

# Štatistické modelovanie hypertextu

## Diplomová práca

**Bc. Ján Suchal**

Fakulta informatiky a informačných technológií  
Slovenská technická univerzita v Bratislave

Vedúca práce: Mgr. Gabriela Kosková PhD.

# Obsah prezentácie

## 1 Technická prezentácia

- Motivácia
- Zber a príprava dát
- Opis metódy zhukovania
- Opis metriky kvality zhukovania
- Porovnanie s iným zhukovacím algoritmom
- Zhrnutie

## 2 Používateľská prezentácia

- Typický priebeh zhukovania
- Ukážka vizualizácie výsledkov
- Zhodnotenie

# Motivácia a ciele

- Motivácia

- potreba automatickej organizácie veľkého množstva informácií
- zhukovanie dokumentov na základe významu
- použitie hypertextu namiesto textu

- Ciele

- vytvoriť nástroj na zhukovanie hypertextových dokumentov
- vytvorený nástroj formou experimentov overiť
- porovnať štatistický a štandardný zhukovací algoritmus

# Zber a príprava dát

- Zdroj: <http://en.wikipedia.org/>
  - kvalitné a dobre previazané dokumenty
  - jednotné formátovanie
- Zberač a transformátor
  - implementovaný v jazyku PHP
- Úložisko údajov
  - relačná databáza MySQL

# Zberač

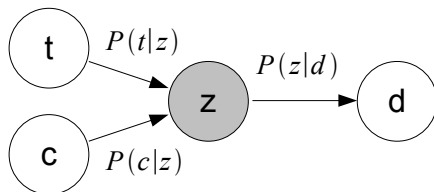
- Inkrementálne sťahovanie stránok
- Ukladanie nových nespracovaných odkazov
- Len obsahové stránky v rámci Wikipédie
  - prefix `en.wikipedia.org/wiki/`
- Výsledok zberu
  - podmnožina stránok Wikipédie

# Transformátor

- Odstránenie nepodstatných častí stránok
  - navigačné lišty, formátovanie, ...
- Transformácia slov na korene slov
  - Porter stemmer algoritmus
- Váhovanie slov podľa HTML značiek
  - title, h1 až h6, em, strong, a
- Výsledky transformácie
  - invertovaný index dokumentov
  - tabuľka prepojení medzi dokumentami

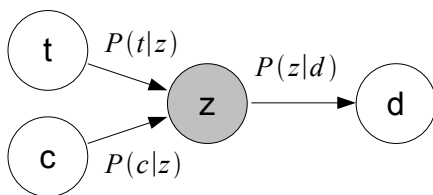
# Pravdepodobnostná latentná sémantická analýza

- Z invertovaného indexu  $P(t|d)$  a  $P(c|d)$



- $P(t|d) = \sum_z P(t|z)P(z|d)$      $P(c|d) = \sum_z P(c|z)P(z|d)$
- Parametre  $P(t|z)$ ,  $P(c|z)$  a  $P(z|d)$  počítané iteračným EM algoritmom na maximalizáciu vierohodnosti

# Zhlukovanie



- $P(z|d)$  zohľadňuje ako slová dokumentov, tak aj odkazy medzi nimi
- $P(z|d)$  určuje do akej miery daný aspekt popisuje daný dokument
- Najvyššie  $P(z|d)$  pre daný dokument určuje zhuk, do ktorého dokument patrí
  - dominantný aspekt

# Metrika kvality zhľukovania

- Nevýhoda použitého zhľukovania
  - empirické určovanie počtu hľadaných aspektov
- Ideálny výsledok zhľukovania
  - každý dokument opísaný len jedným aspektom
- Entropia  $E_i = \sum_{j=1}^k P(z_j|d_i) \log P(z_j|d_i)$ 
  - klesá pri jednoznačnejšom opise dokumentu aspektami
- Kumulatívna entropia
  - suma entropií pre všetky dokumenty

# Porovnanie s iným zhlukovacím algoritmom

- Zvolený algoritmus – k-means
  - jednoduchý princíp, dobre známe výsledky
- Upravený k-means algoritmus
  - berie do úvahy aj prepojenia medzi dokumentami
- Problémy k-means zhlukovania
  - nezvláda vysokú dimenzionalitu textových vstupov
  - nerozlišuje váhu slov v rámci jednotlivých zhlukov
- Dôsledok
  - vzdialenostné zhlukovacie algoritmy nedokážu riešiť vysokodimenzionálne problémy

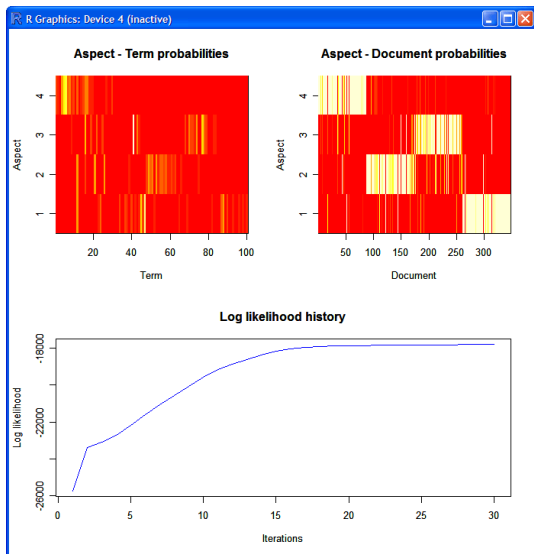
# Výsledky práce

- Zozbieraná a predpripravená množina hypertextových dokumentov
  - 4000 dokumentov, 100 000 koreňov slov
  - 150 000 odkazov, 600 000 prepojení
- Vytvorený a overený softvérový balík na zhlukovanie hypertextových dokumentov
  - štatistické a k-means zhlukovanie
  - možnosť použiť hypertext alebo len čistý text
- Navrhnutá metrika merania kvality zhlukovania
  - kumulatívna entropia
- Porovnanie štatistického zhlukovacieho algoritmu s k-means algoritmom
  - identifikované slabé miesta algoritmov založených na vzdialenostných metrikách

# Typický priebeh zhlukovania

- Vstupy
  - invertovaný index, tabuľka prepojení
  - matica výskytov
- Zhlukovanie
  - štatistické alebo k-means
  - možnosť zohľadniť prepojenia dokumentov
- Výstupy
  - pravdepodobnostné matice
  - zhluky dokumentov
- Analýza a vizualizácia výsledkov
  - výpis dominantných aspektov dokumentov
  - meranie kumulatívnej entropie
  - pravdepodobnostné mapy
  - priebeh maximalizácie vierohodnosti modelu

# Vizualizácia priebehu a výsledku zhlukovania



# Vlastnosti výsledného produktu

- Výhody
  - Silná kontrola nad parametrami zhlukovania
  - Jednoduchá rozšíriteľnosť
  - Použiteľnosť aj v iných doménach
- Nevýhody
  - Zložitosť ladenia vstupných parametrov
  - Pomalý výpočet pre veľké vstupy