cache] - Live, Open Directory, Ask, Gigablast
3. Department of Cybernetics, **Czech Technical University** in Prague
   Research is conducted in two centers: (a) Gerstner Laboratory for Intelligent Decision Making Knowledge Based Systems, Multi-Agent Systems, Machine Learning, Mobile Robotics and (b ...
   cyber.felk.cvut.cz - [cache] - Live, Ask, Open Directory, Gigablast
4. **Czech Technical University**
   Department of Mathematics. Includes information about the faculty, staff list, courses and graduate study.
   math.feld.cvut.cz - [cache] - Live, Open Directory, Ask, Gigablast
5. **Czech Technical University, Prague**
   Department of Computer Science and Engineering, Faculty of Electrical Engineering.
   cs.felk.cvut.cz - [cache] - Open Directory, Gigablast

Suffix Tree Clustering

Shlukování na základě shodných frází

- STC - Suffix Tree Clustering
- Algoritmus shlukování pomocí suffixového stromu popsal O. Zamir[1]
- Hlavním úkolem této práce je implementovat a otestovat tento algoritmus pro české texty

Možná úskalí českého jazyka

- Bohaté tvarosloví
- Volnější slovosled
- Psaní bez diakritiky

Postup algoritmu

Předzpracování textu

- Odstranění zvláštních znaků, převod na malá písmena
- Rozdělení na věty a na slova
- Odstranění předpon, přípon, množného čísla atd.

Nalezení základních shluků

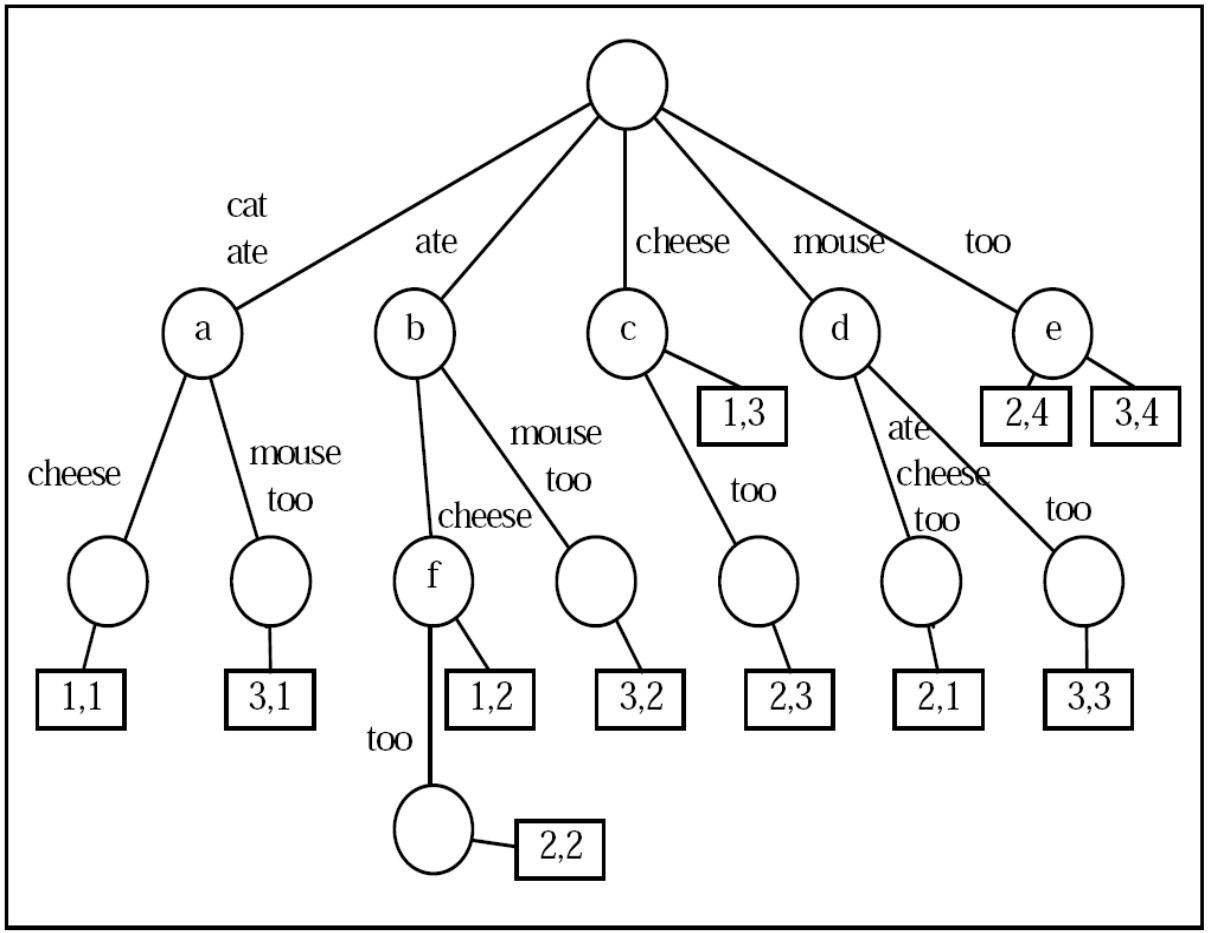
- Vygenerování suffixtree
- Nalezení základních shluků a jejich ohodnocení

$$s(B) = |B| \cdot f(|P|)$$

- $|B|$ - počet dokumentů ve shluku B
- $|P|$ - počet slov ve frázi P
- Z hodnocení fráze vynecháme příliš častá a příliš řídká slova

Suffixtree

„cat ate cheese“, „mouse ate cheese too“, „cat ate mouse too“



Postup algoritmu

Předzpracování textu

- Odstranění zvláštních znaků, převod na malá písmena
- Rozdělení na věty a na slova
- Odstranění předpon, přípon, množného čísla atd.

Nalezení základních shluků

- Vygenerování suffixtree
- Nalezení základních shluků a jejich ohodnocení

$$s(B) = |B| \cdot f(|P|)$$

- $|B|$ - počet dokumentů ve shluku B
- $|P|$ - počet slov ve frázi P
- Z hodnocení fráze vynecháme příliš častá a příliš řídká slova

Postup algoritmu

Kombinace základních shluků

- Dokumenty mají často více společných frází
- Na množině shluků definujeme ekvivalenci:

$$B_n \equiv B_m \Leftrightarrow (|B_m \cap B_n|/|B_m| > 0,5) \wedge (|B_m \cap B_n|/|B_n| > 0,5)$$

- Třídy této ekvivalence jsou výsledné shluky

Složitost STC

- Předpokládáme n dokumentů na vstupu
- Velikost dokumentu je omezená
- Tedy i hloubka suffixového stromu je omezená
- Časová složitost výpočtu základních shluků je $O(n)$
- Kombinování shluků je teoreticky $O(n^2)$, ale když se omezíme na porovnávání ostatních shluků vůči k nejvýše hodnoceným je složitost $O(n)$
- Celkem $O(n)$

Shrnutí

STC

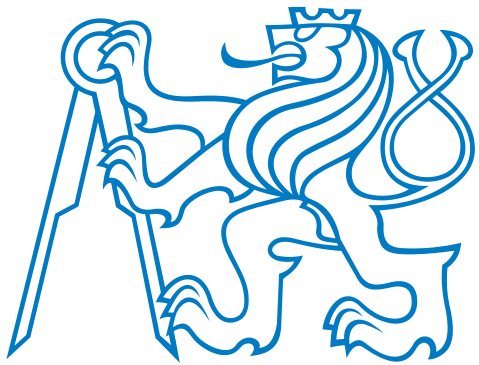
- $O(n)$ algoritmus pro hledání shluků v posloupnosti dokumentů
- Implementace v jazyce Python

Úpravy pro české dokumenty

- Pro výpočet četnosti slov jsem použil převod na základní tvar (*lemmatizace*)
- V hodntící funkci fráze je vhodné penalizovat zájmena, předložky, spojky, částice a citoslovce

Reference

- Zamir, O. and Etzioni, O., *Web document clustering: a feasibility demonstration*
- Vivisimo, *http://clusty.com/*



Děkuji za pozornost