

Obohacovanie položiek TV programu o metadáta

Peter Dubec - bakalárska práca
vedúci: prof. Mária Bieliková
2014/2015

Motivácia práce

- malé množstvo opisných metadát v slovenských televíznych programoch
- kvalita poskytovaných služieb priamo závisí od množstva a kvality dostupných metadát
- práca je súčasťou väčšieho projektu zameraného na odporúčanie položiek televízneho programu

Ciele práce

- preskúmanie dostupných zdrojov za účelom nájdania vhodných zdrojov metadát
- Obohatenie položiek televízneho programu so zámerom zvýšenia ich informačnej hodnoty a umožnenia personalizovaného odporúčania.

Získanie zdrojového televízneho programu

1. televízny program spoločnosti Orange
2. televízny program spoločnosti GoData
3. televízny program filmovej databázy CSFD

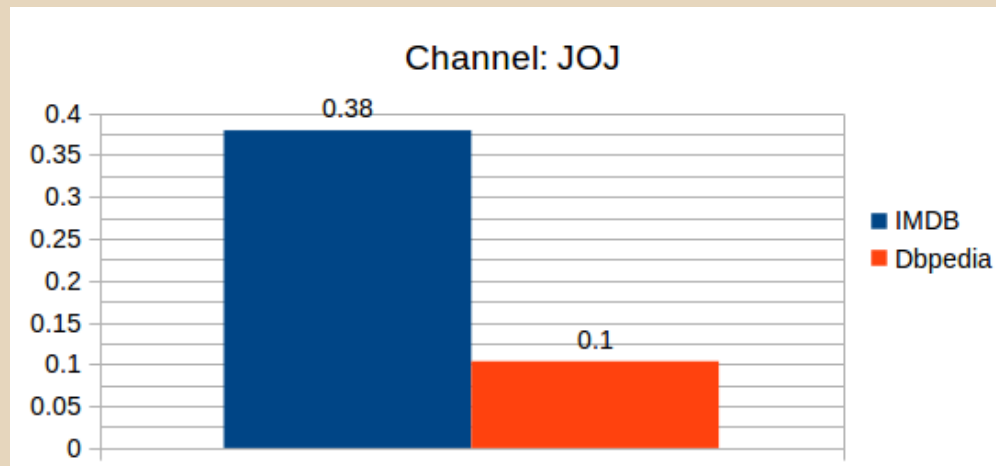
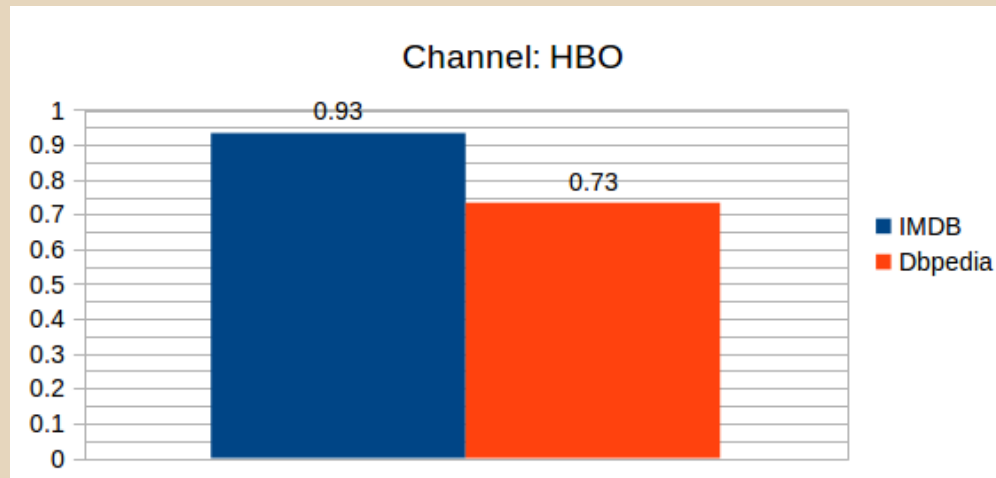
Realizované metódy získavania metadát

1. Extrakcia existujúcich metadát zo zdrojov dostupných na webe
 - online filmové databázy (CSFD, IMDB)
 - priestor Linked Open Data (DBpedia, FreeBase)
2. Tvorba nových metadát
 - automatická klasifikácia dokumentárnych filmov

Extrakcia existujúcich metadát

- extrakcia z online filmových databáz CSFD a IMDB
- priestor Linked Open Data (DBpedia, FreeBase)
 - identifikovaný problém veľmi nízkeho pokrytia

Pokrytie entít v skúmaných zdrojoch



Automatická klasifikácia dokumentárnych filmov

- klasifikácia dokumentárnych filmov na základe ich popisov v anglickom jazyku (SVM)
- klasifikácia do 30 kategórií
- vytvorenie datasetu

Vytvorenie trénovacej množiny

- uskutočnenie hromadného experimentu pre vytvorenie trénovacej množiny
- 18 participantov = 135 unikátnych položiek
- celkovo 350 unikátnych manuálne klasifikovaných položiek

Priebeh automatickej klasifikácie

- klasifikácia pomocou implementácie SVM v jazyku Ruby (LibSVM)
 1. Odstránenie stop slov a interpunkcie z popisu položiek
 2. Vytvorenie vektorovej reprezentácie položiek
 3. Natrénovanie klasifikačného modelu
 4. Automatická klasifikácia (určovanie pravdepodobností)

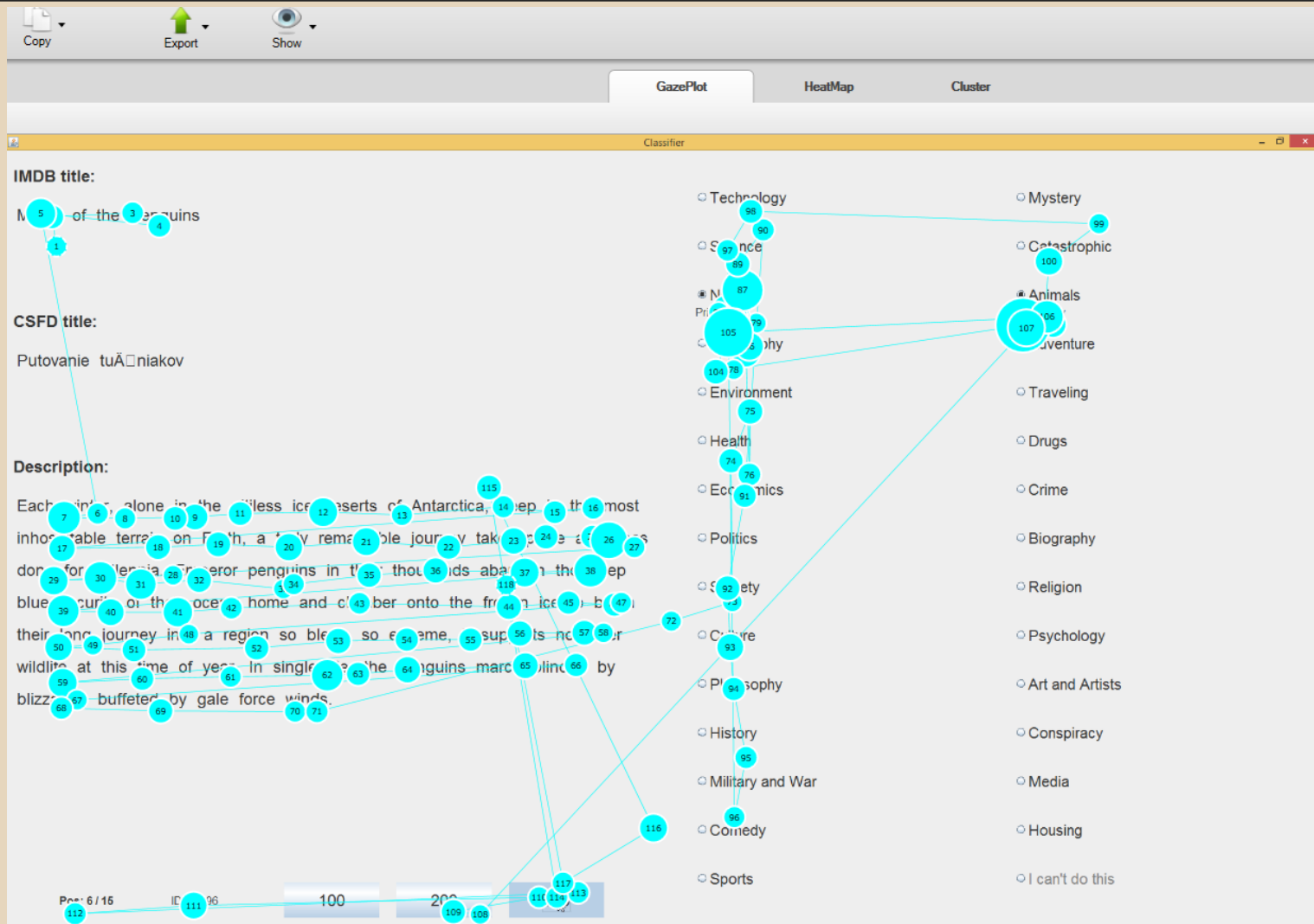
Overenie metódy automatickej klasifikácie

Poradie	Veľkosť trénovacej množiny	Správne % TOP 3	Správne % TOP 2	Správne % TOP1
1	617	57	43	27
2	614	70	50	30
3	633	57	40	17
4	618	70	53	33
5	641	60	47	27
6	656	33	30	17
7	653	47	43	23
Priemer	633	56%	44%	25%

Klasifikácia s využitím sledovania pohľadu

- klasifikácia pomocou využitia sledovania pohľadu
- uskutočnenie 2 experimentov zameraných na zber dát

Klasifikácia s využitím sledovania pohľadu



Klasifikácia s využitím sledovania pohľadu

Poradie	Veľkosť trénovacej množiny	Správne % TOP3	Správne % TOP2	Správne % TOP1
1	260	59	41	35
2	272	29	29	24
3	254	53	41	29
4	260	65	47	41
5	304	35	29	24
Priemer	270	48%	37%	31%

Zhodnotenie

- prieskum oblasti televíznych programov a zdrojov vhodných metadát
- metóda schopná automatickej extrakcie metadát
- vytvorenie datasetu dokumentárnych filmov
- vytvorenie datasetu manuálne klasifikovaných filmov
- metóda automatickej klasifikácie dokumentárnych filmov
- klasifikácia s využitím sledovania pohľadu

Doplňujúca otázka:

Ako by sa zmenila použitá metodika overenia v prípade, že by bolo snahou optimalizovať aj parametre použitej metódy strojového učenia, resp. porovnať rôzne metódy navzájom?