9 VWM – Techniky pro vyhledávání textových, webových a multimediálních dokumentů: modely, algoritmy, aplikace. Optimalizace webových stránek pro vyhledávače.

Matej Choma

June 5, 2019

Tento handout sa zaoberá vyhľadávaním v dokumentoch a SEO. **Dokument** je full-text objekt, alebo full-text anotácia iného objektu na webe. Dokumenty sa skladajú z **termov**. Množina všetkých rôznych termov v súbore dokumentov sa nazýva **slovník**. Poznáme nasledujúce typy získavania informácií z webu, ktoré sa často kombinujú:

Query – jednorazové hľadanie na základe presného dotazu. Algoritmus vráti množinu (nie nutne usporiadanú) vyhovujúcich dokumentov.

Browsing – nemáme presný dotaz, manuálne prehľadávame explicitný, alebo virtuálny graf.

Filtering – algoritmus vracia dynamicky sa meniacu odpoveď pre statický dotaz (subscription na YouTube), alebo implicitný dotaz (recommendations na YouTube).

Pre vyhľadávanie v súbore dokumentov je možné použiť tradičné string-matching algoritmy ako Knuth-Morris-Pratt, Aho-Corasick, etc... Ich beh bude ale veľmi pomalý. Riešením je preprocesovanie a indexácia dokumentov.

Bez **preprocesovania** slovník obsahuje veľké množstvo termov – **stop words** – ktoré nám nič nehovoria o obsahu dokumentu. Patria sem neplnovýznamové slová, veľmi často sa opakujúce slová, ale taktiež čísla, ... Po odstránení stop words stále ostane v dokumentoch množstvo rovnakých slov, ale v rôznom tvare. Proces, ktorý nahradzuje slová ich základným tvarom sa nazýva **lemmatizácia**.

Na hodnotenie kvality odpovedí vyhľadávacieho modelu sa používajú dve metriky – precision a recall. Pre konkrétny model môžeme upravovať ich hodnoty upravovaním veľkosti odpovede, no **tieto metriky sú nepriamo úmerné**.

Precision – pravdepodobnosť, že dokument vo výsledku je relevantný – $P = \frac{\#RelevantAnswers}{\#Answers}$

Recall – pravdepodobnosť, že relevantný dokument je vo výsledku – $R = \frac{\#RelevantAnswers}{\#Relevant}$

1 Boolovský model

Boolovský model vytvára **term-by-document** maticu, kde vektory v riadkoch sú binárne vektory označujúce, v ktorých dokumentoch sa term nachádza. Táto matica je veľmi riedka a preto sa implementuje pomocou invertovaných listov – vektor termu je nahradený listom obsahujúcim usporiadané ID dokumentov, v ktorých sa nachádza (Fig. 1). Dotaz na model je množina termov s operátormi AND, OR a NOT. Tieto operátory reprezentujú operácie prienik, zjednotenie a doplnok na množinách ID dokumentov.

\+ Model je jednoduchý, dokumenty do výsledku buď patria, alebo nie.

\+ Implementácia pomocou invertovaného indexu je efektívna.

± Výsledky reflektujú dokonale dotaz. Užívateľ musí vedieť presne čo hľadá.

− Výsledné dokumenty majú všetky rovnakú váhu.

- q = Brutus AND Caesar
- result = (2, 8)

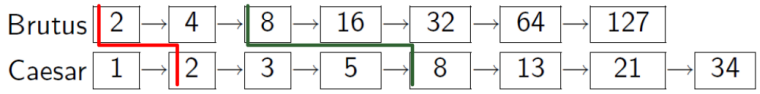

Figure 1: Invertovaný index a dotaz v boolovskom modeli.

2 Vektorový model

Vektorový model hľadá podobné dvojice dokumentov. Dokument d_j je reprezentovaný ako vektor dimenzie m (počet termov v slovníku), ktorý obsahuje váhu (relevanciu) každého termu zo slovníku v dokumente d_j. Každá dimenzia priestoru tak patrí jednému termu.

Váha termu t_i v dokumente d_j sa počíta ako

$$w_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot \log_2(\frac{n}{df_i}),$$

kde

$$tf_{i,j} = \frac{f_{i,j}}{\max_{k \in \{1,\dots,n\}} f_{i,k}}$$

a $f_{i,j}$ je počet výskytov termu t_i v dokumente d_j a df_i je počet dokumentov obsahujúcich term t_i.

Dotaz q má formu dokumentu. Vektory dokumentov sa porovnávajú medzi sebou na základe ich smeru, konkrétne pomocou kosínusovej vzdialenosti. Takže pokiaľ zreťazíme dokument sám so sebou a porovnáme ho s originálnym, budú úplne podobné, na rozdiel od úplne rozdielnych pri použití euklidovej vzdialenosti. Vďaka tejto metrike sú výsledky zoradené podľa miery podobnosti a je možné určiť hranicu, ako veľmi/málo podobné výsledky ešte vrátiť. Pri implementácií sa opäť používa invertovaný index.

$$\text{CosSim}(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{m}(w_{i,j} \cdot w_{i,q})}{\sqrt{\sum_{i=1}^{m} w_{i,j}^2 \cdot \sum_{i=1}^{m} w_{i,q}^2}}$$

\+ Model vie vyhľadávať na základe príkladu – nie je potrebné špeciálne špecifikovať dotaz.

\+ Výsledky sú zoradené.

\+ Modej je jednoduchý a implementácia pomocou invertovaného indexu je efektívna.

– Syntax nie je bratá do úvahy.

– Geometrizáciou sa v modeli stráca presná sémantická informácia.

– Nevie si poradiť s problémom synoným a homoným.

2.1 LSI vektorový model – latent semantic indexing

LSI vektorový model je úpravou vektorového modelu, ktorej cielom je adresovať problém synoným a homoným a znížiť dimenzionalitu vektorov. LSI pracuje s konceptami ako s lineárnymi kombináciami termov. Využíva metódu SVD (singular value decomposition) na rozklad term-by-document matice A.

$$A = U\Sigma V^T \approx U_k \Sigma_k V_k^T$$

Vektory v U reprezentujú koncepty, ktoré sú na základe zmenšujúcich sa hodnôt na diagonále Σ zoradené podľa dôležitosti. Tá rýchlo klesá a tak je dostatočné brať do úvahy iba prvých k konceptov (pre $10^4 - 10^5$ termov sa vyberá $10 - 10^3$ konceptov). Ako $concept - by - document$ matica sa berie

$$D_k = \Sigma_k V_k^T$$

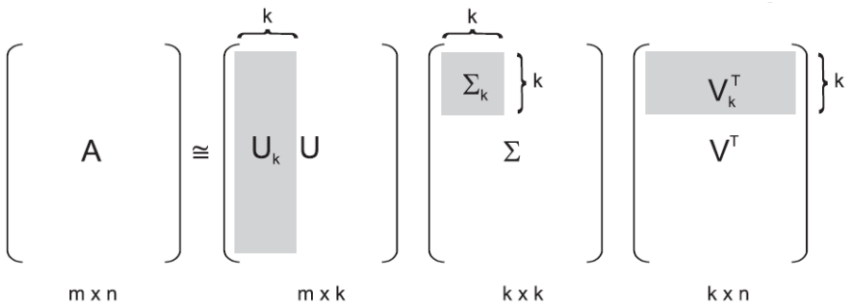

Figure 2: Redukcia dimenzionality pomocou SVD.

a dotaz sa zobrazí do priestoru konceptov pomocou

$$q_k = U_k^T q.$$

Dotazovanie a porovnávanie vektorov dokumentov prebieha rovnako ako vo vektorovom modeli pomocou kosínusovej vzdialenosti. Concept-by-document matica je hustá, takže použitie invertovaného indexu už nie je možné. Nakoľko je ale dimenzia vektorov rádovo v tisícoch, je model upočítateľný i bez indexovania.

\+ Model odkrýva skrytú sémantiku v súbore dokumentov.

\+ Model čiastočne rieši problém synoným a homoným.

\+ Redukcia dimenzionality.

\− Koncepty nemajú nič spoločné s lingvistikou, sú iba lineárnymi kombináciami termov.

\− Preprocesovanie je výpočtovo náročné. Výpočet SVD je v čase $\mathcal{O}(n^2 + m^3)$, kde n je počet dokumentov a m je počet termov v slovníku.

3 SEO

SEO (Search Engine Optimization) je spôsob úpravy webových stránok, aby boli dobre hodnotené webovými vyhľadávačmi. Obsah webovej stránky teda nemusí byť dobrý a prehľadný iba pre ľudského užívateľa, ale aj pre stroj. Vyhľadávače hodnotia webové stránky (angl. *web page*) namiesto celých webovoých sídiel (angl. *web site*), preto je potrebné optimalizovať každú stránku osobitne. Pri SEO sa zameriava na tieto prvky.

Názvy súborov – mali by byť krátke a relevantné ku obsahu stránky.

`<title>` **tag** – najdôležitejšia informácia o stránke pre vyhľadávače aj ľudí. Mal by to byť stručný (kratší ako 64 znakov) názov, sumarizujúci obsah stránky.

header tagy – informácia v názvoch je pre vyhľadávače dôležitá podľa levelu názvu `<h1>`,...,`<h6>`.

meta tagy – teraz sa používajú menej ako v minulosti. Stále je to ale spôsob ako predať prehľadnú informáciu užívateľovi vyhľadávača (Fig. 3).

textové modifikátory – zvýraznené kľúčové slová pomocou tagov `<strong>`,`<em>`,`<b>`,`<i>`,`<u>` sú odlíšené nie len vizuálne ale aj pre vyhľadávače.

optimalizácia obrázkov – vyhľadávače zatiaľ nerozumejú dobre obsahu obrázkov, preto je dôležité vyplniť tag `alt`.

interné odkazy – používanie interných odkazov robí navigáciu na stránke prehľadnejšiu pre užívateľa aj crawler. Kotviaci text sa berie ako dôležité kľúčové slovo.

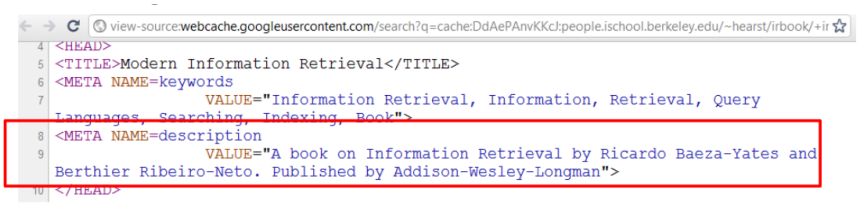

Figure 3: Meta tagy.

Obsah webovej stránky by mal byť kvalitný, originálny (duplikáty sú penalizované, aj originál aj plagiát), aktualizovaný, "pekný" a napísaný pre čitateľa, nie pre webový vyhľadávač. Pre vyhľadávač je dobré používať kľúčové slová vo vyššie zmienených tagoch aj v texte. Tieto by mali byť všetobecnejšie na hlavných stránkach a konkrétne na podstránkach. Iné ako html dokumenty na stránkach (PDF, MS Office, ...) by mali byť optimalizované rovnako, ako keby boli webovými stránkami.

Štruktúra webového sídla by mala obsahovať nasledujúce.

sitemap – html/xml/php súbor, ktorý pomáha s navigáciou robotov.

robots.txt – dáva informáciu, ktoré stránky by mali byť prejdené crawlerom a ktoré nie.

kontaktné údaje a zásady ochrany osobných údajov – robia webové sídlo dôveryhodnejším.

.htaccess – súbor s konfiguárciou pre Apeche server (ak web beží na ňom). Chránenie zložiek heslom,...

backlinks – odkazy na web z dobre hodnotených webov zvyšujú jeho hodnotenie.

SMO – social media opitmization – fan page na FB, YouTube,... Zvyšuje povedomie o webe medzi fanúšikmi.

Google Analytics je rozšírené riešenie od Google, používané na podrobné sledovanie aktivity na webe.

Neetické SEO je spôsob ako rýchlo a na krátku dobu dosiahnuť dobré hodnotenie od vyhľadávačov. Používa spamovanie kľúčových slov, odkazov, neviditeľný text,... Takéto SEO je často odhalené vyhľadávačmi a web zabanovaný.