

Manažment znalostí (3)

OBSAH PREDNÁŠKY

- Vyhodnocovanie systémov pre vyhľadávanie informácií (IR)
 - Rôzne kritériá pre hodnotenie IR systémov
- Hodnotenie efektívnosti vyhľadávania (spokojnosť používateľa)
 - Presnosť, návratnosť, F miera – hodnotenie výsledkov bez usporiadania (unranked retrieval set)
 - Grafické znázornenie závislosti presnosti a návratnosti – hodnotenie usporiadaných výsledkov vyhľadávania (ranked retrieval set)
 - Sumarizačné miery efektívnosti vyhľadávania
- Prezentácia výsledkov vyhľadávania - sumáre

Vyhodnocovanie systémov pre vyhľadávanie informácií (2)

- Čo by sme ale radi merali u IR systému je miera spokojnosti jeho používateľov
 - Tú ovplyvňuje mnoho faktorov (rýchlosť odozvy, rôzne aspekty GUI, kvalita sumárov a pod.)
- Pod vyhodnotením IR systému ale najčastejšie rozumieme mieru relevancie množiny vrátených dokumentov k informačnej potrebe vyhľadávajúceho
- Pre daný systém IR (resp. jeho konkrétnu vyhľadávaciu stratégiu S) sa kvantifikuje
 - podobnosť medzi množinou dokumentov vyhľadaných pomocou S
 - a množinou dokumentov stanovených ako relevantné k danej otázke expertmi

Štandardné testovacie kolekcie pre IR

- Dátové kolekcie pripravované a spravované už dlhé roky NIST (*National Institute of Standards and Technology*)
 - TREC (*Text REtrieval Conference*) – 1,89 mil. dokumentov, 450 informačných potrieb (zvaných *topics* – špecifikovaných podrobnými pasážami textu)
 - Ohodnotenie top-k dokumentov vrátených nejakým IR systémom, nie všetkých dokumentov v kolekcii
 - Gov2 – 25 mil. web stránok, najväčšia testovacia kolekcia (stále však cca. 1000 krát menšia ako indexujú bežné vyhľadávače)
- Reuters-21578 a Reuters-RCV1 – kolekcie novinových článkov používané najmä pre klasifikáciu textov, podobne aj
- 20 Newsgroups – kolekcia Usenet news skupín (1000 článkov z každej)

Vyhodnocovanie systémov pre vyhľadávanie informácií (1)

- Merateľné kritériá pre porovnávanie IR systémov:
 - Rýchlosť indexovania (počet dokumentov/hodinu, pre určitú distribúciu veľkosti indexovaných dokumentov)
 - Rýchlosť vyhľadávania (aké je oneskorenie, ako funkcia veľkosti indexu)
 - Veľkosť kolekcie indexovaných dokumentov, resp. aj jej distribúcia medzi rôzne oblasti zamerania
- Trocha ťažšie môže byť merateľná napr. výrazová sila dopytovacieho jazyka, resp.
 - rýchlosť vyhľadávania pre zložité dopyty

2

Vyhodnocovanie efektívnosti vyhľadávania (1)

- Pre takéto vyhodnotenie IR systému teda potrebujeme:
 - Kolekciu dokumentov
 - Testovaciu množinu informačných potrieb (testovacie dopyty)
 - Spravidla binárne ohodnotenie dokumentov v kolekcii z pohľadu ich ne/relevancie ku každému dopytu z testovacej množiny dopytov
- Informačná potreba je teda najprv transformovaná do podoby dopytu, ale relevantnosť je meraná voči informačnej potrebe, nie voči dopytu
 - Příklad informačnej potreby: „Chcem zistiť či konzumácia červeného vína efektívnejšie znižuje riziko srdcového infarktu než konzumácia bieleho vína.“
 - Dopyt: „víno červené biele srdcový infarkt zníženie rizika“
 - Z jednoslovného dopytu je pre IR systém veľmi ťažké zistiť informačnú potrebu, ale používateľ nejakú vždy má („python“?)

4

Základné miery efektívnosti vyhľadávania pre neusporiadanú množinu výsledkov

- Nech:
 - q je daný dopyt reprezentujúci informačnú potrebu
 - R je množina relevantných dokumentov ku q
 - $|R|$ je počet relevantných dokumentov ku q
 - A je množina dokumentov, ktoré vyhľadávací systém používajúci stratégiu S vráti ako odpoveď na q
 - $|A|$ je počet dokumentov vrátených S ako odpoveď na q
 - R_A je prienik množín R a A

$$\text{Návratnosť (recall)} \quad N = \frac{|R_A|}{|A|} \quad \text{Presnosť (precision)} \quad P = \frac{|R_A|}{|R|}$$

6

Základné miery efektívnosti trochu ináč

	Relevantné dokumenty	Nerelevantné dokumenty
Dokumenty vrátené IR systémom na q	true positive (tp)	false positive (fp)
Dokumenty, ktoré neboli v odpovedi IR systému na q	false negative (fn)	true negative (tn)

$$\text{Návratnosť (recall)} \quad N = \frac{tp}{(tp + fn)}$$

$$\text{Presnosť (precision)} \quad P = \frac{tp}{(tp + fp)}$$

7

Prečo nie presnosť klasifikácie?

	Relevantné dokumenty	Nerelevantné dokumenty
Dokumenty vrátené IR systémom na q	true positive (tp)	false positive (fp)
Dokumenty, ktoré neboli v odpovedi IR systému na q	false negative (fn)	true negative (tn)

$$\text{Accuracy} \quad A = \frac{tp + tn}{(tp + fp + tn + fn)}$$

- Presnosť klasifikácie (accuracy) sa využíva pri klasifikácii, vyjadruje percento správne zaradených príkladov (v tomto prípade dokumentov)
- Pre IR ale nie je vhodná, lebo spravidla 99,9% dokumentov sú nerelevantné

8

Závislosť medzi presnosťou a návratnosťou (1)

- Treba si uvedomiť, že presnosť a návratnosť sú v *zásade protichodné požiadavky*, t.j. akýkoľvek IR systém sa musí snažiť o vhodný *kompromis medzi nimi*
- **Návratnosť** je neklesajúca funkcia počtu dokumentov vrátených na daný dopyt
- **Presnosť** naproti tomu klesá s pribúdajúcim počtom vrátených dokumentov aj v dobrom IS systéme
- Preto sa zvykne používať aj zložená *miera F*, ktorá práve popisuje kompromis medzi presnosťou a návratnosťou;

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

9

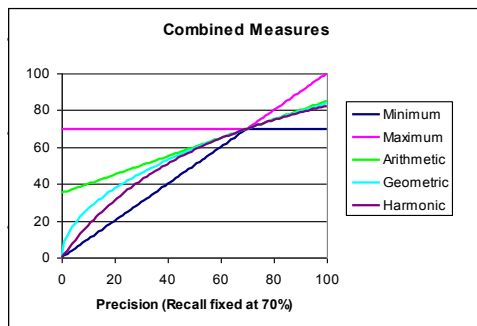
Závislosť medzi presnosťou a návratnosťou (2)

$$\beta^2 = \frac{1-\alpha}{\alpha} \quad \text{kde } \alpha \in [0,1] \Rightarrow \beta \in [0,\infty]$$

- Ak je pre nás rovnako dôležitá presnosť aj návratnosť, potom treba dať $\alpha = 0,5 \Rightarrow \beta = 1$
- $F_{\beta=1} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} = \frac{2PR}{P + R}$
- Hodnoty $\beta < 1$ zdôrazňujú presnosť, hodnoty $\beta > 1$ zase návratnosť
- Presnosť, návratnosť a F miera majú rozsah hodnôt $[0,1]$, ale môžu sa uvádzať aj v percentách

10

F_1 a ďalšie možné odvodené mierky



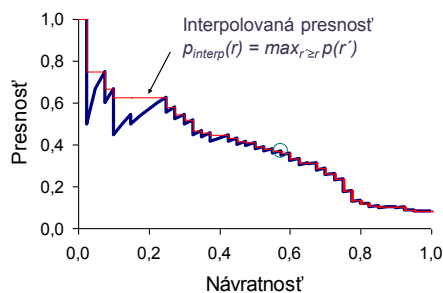
11

Hodnotenie usporiadaných výsledkov vyhľadávania

- Avšak používateľ obvyčajne nevidí celú množinu A (odpoveď na svoj dopyt q) naraz, ale postupne, dokumenty sú usporiadané podľa stupňa relevancie (*ranked retrieval set*)
- Teda návratnosť a presnosť sa z pohľadu používateľa postupne menia
- Priebeh presnosti, ako funkcie závislej od návratnosti sa zvykne zobrazovať graficky
- Ide o tzv. **krivku presnosť – návratnosť**

12

Krivka presnosť – návratnosť



13

Príklad (1)

- Nech množina všetkých relevantných dokumentov R na dopyt q_1 je nasledovná:
 $R_1 = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- Odpoveď vyhľadávacej stratégie S obsahuje túto postupnosť dokumentov:

1. d_{123}	6. d_9	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56}	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25}	15. d_3
- Zostrojte krivku presnosť – návratnosť pre danú vyhľadávaciu stratégiu S a otázku q_1

14

Príklad (1)

Normovaná krivka presnosť - návratnosť

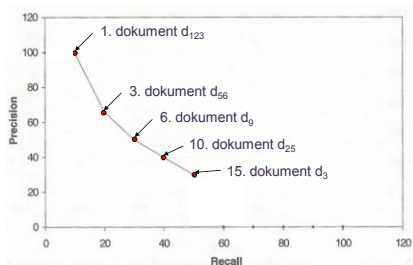


Figure 3.2 Precision at 11 standard recall levels.

15

Normovaná krivka presnosť - návratnosť

- Definuje sa 11 normovaných úrovni návratnosti: 0%, 10%, 20%, ..., 100%
- Nech $r_j, j \in \{0, 1, 2, \dots, 10\}$ je referencia na j -tu štandardnú úroveň návratnosti (napr. r_5 je úroveň návratnosti 50%)
- Potom hodnota $P(r_j)$ sa interpoluje podľa vzorca:

$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$
- čo znamená že interpolovaná presnosť na j -tej štandardnej úrovni návratnosti je maximálna známa presnosť pre ľubovoľnú návratnosť z intervalu $\langle r_j, r_{j+1} \rangle$

16

Príklad (2)

- Nech množina všetkých relevantných dokumentov R_2 na dopyt q_2 je nasledovná:
 $R_2 = \{d_3, d_{56}, d_{129}\}$
- Odpoveď vyhľadávacej stratégie S obsahuje tú istú postupnosť dokumentov ako v predchádzajúcom prípade, t.j.:

1. d_{123}	6. d_9	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56}	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25}	15. d_3
- Zostrojte **normovanú** krivku presnosť – návratnosť pre danú vyhľadávaciu stratégiu S a otázku q_2

17

Príklad (2)

Normovaná krivka presnosť - návratnosť

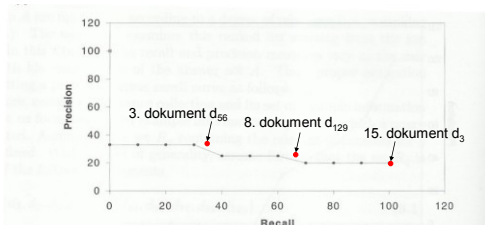


Figure 3.3 Interpolated precision at 11 standard recall levels relative to $R_2 \{d_3, d_{56}, d_{129}\}$.

18

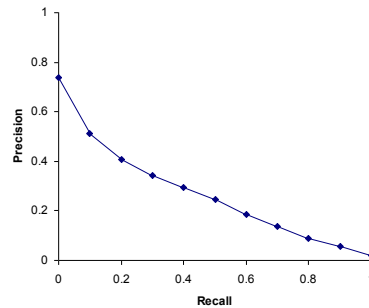
Vyhodnocovanie efektívnosti vyhľadávania (4)

- Vyhodnotenie vyhľadávacej stratégie S sa však spravidla nerobí na jednom dopyte q , ale na množine dopytov
- Definuje sa 11 normovaných úrovní návratnosti r : 0%, 10%, 20%, ..., 100%
- Vypočítajú sa presnosti na jednotlivých normovaných úrovniach návratnosti $P(r)$ pre všetky dopyty z testovacej množiny dopytov
- Nech počet dopytov je N_q , potom výsledná krivka bude zachytávať priemerné hodnoty presností pre jednotlivé normované úrovne návratnosti, vypočítané podľa vzorca:

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q} \quad \bullet \quad \text{Kde } P_i(r) \text{ je presnosť } S \text{ pri } i\text{-tom dopyte na úrovni návratnosti } r$$

19

Príklad (dobrý výsledok z TREC 8)



20

Príklad (3)

Vyhľadavanie založené na ontológii (1)

- Predpokladali sme, že množina konceptov pre daný dopyt je známa
- Množina konceptov asociovaných s daným dokumentom sa vyberie z databázy
- Tieto dve množiny sa porovnávajú nasledovnou mierkou podobnosti daného dokumentu D_i a dopytu Q :

$$sim_{onto}(Q, D_i) = \begin{cases} |Q_{con} \cap D_{i,con}| & \text{if } |Q_{con} \cap D_{i,con}| \neq 0 \\ k (=0,1) & \end{cases}$$

21

Príklad (3)

Vyhľadavanie založené na ontológii (2)

- Výsledná podobnosť sa vypočíta ako súčin podobnosti založenej na ontológii a podobnosti vypočítanej podľa vektorového modelu, (prípadne LSI modelu)

$$sim(Q, D_i) = sim_{onto}(Q, D_i) * sim_{TF-IDF}(Q, D_i)$$

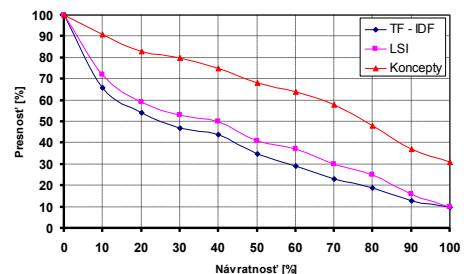
22

Príklad (3) – Použitá kolekcia dokumentov

- Kolekcia nazvaná **Cystická fibróza** (získaná z MEDLINE)
 - Kolekcia pozostáva z 1239 dokumentov
 - Minimálna veľkosť dokumentu 0.12 kb, maximálna veľkosť 3.8 kb a priemerná veľkosť 1.045 kb
 - Ku kolekčii existuje aj súbor so 100 dopytmi
 - Pre každý dopyt je známa množina relevantných dokumentov
 - Každý dokument v odpovedi je ohodnotený číslom 0 až 8 (4 nezávislí experti hodnotili mieru relevancie 0-2)
 - Existuje 821 konceptov a priemerný počet konceptov priradených dokumentu je 2.8
 - Priemerný počet dokumentov asociovaných s jedným konceptom je 4.2

23

Príklad (3) – Výsledky



24

Sumarizačné mierky efektívnosti vyhľadávania (1)

1. Priemerná presnosť pri nájdených relevantných dokumentoch (MAP - mean average precision)

- Táto miera favorizuje vyhľadávacie stratégie, ktoré rýchlo nájdu relevantné dokumenty

$$\bar{P}_{q_1} = \frac{1+0.66+0.5+0.4+0.3}{5} = 0.57 \quad \bar{P}_{q_2} = \frac{0.33+0.25+0.2}{3} = 0.26$$

2. R-presnosť (RP) je presnosť vyhľadávacej stratégie S na |R|-tej pozícii, t.j. pri |R|-tom vrátenom dokumente

$$RP_{q_1} = \frac{4}{10} = 0.4 \quad RP_{q_2} = \frac{1}{3} = 0.33$$

- Táto miera vlastne nie je sumarizačnou, popisuje iba jeden bod krivky presnosť - návratnosť, prax však ukazuje, že je vysoko korelovaná s MAP

25

Sumarizačné mierky efektívnosti vyhľadávania (2)

3. Presnostné histogramy sa používajú na porovnanie presnosti dvoch stratégií vyhľadávania (S_1 a S_2) pre viaceré dopyty $i = 1 \dots N_q$

$$RP_{S_1/S_2}(i) = RP_{S_1}(i) - RP_{S_2}(i)$$

4. Štatistiky v sumarizačnej tabuľke – napr. počet otázok, celkový počet vrátených dokumentov, z nich celkový počet relevantných dokumentov, a pod.

26

Používateľsky orientované mierky efektívnosti vyhľadávania

- U je podmnožina R takých dokumentov, ktoré sú používateľovi už známe
- $R_k = A \cap U$ je množina používateľovi známych dokumentov v odpovedi A
- R_U je množina relevantných dokumentov v odpovedi A, ktoré používateľovi neboli predtým známe

5. Pokrytie (coverage) C je definované nasledovne:

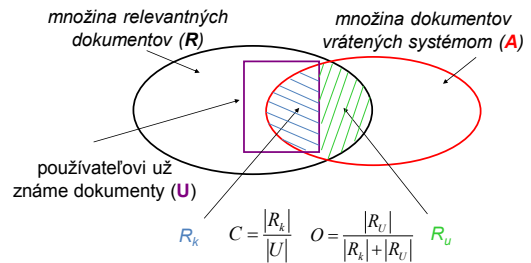
$$C = \frac{|R_k|}{|U|}$$

6. Novosť (novelty) O je definovaná nasledovne:

$$O = \frac{|R_U|}{|R_k| + |R_U|} = \frac{|R_U|}{|R_A|}$$

27

Vysvetlenie významu množín pri používateľsky definovaných mieraach efektívnosti vyhľadávania



28

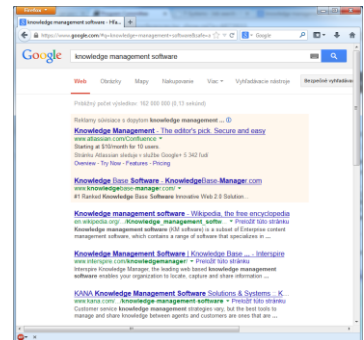
Vylepšovanie bežiacého IR systému

- Používateľské štúdie sú dobrý nástroj, najmä v čase návrhu, ale sú časovo náročné a nákladné
- Pre bežiaci IR systém sa najčastejšie používa metóda zvaná **A/B test**:
 - Pre takýto test sa spraví práve jedna zmena (systém B) aktuálneho systému (systém A), ktorej vplyv chceme odhadovať
 - Časť používateľských požiadaviek (1 až 10%) sa presmeruje na zmenený systém B, zvyšné spracúva aktuálne bežiaci systém A
 - Porovnávajú sa sledované parametre (napr. frekvencia klikaní na prvý odkaz v zozname) medzi systémami A a B
 - Pre dostatočne veľkom počte používateľov možno takýmto spôsobom lacno a rýchle overiť vplyv navrhovanej zmeny

29

Prezentácia výsledkov vyhľadávania

- Systém IR vráti usporiadaný zoznam dokumentov (podľa miery relevancie)
- Preddefinovaný počet dokumentov s krátkym popisom – **sumárom**
- Obsah sumáru je veľmi dôležitý – viac o sumároch viď ďalej



Sumáre vo výsledkoch IR systémov (1)

- Dva základné druhy sumárov:
 - **statické** – nezávislé na dopyte ktorý viedol k vyhľadaniu daného dokumentu, stále rovnaký
 - **dynamické** – prispôbené konkrétnemu dopytu, snažia sa ukázať, prečo bol daný dokument vybraný ako relevantný k dopytu
- **Statické sumáre** typicky predstavujú časť dokumentu
 - napr. prvých 50 slov, uložené do cache v čase indexácie
 - výber reprezentatívnej množiny viet z dokumentu - použitie NLP pre skórovanie viet a výber najlepších
 - sofistikované techniky sumarizácie textov – používané v experimentálnych systémoch

31

Sumáre vo výsledkoch IR systémov (2)

- **Dynamické sumáre** Prezentujú jedno alebo viac „okien“ v dokumente, ktoré obsahujú niekoľko termov z dopytu
 - vyžaduje rýchle vyhľadanie okien v cache pamäti dokumentov
 - Skórovanie nájdených okien vzhľadom na dopyt (príznamy ako veľkosť a poloha okna v dokumente)
- Aké sumáre používa vyhľadávač Google?

32