Diversita a diversifikace v doporučovacích systémech

Tereza Miklóšová

Úvod do doporučovacích systémů a uživatelských preferencí - NSWI166

Zdrojové články

- Efficient Diversification of Web Search Results (IASelect, xQuAD, OptSelect)
- The Use of **MMR**, Diversity-Based Reranking for Reordering
- Coverage, Redundancy and Size-Awareness
  - Záznam o představení tématu na konferenci
- Novelty and Diversity Enhancement and Evaluation in Recommender Systems
- Explicit Search Result Diversification through Sub-Queries (xQuAD)

Obsah

- Definice
- MMR
- xQUAD
- Další přístupy k řešení problému
- Shrnutí

Definice

- Liší se požadavky

- *Hledaný výraz může být nejednoznačný*
- Diversita představuje pokrytí všech možných významů
- Příklad:

„apple“

Definice

- *Hledaný výraz není dostatečně konkrétní*
- Diversita představuje nabídnutí různorodých položek

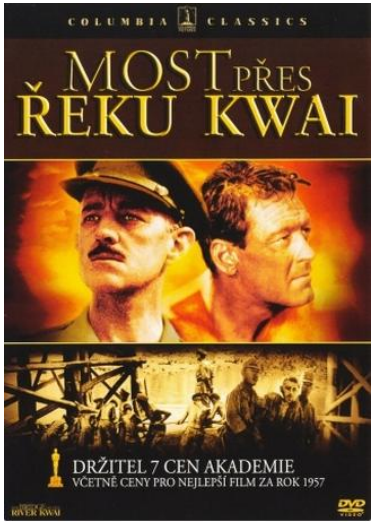
- Příklad:

„shirt“

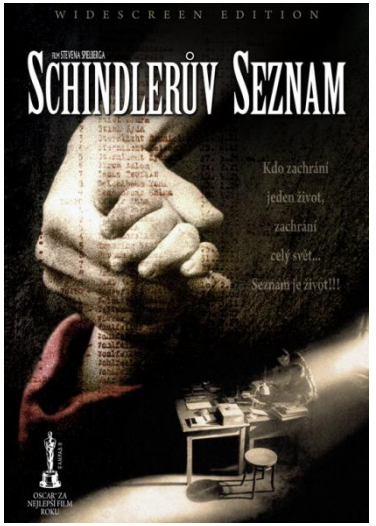
Definice

- Doporučovací systémy

romantický
komedie



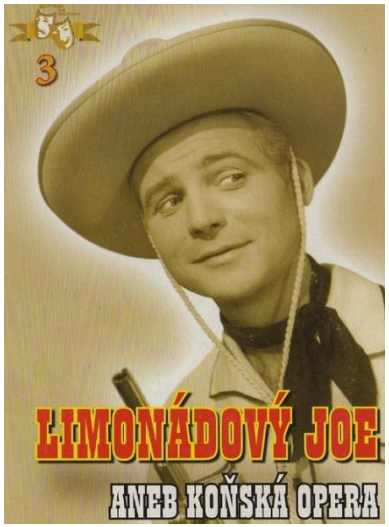
válečný
dobrodružný



biografický
historický

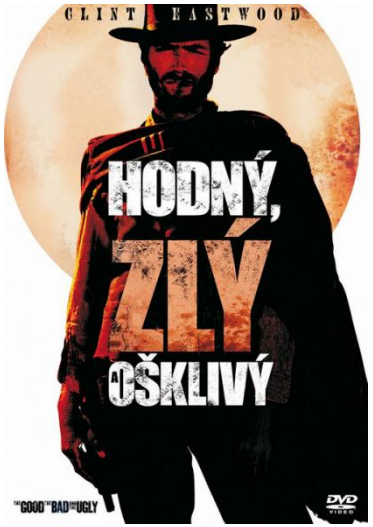
Definice

- Doporučovací systémy znající naše preference



western
komedie

sci-fi
western



western
dobrodružný

- Příliš mnoho westernů…

Finální definice

- Diversifikace výsledku hledání se provádí za účelem poskytnutí širokého pokrytí jednotlivých předmětů v dotazu a zároveň má za úkol snížit celkovou přebytečnost informace o předmětech, které již ve výsledném řazení pokryty jsou

- Diversita v doporučovacích systémech se zaměřuje na různé zájmy uživatele, pomáhá zvýšit atraktivitu a užitečnost doporučení

MMR – Maximal marginal relevance

- Vybraná položka je co nejrelevantnější vzhledem k dotazu, ale zároveň je co nejrozdílnější od těch, které už ve výsledku máme

$$MMR \stackrel{\text{def}}{=} Arg \max_{D_i \in R \setminus S} \left[\lambda (Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

- **Sim**… jakákoliv vhodná podobnostní metrika (Jaccard, cosine,…)
- D_i… dokument, který zvažujeme
- D_j… dokumenty již vybrané
- **λ**… [0,1] (**λ**=0 … maximální diversita)
- **S**…již vybrané dokumenty
- **R**…seřazený seznam vrácený z vyhledávacího systému podle **dotazu Q**

Využití MMR

- Hodnocení závislé na diversitě
- Shrnutí dokumentů
- Otázky a odpovědi (stackovrflow, fóra,…)

xQuAD

- e**X**plicit **Qu**ery **A**spect **D**iversification

- 1) důležitost předmětů (poddotazů) v dotazu
- 2) pokrytí dokumentu vzhledem k jednomu nebo více předmětu v dotazu
- 3) novota dokumentu vzhledem k již vybraným dokumentům ve výsledku
- 4) relevance dokumentu vzhledem k původnímu dotazu

xQuAD algoritmus

q… dotaz
R(q)…seznam dokumentů vrácený vzhledem k dotazu
Q(q)…množina poddotazů qi v q
r(d,q)…relevance dokumentu d vzhledem k dotazu q, stejně pro r(d, qi)
iX(qi,q)…funkce odhadující důležitost poddotazu qi v q
τ…počet dokumentů ve výsledném diversifikovaném pořadí
ω…operátor pro vyvažování mezi relevancí a diversitou výsledku
m(qi)… míra novoty dokumentu odpovídajícímu předmětu qi z dotazu

xQuAD$[q, R(q), Q(q), r, i_X, \tau, \omega]$

1 $S(q) \leftarrow \emptyset$
2 **while** $|S(q)| < \tau$ **do**
3 **for** $d \in R(q)$ **do**
4 $r(d, q, Q(q)) \leftarrow r(d, q) \times \left(\sum_{q_i \in Q(q)} i_X(q_i, q) r(d, q_i)/m(q_i)\right)^{\omega}$
5 **end for**
6 $d^* \leftarrow \arg\max_d \; r(d, q, Q(q))$
7 **for** $q_i \in Q(q)$ **do**
8 $m(q_i) \leftarrow m(q_i) + r(d^*, q_i)$
9 **end for**
10 $R(q) \leftarrow R(q) \setminus \{d^*\}$
11 $S(q) \leftarrow S(q) \cup \{d^*\}$
12 **end while**
13 **return** $S(q)$

Porovnání MMR a xQuAD z článku

Table 1: Comparative performance with a uniform aspect importance estimator.

	α-NDCG		u-MAP-IA		i-MAP-IA	
	@10	@100	@10	@100	@10	@100
BM25	0.4505	0.5308	0.2286	0.1710	0.1416	0.1969
+MMR	0.4364	0.5102	0.2289	0.1700	0.1380	0.1841
+IA-Select	0.3392	0.4141	0.1592	0.1141	0.0868	0.1271
+QFilter	0.4509	0.5200	0.2300	0.1856	0.1416	0.1934
+xQuAD$_U$	**0.5727**$^{\blacktriangle}_{\blacktriangle}$	**0.6120**$^{\triangle}_{\blacktriangle}$	**0.2760**	**0.2240**	**0.1825**	**0.2235**
DPH	0.4633	0.5476	0.2464	0.1827	0.1620	0.2134
+MMR	0.4087$^{\blacktriangledown}$	0.4273$^{\blacktriangledown}$	**0.2876**	**0.2422**	0.1479	0.1805
+IA-Select	0.3585	0.4340	0.1765	0.1318	0.1029	0.1403
+QFilter	0.4634	0.5342	0.2466	0.1947	0.1620	0.2103
+xQuAD$_U$	**0.5935**$^{\blacktriangle}_{\blacktriangle}$	**0.6151**$^{\triangle}_{\triangle}$	0.2871	0.2371	**0.1998**	**0.2424**

BM25, DPH…hodnotící funkce používané při vracení dokumentů k odhadu relevance dokumentu vzhledem k dotazu

α-NDCG…evaluační metrika, při větším α je více preferována diversita

u-MAP-IA…evaluační metrika, všechny poddotazy jsou stejně důležité

i-MAP-IA…evaluační metrika, odhaduje důležitost poddotazů

MAP-IA = intent-aware mean average precision NDCG = normalised discounted cumulative gain

Další diversifikační přístupy

- A Binomial Framework for Genre Diversity
  - The Binomial Diversity Metric
    - BinomDiv(R) = Coverage(R) · NonRed(R)
  - Coverage(R) … produkt žánrů nevybraných do doporučení, pravděpodobnost, že nebudou náhodně vybrány do výběru
  - NonRed(R) … zbytková tolerance pro nadbytečnost daného žánru, každý žánr má jinou toleranci – jak pravděpodobné by bylo pro žánr dostat se do náhodného seznamu k-krát

$$f_{BinomDiv}(i; S) = (1 - \lambda)\, rel(i) + \lambda\, div(i; S)$$

…diversita předmětu i z množiny S

$$div(i; S) = BinomDiv(S \cup \{i\}) - BinomDiv(S)$$

…když už v seznamu jsou nějaké doporučené předměty s předmětem i

Další diversifikační přístupy

- OptSelect
  - Diversifikuje vrácené dokumenty při vložení dotazu
  - Nižší časová složitost než xQuAD ($n \log_2 k$ x nk)

$$\widetilde{U}(S|q) = \sum_{d \in S} \sum_{q' \in S_q} (1-\lambda) P(d|q) + \lambda P(q'|q)\, \widetilde{U}(d|R_{q'})$$

Děkuji za pozornost