

Diplomový projekt

Bc. Ján Suchal

Štatistické modelovanie hypertextu

Vedúca projektu:

Mgr. Gabriela Kosková, PhD.

Obsah prezentácie

- Motivácia a ciele
- Zber a príprava dát
- Opis metódy zhukovania
- Opis metriky na meranie kvality zhukovania
- Možné smery ďalšej práce
- Zhrnutie

Motivácia a ciele

- Motivácia

- Potreba automatickej organizácie čoraz väčšieho množstva informácií
- Zhlukovanie dokumentov na základe významu
- Použitie hypertextu namiesto textu

- Ciele

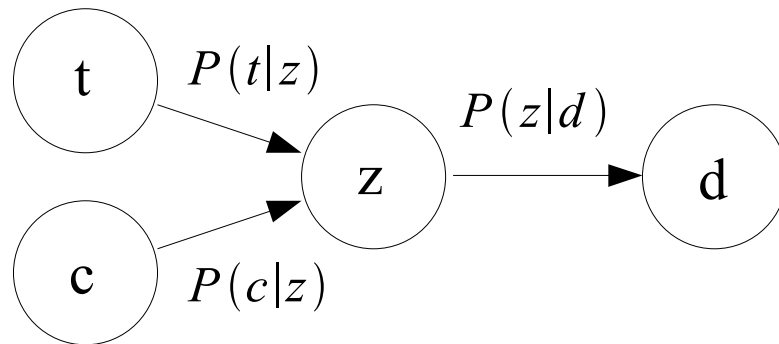
- Vytvoriť nástroj na zhlukovanie hypertextových dokumentov
- Vytvorený nástroj formou experimentov overiť na netriviálnej množine dát

Zber a príprava dát

- Zdroj: <http://en.wikipedia.org/>
- Získaná množina dokumentov
 - 4000 dokumentov, 100 tisíc koreňov slov
 - 150 tisíc odkazov, 600 tisíc prepojení
- Váhovanie slov podľa HTML značiek
 - `title`, `h1` až `h6`, `em`, `strong`, `a`
- Uložené ako invertovaný index

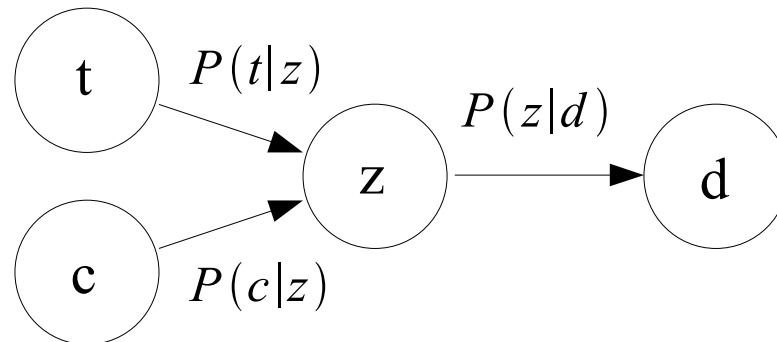
Pravdepodobnostná latentná sémantická analýza

- 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 $P(t|d)$, $P(c|d)$ a $P(z|d)$ vypočítame 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 zhluk, do ktorého dokument patrí

Metrika kvality zhukovania

- Nevýhoda použitého zhukovania
 - potreba empiricky určiť počet hľadaných aspektov
- Ideálny výsledok zhukovania
 - 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

Možné smery d'alsej práce

- Porovnanie s d'alšími zhlukovacími algoritmami
- Využitie výsledkov zhlukovania pri vyhľadávaní informácií
 - kolaboratívne filtrovanie, personalizácia vyhľadávania

Zhrnutie

- Zber množiny hypertextových dokumentov
- Vytvorený a overený softvérový balík na štatistické zhľukovanie dokumentov
- Navrhnutá metrika merania kvality zhľukovania
- 2 možné smery pokračovania v práci
 - porovnanie zhľukovacích algoritmov
 - využitie zhľukovania na vyhľadávanie informácií

Ďakujem za pozornosť