Dokumentografické informační systémy (DIS)

DIS = databáze textů (dokumentů) + software umožňující efektivní vyhledávání

na rozdíl od klasických databází jsou položky nestrukturované

nebo jen částečně strukturované (klasická strukturovaná –

např. autor, název, rok vydání; vlastní text nestrukturovaný)

tyto databáze jsou obvykle značně rozsáhlé

vyhledávací problém: najít k uživatelskému požadavku

relevantní dokumenty

formy vyhledávání: listování v databázi

formulace dotazu

dotaz = funkce, jejíž hodnotou je podmnožina dokumentů

dokumenty, které jsou odpovědí na dotaz, se nazývají hity –

nemusí být ale nutně relevantní (hit vyhovuje dotazu,

relevantní dokument vyhovuje požadavku uživatele)

formálně: model DIS je soubor pojmů či nástrojů pro

reprezentaci dokumentů

reprezentaci dotazu

reprezentaci pravidel a procedur umožňujících určit shodu

mezi požadavkem uživatele a dokumenty, které

vyhovují tomuto požadavku

informační služby DIS: zadat dotaz
vyhodnotit dotaz (případně upřesnit)
zpřístupnit odpovídající texty

problémy při vyhledávání:
najít vhodné algoritmy a datové struktury
určit, co je relevantní a co ne
zajistit efektivnost zpracování (tj. rychlost, přesnost a úplnost)
zajistit uspořádání výstupů podle relevance

koeficient přesnosti (precision) $P = \frac{R_Q}{A_Q}$

koeficient úplnosti (recall) $R = \frac{R_Q}{R_F}$

kde Q je dotaz
R_Q je počet vybraných relevantních dokumentů
R_F je počet všech relevantních dokumentů v kolekci
A_Q je počet všech vybraných dokumentů

tj. $P = P$(vybraný dokument je relevantní)
$R = P$(relevantní dokument byl vybrán)

ideální je $P = R = 1$, ale to se v praxi nestává, obvykle jsou ve vztahu nepřímé úměrnosti
preferuje se vyšší P

Modely DIS

1) Booleův model

reprezentace dokumentu: množina termů %

dotaz: Booleův výraz obsahující termy, logické spojky, eventuálně další speciály

(dotazy konjunktivní, disjunktivní, KNF, DNF, ...)

vyhledané dokumenty obsahují kombinaci termů specifikovanou dotazem

všechny dokumenty vyhovující dotazu jsou chápány jako stejně důležité, nelze je uspořádat podle hodnoty relevance

2) Vektorový model

dokumenty i uživatelské dotazy jsou vektory

podobnost dokumentu s dotazem je dána níjakým vzorcem

mějme množinu termů v dotazích $t_1, t_2, \ldots, t_n$

dokumenty $D_1, \ldots, D_m$

dotaz Q

reprezentace dokumentů:

$$D = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1n} \\ w_{21} & w_{22} & \ldots & w_{2n} \\ - & - & - & - \\ w_{m1} & w_{m2} & \ldots & w_{mn} \end{pmatrix}$$

1. Komu se říká indexace

termy jsou výrazy (klíčová slova), které charakterizují dokument

účelem je tedy co nejvíce zredukovat, přitom se zpracovávají nevýznamná slova (předložky, spojky, ...)

je to spojeno s určitou analýzou - termy jsou slova v základním tvaru

indexování dokumentů se provádí buď ručně nebo automaticky např. na základě četnosti slov v daném jazyce nebo oboru

podobně účelem termů je také rozlišit jeden dokument od druhého, nehodí se jako termy ta slova, která se vyskytují ve všech nebo téměř ve všech dokumentech

(bez. inverzní četnost termu v dokumentech - počet dokumentů, ve kterých se daný objevuje)

kde w_{ij} je váha j-tého termu v i-tém dokumentu

v nejjednodušším případě $w_{ij} = \begin{cases} 1 & \text{když } t_j \text{ se v } D_i \text{ vyskytuje} \\ 0 & \text{jinak} \end{cases}$

(tzv binární model)

jinak w_{ij} může být např. počet výskytů termu v dokumentu

reprezentace dotazu: $Q = (q_1, q_2, \dots, q_n)$

kde q_j je váha j-tého termu v dotazu

v nejjednodušším případě $q_j = \begin{cases} 1 & \text{když } t_j \text{ hledám} \\ 0 & \text{jinak} \end{cases}$

jinak q_j vyjadřuje důležitost termu v dotazu

(analogicky w_{ij} důležitost dokumentu vzhledem k hledanému termu)

podobnost dokumentu D_i s dotazem Q se počítá např. jako skalární součin $f(Q, D_i) = \sum_{k=1}^{n} q_k \cdot w_{ik}$

(ale existují i jiné míry)

pokud Q má tytéž termy jako D_i, pak $f(Q, D_i)$ = maximální

pokud D_i neobsahuje žádný z termů dotazu, pak $f(Q, D_i) = 0$

tento model umožňuje seřadit vyhledané texty podle podobnosti s dotazem, tj. podle "relevance"

pro netextové modely je to ještě výraznější

skórování však je deterministické na základě počtu výskytů

3) Pravděpodobnostní model

reprezentace dokumentů a dotazu je dvojná jako u vektorového modelu, ale váhy se stanovují na základě pravděpodobnosti
existuje v binární i netriviální podobě
předpokládá se nezávislost výskytu termů v dokumentech
(což je dost nereálné a bohy je to hlavní slabina modelu)
podobnost je vyjádřena např. skalárním součinem
požaduje se, aby platilo:

$$f(Q, D_i) > f(Q, D_j) \Leftrightarrow P(R/D_i) > P(R/D_j),$$

tj. dokument D_i je podobnější dotazu Q než dokument D_j,
právě když D_i je "relevantnější" než D_j.

⊛

$P(R/D_i)$ = P(vybraný dokument D_i je vzhledem k dotazu relevantní)

z této podmínky se odvozují optimální váhy termů v dokumentech

⊛ jinak: dokumenty s vyšší pravděpodobností relevance mají vyšší koeficient podobnosti

předpoklady (hledané termy jsou v dotazu i v dokumentech na pozicích n množin)

$D = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{pmatrix}$, kde $D_i = (d_{i1}, d_{i2}, \dots, d_{in}, d_{i,n+1}, \dots)$

d_{ij} = počet výskytů j-tého termu

$Q = (q_1, q_2, \dots, q_n, q_{n+1} \dots)$ kde $q_i = \begin{cases} 1 & \text{pro } i = 1, \dots, n \\ 0 & \text{pro } i > n \end{cases}$, tj. dotaz je binární

$f(Q, D_i) = \sum_{k=1}^{n} q_k w_{ik}$, kde w_{ik} jsou váhy vypočtené

z četností d_{ik} tak, aby byl splněn požadavek

$$f(Q, D_i) > f(Q, D_j) \Leftrightarrow P(R|D_i) > P(R|D_j)$$

označíme:

$P(R|D_i) = P($"vybraný" dokument D_i je relevantní vzhledem k $Q) =$

$= P($relevance dokumentu $D_i)$

$P(R) = P($relevance$) = P($náhodně vybraný dokument je relevantní$)$

$P(\bar{I}) = P($irelevance$)$

$P(D_i|R) = P($dokumentu D_i ~~se vyskytuje~~ mezi relevantními$)$

$P(D_i|\bar{I}) = P($ —"— irelevantními$)$

za předpokladu nezávislosti výskytu termů platí

$$P(D_i|R) = P(d_{i1}, \ldots, d_{in}|R) = \prod_{k=1}^{n} P(d_{ik}|R)$$

kde $P(d_{ik}|R) = P($v relevantním dokumentu D_i se k-tý term vyskytuje d_{ik} - hezký$)$

podobně $P(D_i|\bar{I})$

použijeme Bayesův vzorec pro $P(R|D_i)$ a $P(R|D_j)$ a vyjádříme, kdy platí $P(R|D_i) > P(R|D_j)$

dosadíme do nerovnosti

$$P(R|D_i) > P(R|D_j)$$

$$\frac{P(D_i|R)P(R)}{P(D_i|R)P(R)+P(D_i|\bar{I})P(\bar{I})} > \frac{P(D_j|R)P(R)}{P(D_j|R)P(R)+P(D_j|\bar{I})P(\bar{I})}$$

$$P(D_i|R)P(R)P(D_j|R)P(R) + P(D_i|R)P(R)P(D_j|\bar{I})P(\bar{I}) >$$

$$> P(D_j|R)P(R)P(D_i|R)P(R) + P(D_j|R)P(R)P(D_i|\bar{I})P(\bar{I})$$

$$P(D_i|R)P(D_j|\bar{I}) > P(D_j|R)P(D_i|\bar{I})$$

$$\frac{P(D_i|R)}{P(D_i|\bar{I})} > \frac{P(D_j|R)}{P(D_j|\bar{I})}$$

ľj. $\prod_{k=1}^{n} \frac{P(d_{ik}|R)}{P(d_{ik}|\bar{I})} > \prod_{k=1}^{n} \frac{P(d_{jk}|R)}{P(d_{jk}|\bar{I})}$ predpokladáme

$$\sum_{k=1}^{n} \log \frac{P(d_{ik}|R)}{P(d_{ik}|\bar{I})} > \sum_{k=1}^{n} \log \frac{P(d_{jk}|R)}{P(d_{ik}|\bar{I})}$$

w_{ik} w_{jk}

$$f(Q, D_i') = \sum_{k=1}^{n} q_k w_{ik} > \sum_{k=1}^{n} q_k w_{jk} = f(Q, D_j')$$

(položíme všechna q_k jsou jedničky)

Příklad: Požadavek Q ... 1 term

relevantní dokumenty: $D_1 = (2, \ldots)$

$D_2 = (2, \ldots)$

$D_3 = (1, \ldots)$

$D_4 = (1, \ldots)$

$D_5 = (0, \ldots)$

irelevantní dokumenty: $D_6 = (1, \ldots)$

$D_7 = (0, \ldots)$

$D_8 = (0, \ldots)$

$D_9 = (0, \ldots)$

$D_{10} = (0, \ldots)$

binární model: všechny frekvence ≥ 1 nahradíme 1

$\tilde{D}_1 = (1, \ldots)$

$\tilde{D}_2 = (1, \ldots)$

atd.

požadavek $Q = (w, 0, \ldots 0)$

kde $w = \log\left(\frac{p}{1-p}\right) - \log\left(\frac{q}{1-q}\right)$

$p = P$(relevantní dokument obsahuje term)

$q = P$(irelevantní — " —)

nechť např. $p = \frac{4}{5}$, $q = \frac{1}{5}$

pak $w = \log 4 - \log \frac{1}{4} = \log \frac{4}{\frac{1}{4}} = \log 16$

odpověď na dotaz: dokumenty $\tilde{D}_1, \tilde{D}_2, \tilde{D}_3, \tilde{D}_4, \tilde{D}_6$

nebinární model: vypočteme váhy termů v dokumentech

$$w_{i1} = \log \frac{P(d_{i1}|R)}{P(d_{i1}|\bar{I})} \qquad d_{i1} = 0, 1, 2$$

$P(0|R) = \frac{1}{5}$ $P(0|\bar{I}) = \frac{4}{5}$ ... váha $= \log \frac{1}{4}$

$P(1|R) = \frac{2}{5}$ $P(1|\bar{I}) = \frac{1}{5}$ ... váha $= \log 2$

$P(2|R) = \frac{2}{5}$ $P(2|\bar{I}) = 0$ .... váha $= \infty$

máme: $D_1' = (\infty, \dots)$

$D_2' = (\infty, \dots)$

$D_3' = (\log 2, \dots)$

$D_4' = (\log 2, \dots)$

$D_5' = (\log \frac{1}{4}, \dots)$

$D_6' = (\log 2, \dots)$

$D_7' = D_8' = D_9' = D_{10}' = (\log \frac{1}{4}, \dots)$

přiadavek: $Q = (1, 0 \dots 0)$

odpověď na dotaz: D_1', D_2' pak D_3', D_4', D_6'

nalezec ostatní

nebinární model dává lepší představu o důležitosti dokumentů

nevýhoda: váhy termů v dokumentech jsou většinou nenulové, i když odpovídající přehvenze výskytu jsou nulové

výsledek podobnosti je nepřehledný

Řešení: použít váhy d_{ik}'', kde

$$w_{ik}'' = \underbrace{\log \frac{P(d_{ik}|R)}{P(d_{ik}|\bar{I})}}_{w_{ik}} - \underbrace{\log \frac{P(Q_k|R)}{P(Q_k|\bar{I})}}_{w_{kQ}}$$

$P(Q_k|R)$ = P(relevantní dokument má nulovou přehvenzi výskytu k-tého termu)

zejména $w_{ik}'' = 0 \iff d_{ik} = 0$

Platí: nechť $D_i' = (w_{i1}, \dots, w_{ik}, \dots, w_{in})$

kde $w_{ik} = \log \frac{P(d_{ik}|R)}{P(d_{ik}|\bar{I})}$

$$D_i'' = (w_{i1}'', \dots, w_{ik}'', \dots, w_{in}'')$$

kde $w_{ik}'' = \log \frac{P(d_{ik}|R)}{P(d_{ik}|\bar{I})} - \log \frac{P(Q_k|R)}{P(Q_k|\bar{I})}$.

požadavek $Q = (q_1, \dots, q_k, \dots, q_n)$

Pak $f(Q, D_i') > f(Q, D_j') \Leftrightarrow f(Q, D_i'') > f(Q, D_j'')$

důkaz: označíme $d_{k0} = \log \frac{P(O_k|R)}{P(O_k|I)}$

$$f(Q, D_i'') - f(Q, D_j'') = \sum_{k=1}^{n} q_k d_{ik}' - \sum_{k=1}^{n} q_k d_{k0}$$

$$= \sum_{k=1}^{n} q_k (w_{ik} - w_{k0}) - \sum_{k=1}^{n} q_k (w_{jk} - w_{k0}) =$$

$$= \sum_{k=1}^{n} q_k (w_{ik}' - w_{jk}') = \sum_{k=1}^{n} q_k w_{ik} - \sum_{k=1}^{n} q_k w_{jk} =$$

$$= f(Q, D_i') - f(Q, D_j')$$

v Příkladě: $d_{10} = \log \frac{P(O|R)}{P(O|I)} = \log \frac{1}{4}$

$D_1'' = (\infty, \dots)$

$D_2'' = (\infty, \dots)$

$D_3'' = (\log 2 - \log \frac{1}{4}, \dots)$

$D_4'' = (\log 2 - \log \frac{1}{4}, \dots)$

$D_5'' = (0, \dots)$

$D_6'' = (\log 2 - \log \frac{1}{4}, \dots)$

$D_7'' = D_8'' = D_9'' = D_{10}'' = (0, \dots)$

Zobecnění: nebinární model dělí dokumenty vzhledem ke hraničnímu termu do tříd - v téže třídě jsou dokumenty s touž frekvencí výskytu termu někdy se také frekvence normalizuje délkou dokumentu

Jakákoli třída může být prázdná

jiné třídy: dokumenty s frekvencí v určitém intervalu

např: $\langle 0 \rangle, (0, f_1 \rangle, (f_1, f_2 \rangle \dots (f_{t-2}, \infty)$ ... t tříd

váha k-tého termu v dokumentu j-té třídy se definuje jako

$$w_{jk} = \log \frac{P(d_k \in \bar{I}_j \mid R)}{P(d_k \in \bar{I}_j \mid \bar{I})}$$

d_k ... normalizovaná frekvence výskytu

byla váhy zase splňují podmínku optimality

Předpoklad nezávislosti výskytu termů v dokumentech není příliš realistický.

Řešení:

a) termy dokumentu se rozdělí do skupin, termy uvnitř téže skupiny jsou navzájem závislé, termy z různých skupin jsou navzájem nezávislé
každá skupina se nahradí "supertermem" %.

b) použije se faktorová analýza – to je časově náročné i pro malé a středně velké soubory dokumentů

c) daná množina termů se transformuje na množinu "téměř nezávislých" termů (nebo alespoň méně závislých) – heuristická aproximace faktorové analýzy

d) uvažují se jenom některé závislosti – zjednodušený model

%. dokumenty se rozdělí do tříd, každá třída reprezentuje jinou kombinaci termů ze supertermu s určitou pravděpodobností

e) Bayesovské sítě

Princip je stejný jako u sítě modelující závislosti písmen ve slově

Učí se to na databázi dokumentů

Faktorová analýza

Pozorujeme (měříme) hodnoty nějakých proměnných.

Z výsledků měření plyne, že jsou navzájem silně korelovaný. Z toho usuzujeme, že jsou projevem nějaké další náhodné veličiny, která je v pozadí a projevuje se jejich prostřednictvím. Tyto proměnné se pak nazývají faktor.

Příklad: proměnné – paměť
pozornost
schopnost úsudku
atd.

faktor – inteligence

Faktorů může být víc.

Úkolem je najít co nejmenší počet faktorů, které vysvětlují dostatečně přesně to, co původní proměnné, ale je jich méně a jsou navzájem nezávislé.

Stanovení faktorů je založeno na korelační matici proměnných.

Původně byla faktorová analýza navržena pro situace, kdy hodnoty faktorů nemůžeme přímo měřit, jako konkrétní projevy ale ano.

Odhady parametrů:

odvodili jsme optimální váhy termů v dokumentech

jako $$w_{ik} = \log \frac{P(d_{ik}|R)}{P(d_{ik}|\bar{I})}$$

kde

$P(d_{ik}|R)$ = P (relevantní dokument má d_{ik} vyskytlí k-tého termu)

$P(d_{ik}|\bar{R})$ = P (relevantní -"-)

bylo pravděpodobnostně se počítají ze všech dokumentů v databázi

někdy se dají odhadnout na základě dokumentů, vybraných jako odpověď na dotaz:

vybrané dokumenty se rozdělí na relevantní a irelevantní.

nechť a_0 = počet relevantních dokumentů, ve kterých se hledaný term nevyskytuje

a_1 = —"— vyskytuje s četností 1

a_2 = —"— 2

atd.

b_0 = počet irelevantních dokumentů, ve kterých se hledaný term nevyskytuje

b_1 = —"— vyskytuje s četností 1

b_2 = —"— 2

atd.

pak $$P(d_{ik} = j \mid R) = \frac{a_j}{\sum_{t \geq 0} a_t}$$

$$P(d_{ik} = j \mid \bar{I}) = \frac{b_j}{\sum_{t \geq 0} b_t}$$

Dá se to použít, pouze když dokumentů v odpovědi je dostatečný počet.