Extrakcia kľúčových pojmov z výučbových dokumentov

Jozef Harinek

Motivácia

- Veľký počet dokumentov
- Potreba efektívneho spracovania

- Cieľ - opis domény
- Zdroje
  - vzdelávacie dokumenty
  - anotácie

Stav poznania

- Tvorba doménového modelu
  - automatizovanie tvorby doménového modelu (Šimko, Bieliková, 2009) (Vráblecová, 2013)
  - Použitie folksonómií na tvorbu ontológií (Schmitz et al., 2006)
- Využitie používateľského obsahu (mimo doménového modelu)
  - na zlepšenie extrakcie RDT z webových zdrojov (Uherčík et al., 2010)
  - na predpovedanie udalostí, ako napr. zemetrasenie (Sakaki et al., 2010)

Problém

- Vylepšiť výsledky extrakcie kľúčových pojmov z výučbových dokumentov; zlepšiť výsledky ATR za použitia používateľských anotácií.

Návrh riešenia

- Základné typy anotácií
  - Tagy
  - Komentáre
  - Zvýraznenia
- Charakteristiky
  - Z textu dokumentu, dopísané používateľom
- Prínos
  - Každý typ anotácie rôzny (najviac tagy)

Návrh riešenia

- Výslednú množinu pojmov získať kombináciou váh RDT získaných z anotácií a z dokumentu

$$w(t,d) = (1-p) * w_{ATR}(d,t) + p * \ w_{annot}(d,t)$$

Návrh riešenia

- Výslednú množinu pojmov získať kombináciou váh RDT získaných z anotácií a z dokumentu

$$w(t,d) = (1-p) * w_{ATR}(d,t) + p * \ w_{annot}(d,t)$$

$$w_{annot}(t,d) = \sum_{a \in A} \sigma_a \ * \sum_{u \in U_t} UR(u) * w_{ATR}(t, E_d)$$

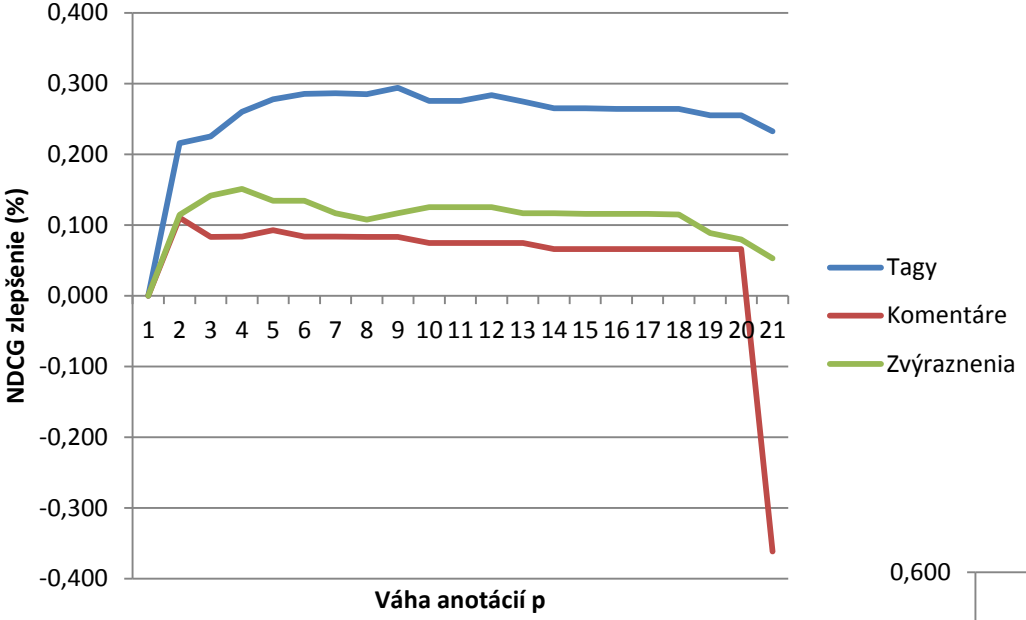
Implementácia a overenie

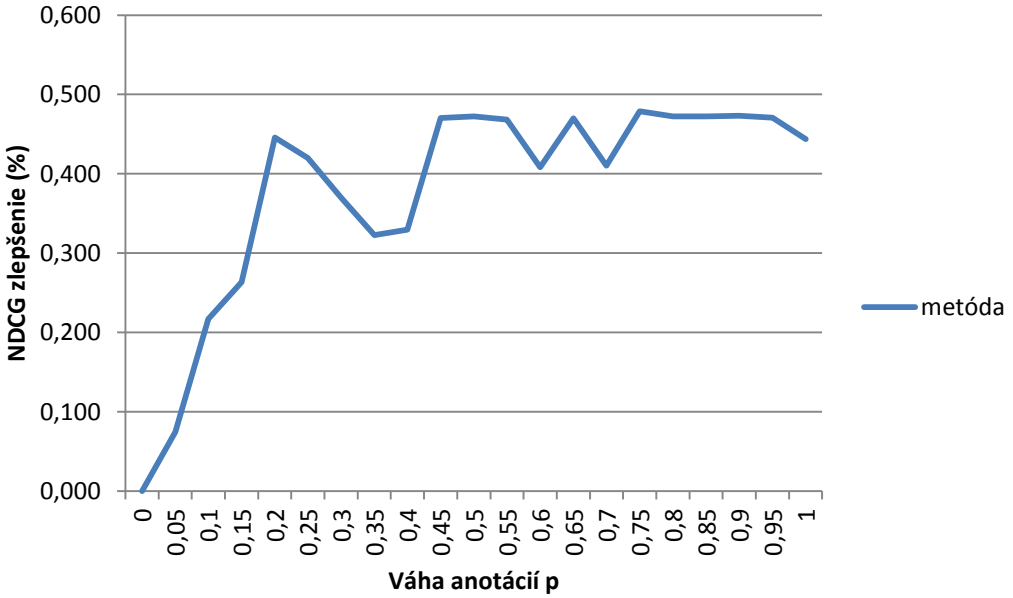
- Webová služba
  - Ruby on Rails
  - Komunikácia cez POST požiadavky (xml, json)
  - Integrácia do systému COME2T (ALEF)
- Overenie
  - Porovnanie so zlatým štandardom
  - A posteriori overenie
  - NDCG, Average Precision

Experiment (zlatý štandard)

- Vstup
  - 185 dokumentov, 21 491 anotácií, 329 študentov, 2 roky
- Výsledky (NDCG)
  - Zlepšenie: 29,4 % tagy 15,1 % zýraznený text, 11 % komentáre
  - Celkové zlepšenie: 47,9 %

Experiment (zlatý štandard)



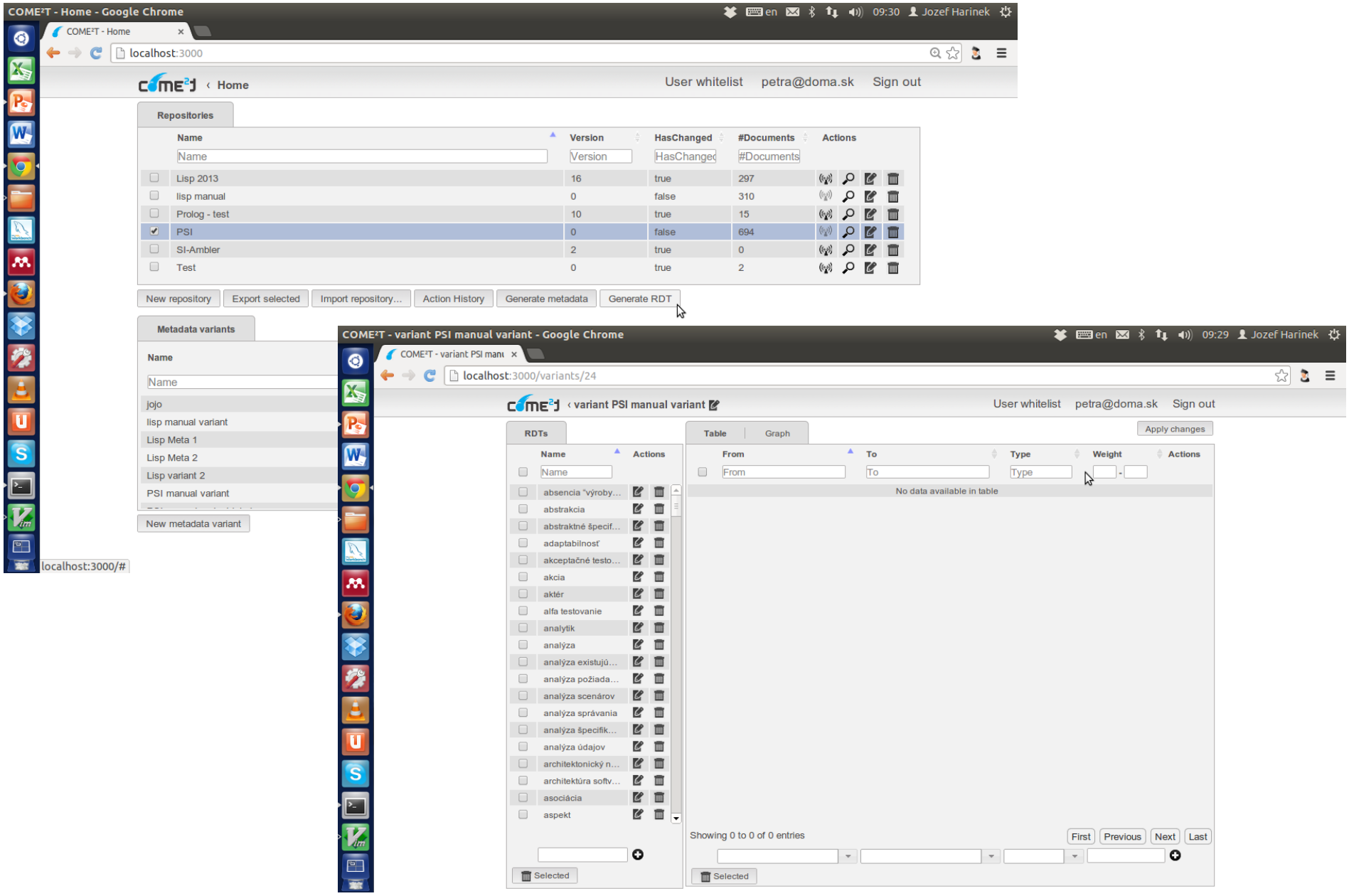


Experiment (A posteriori)

- Vstup
  - 33 dokumentov, top 15% vygenerovaných RDT, 5 expertov

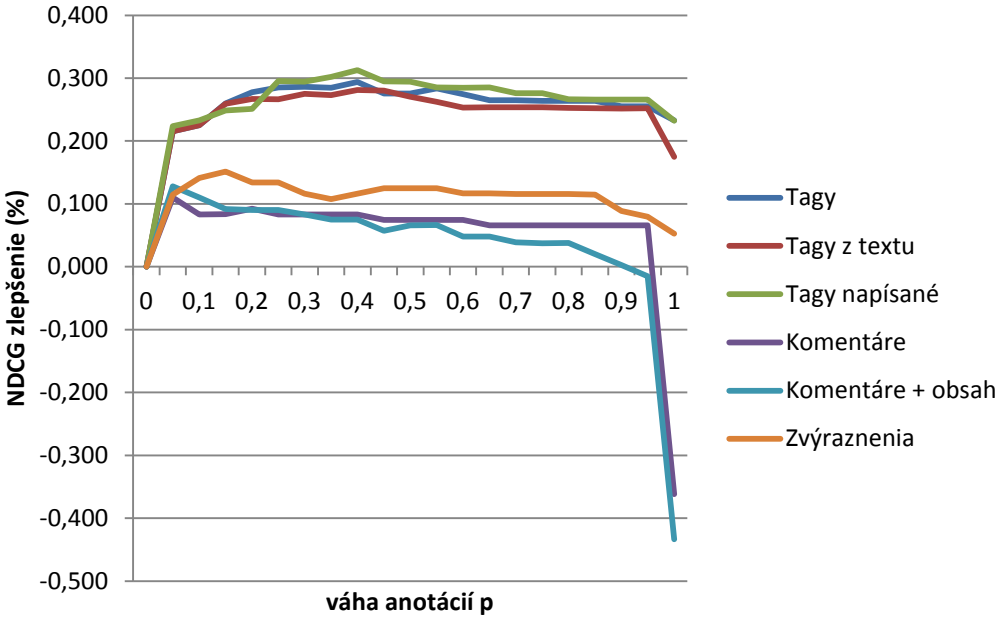
- Výsledky
  - NDCG = 0,914 (0,886) [3,2%];
    Average precision = 0,63 (0,57) [9,9%]
  - NDCG = 0,943 (0,943) [0 %];
    Average precision = 0,714 (0,666) [7,3 %]

COME2T

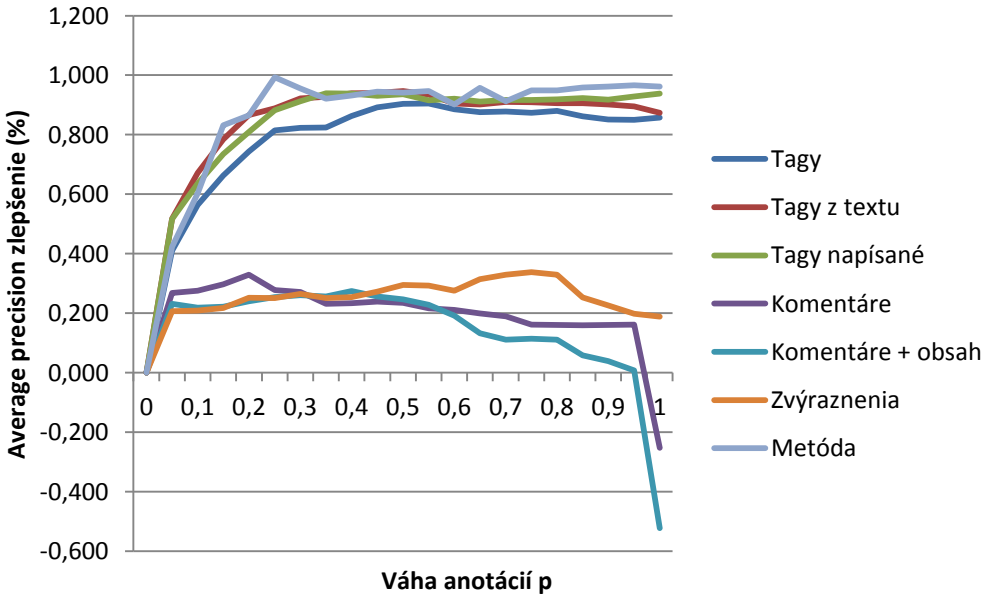

COME²T - Home - Google Chrome
COME²T - Home
localhost:3000
Home
User whitelist
petra@doma.sk
Sign out
Repositories
Name
Version
HasChanged
#Documents
Actions
Lisp 2013 16 true 297
lisp manual 0 false 310
Prolog - test 10 true 15
PSI 0 false 694
SI-Ambler 2 true 0
Test 0 true 2
New repository
Export selected
Import repository...
Action History
Generate metadata
Generate RDT
Metadata variants
Name
jojo
lisp manual variant
Lisp Meta 1
Lisp Meta 2
Lisp variant 2
PSI manual variant
New metadata variant
localhost:3000/#
COME²T - variant PSI manual variant - Google Chrome
localhost:3000/variants/24
variant PSI manual variant
User whitelist
petra@doma.sk
Sign out
RDTs
Table
Graph
Apply changes
From
To
Type
Weight
No data available in table
absencia "výroby...
abstrakcia
abstraktné špecif...
adaptabilnosť
akceptačné testo...
akcia
aktér
alfa testovanie
analytik
analýza
analýza existujú...
analýza požiada...
analýza scenárov
analýza správania
analýza špecifik...
analýza údajov
architektonický n...
architektúra softv...
asociácia
aspekt
Showing 0 to 0 of 0 entries
First
Previous
Next
Last
Selected

Zhrnutie

- Návrh a overenie metódy na extrakciu RDT s využitím používateľských anotácií
- Experimentálne overenie potenciálu používateľských anotácií pri automatickej tvorbe doménového modelu
- Zlepšenie 47.3 % (NDCG)
- Integrácia do COME2T
- Článok prijatý na publikovanie na DocEng’13



Absolútne hodnoty
NDCG = 0,6
Average Precision = 0,225



Vyjadrenie k posudku oponenta

- Konverzia na čistý text
- Chýba bibliografický záznam
- Hranica 20 %
- Použitie iných RDT algoritmov
- Tabuľka 2 -> cieľom bolo ilustrovať rôzne poradie, nie rôznu váhu RDT

Otázky od oponenta

1. Skúšali ste použiť iný frekvenčný interval na odfiltrovanie nedôležitých slov (iný od odfiltrovanie spodných 20 %)? Prečo ste nepoužili interval <N/100, N/10> ako uvádzate v analýze?

Otázky od oponenta

2. Prečo ste sa nepokúsili vyhodnotiť vašu metódu aj s využitím ďalších analyzovaných metód RDT? Potrebovali by ste k tomu ďalšie dáta? Ako by ste postupovali, keby ste chceli použiť vašu metódu s využitím metódy pravdepodobnostného pomeru?