SZABADSZAVAS KERESŐK RANGSOROLÁSA

KRAUSZNÉ PRINCZ Mária

Debreceni Egyetem, ATC Műszaki Kar
Műszaki Alaptárgyi Tanszék
4028 Debrecen, Ótemető u. 2-4
pmaria@delfin.unideb.hu

KIVONAT

Az eredmények rangsorolása az információkereső rendszerek nagyon fontos része. Ez a cikk áttekinti a rangsorolás matematikai alapjait, valamin ismerteti azon elveket, amelyek egy-egy weboldal rangsorbeli helyét meghatározzák. Számos technika létezik, amely egy weboldalnak jobb pozíciót biztosít valamely kereső rangsorolásánál. Ezen technikák két nagy kategóriára oszthatók: Az egyik csoportba azok tartoznak, amelyeket a keresők a jól tervezett weboldal részeként tekintenek, s így támogatnak (weboldalak optimalizálása), valamint azon technikák, amelyek a kereső rangsorolását próbálják manipulálni, ezért a keresők ezen próbálkozásokat büntetik.
Kulcsszavak: szabadszavas keresők, szabadszavas keresők rangsorolása, weboldalak optimalizálása, spamming

BEVEZETÉS

Számos rangsorolási technika létezik, de az egyes keresők pontos rangsorolási algoritmusa szigorúan titok tárgyát képezi két okból is: egyrészt a fejlesztők védeni akarják a módszereiket a versenytársaktól, másrészt nehezebbé kívánják tenni a weboldalak tulajdonosai számára is, hogy manipulálják az oldalaik rangsorbeli helyezését.

A RANGSOROLÁS MATEMATIKAI ALAPJAI

A hagyományos információ-visszakereső rendszerek **indexkifejezések**et használnak arra, hogy névmutatóval ellássanak és visszakeressenek dokumentumokat. Az indexkifejezések olyan kulcsszavak (vagy szókapcsolatok), amelyek jelentéssel bírnak. Az információ-visszakeresés alapja az az elgondolás, hogy a dokumentumok jelentése és a felhasználók információ szükséglete is kifejezhető indexkifejezések halmazaként. A visszakeresésben a problémát az okozza, hogy a dokumentum vagy a felhasználó kérésének jelentéséből is sok elvész, amikor a dokumentum szövegét vagy a felhasználó információ igényét szavak halmazával helyettesítjük. Az információ-visszakereső rendszerek egyik feladata annak az eldöntése, hogy mely dokumentumok relevánsak egy lekérdezéshez, s melyek nem. Ennek megítélése a rangsoroló algoritmustól függ, amely az eredményül kapott dokumentumok között megkísérel egy rangsort felállítani, s amely a visszakereső rendszerek lényeges része.

A klasszikus információ-visszakereső modellek

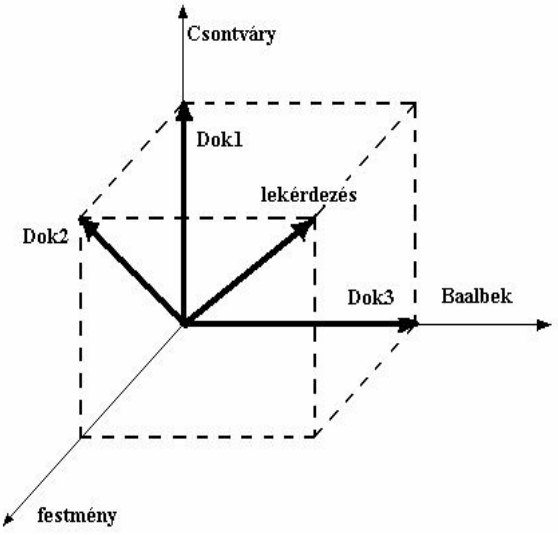
Az információ-visszakeresésben három klasszikus modellt különböztetünk meg: A Boole, a vektortér, és a valószínűségi modellt.

A **Boole modell**ben a dokumentumokat és a lekérdezéseket is indexkifejezések halmazaként tekintjük. A visszakereséskor azt vizsgálja a rendszer, hogy az egyes

dokumentumok tartalmazzák-e a lekérdezésben előforduló kifejezéseket. Boole lekérdezések eredménye mindig pontosan egyezik a lekérdezésben megfogalmazottakkal. Noha egy Boole kereső alapjában véve nem rangsorolja az eredményül kapott dokumentumokat (ehelyett egyszerűen kilistáz minden olyan dokumentumot, amely megfelel a lekérdezésben megfogalmazottaknak), de az eredmény rangsorolható és rendezhető, felhasználva egyéb tulajdonságokat. A leggyakrabban a lekérdezésben található kifejezések dokumentumbeli előfordulási gyakoriságát, a dokumentumon belül az előfordulás helyét, a kifejezések előfordulásának viszonylagos közelségét, a kifejezés szemantikai jelentőségét veszik figyelembe.

A **vektortér modell** a lekérdezés és a keresés eredményének hasonlóságát geometriai perspektívából közelíti. Ez azon a feltételezésen alapul, hogy mind a lekérdezés, mind egy tetszőleges dokumentum vektorként ábrázolható egy sokdimenziós térben, ahol a tér dimenziójának száma a lekérdezésben szereplő kifejezések számával egyenlő, s a lekérdezésben szereplő minden egyes kifejezés egy bázisvektornak (iránynak) felel meg. A dokumentum vektora a bázisvektorok lineáris kombinációjaként áll elő, ahol a bázisvektor együtthatója 1, ha a kifejezés szerepel a dokumentumban, és 0, ha nem. A dokumentumok hasonlóságát a lekérdezéshez a dokumentum és a lekérdezés vektora által bezárt szög adja, amely a két vektor skaláris szorzatából könnyen kifejezhető.

Tartalmazza?	Dokumentum1	Dokumentum2	Dokumentum3	Lekérdezés
Csontváry	igen	igen	nem	igen
Baalbek	nem	nem	igen	igen
Festmény	nem	igen	nem	igen



1.ábra. Példa a 3 dimenziós vektortér modellre

A **valószínűségi modell**ben a dokumentumok és a lekérdezések modellezése is a valószínűségi tételeken alapul, innen kapta a modell is a nevét. A valószínűségen alapuló visszakeresés bizonyos paraméterek valószínűségi becslésén alapul. A

valószínűségi modellek feltételezik, hogy a keresési szavak, kifejezések releváns és nem releváns dokumentumokban fordulnak elő, és megpróbálják azt megbecsülni, hogy egy dokumentum egy lekérdezéshez viszonyítva releváns-e vagy sem, valamint a relevancia visszajelzése révén megpróbálják a releváns dokumentumok számát megnövelni.

Mindhárom klasszikus modellben különböző alternatívákat dolgoztak ki az alapmodell erősítésére. A Boole modell alternatívái a fuzzy halmazok modellje és a kiterjesztett Boole modell. A vektortér modell alternatívái az általánosított vektortér modell, a szemantikus indexelés és a neurális hálók modellje. A valószínűségi modell alternatívája a Bayesian hálók.

Rangsorolás

A legtöbb szabadszavas kereső a Boole és/vagy a vektortér modell variációit használja a rangsorolásnál. Valamennyi kereső úgy rendezi a keresés eredményét, hogy az eredmény lista elejére az általa legfontosabbnak tartott dokumentumok kerüljenek. A rangsorolási algoritmusok keresőnként különböznek.

Keresési kifejezések helye és előfordulási gyakorisága

A szabadszavas keresők rangsorolásnál figyelembe veszik a kulcsszó előfordulásának helyét a dokumentumon belül, valamint az előfordulás gyakoriságát. A rangsorolásnál kiemelten kezelik, ha a keresett kifejezés az alábbi helyek valamelyikén található a dokumentumban:

- a dokumentum nevében (TITLE tag)
- az URL címben
- egyes HTML jelölők között (pl. a címsorokban, kiemelt szövegben)
- a keyword, description, ALT meta elemekben
- a hiperhivatkozások szövegében.

A valóságban azonban bizonyos kulcsszavak túlsúlya nincs mindig arányban az oldal jelentőségével. Éppen ezért egyre több kereső a keresési kifejezések előfordulási helyének feldolgozása mellett új megoldásokat is alkalmaz egy-egy oldal fontosságának meghatározásakor (második generációs szabadszavas keresők).

A keresés irányultságának fogalmi felismerése

Ha egy adott kulcsszó egyszerre több témát is felölel, akkor az eredménylista tetején a népszerűbb téma fog szerepelni, míg a keresett jelentésben esetleg csak a lista végén tűnnek fel oldalak. Ezt a jelenséget témasodródásnak nevezzük (topic-drift). A problémára megoldást jelenthet, ha téma szerint csoportosítjuk, azaz klaszterezzük az oldalakat.

A témasodródás elkerülésére egy másfajta megoldás, ha a kereső a feltett kérdés alapján megpróbálja kitalálni, hogy valójában mire is kíváncsi a felhasználó, s ennek megfelelően szolgáltatja az eredményoldalakat. Ez a megközelítés az alkalmazott

jelentéstan, a természetes nyelvi feldolgozás alkalmazását igényli.

Hivatkozások analízise

Az új rangsoroló algoritmusok közül több a hivatkozások tartalmazta információt használja fel az eredményoldalak rangsorának kialakításához. Ezek közé tartozik Kleinberg HITS algoritmusa és a Google által használt PageRank. Egy oldalnak akkor lesz a PageRankje magas, azaz akkor szerepel előkelőbb helyen a rangsorban, ha rá sokan hivatkoznak, vagy vannak olyan oldalak, amelyeknek nagy a PageRankje, és rámutatnak. A HITS első ránézésre nagyon hasonlónak tűnik a PageRank algoritmushoz, de van egy fontos különbség közöttük: a PageRank egy téma független, csak a linkstruktúra által meghatározott érték, a HITS viszont mindig egy konkrét témára nézve keres (tipikusan egy másik, vektortér-modell alapú kereső által visszaadott találatok között). Ez azt jelenti, hogy egyrészt a HITS eredménye sokkal pontosabb és relevánsabb lesz, másrészt viszont minden egyes kérdésnél újra ki kell számolni, tehát kevésbé hatékony. Mindkét algoritmus hibája, hogy nem szünteti meg a témasodródást. A hivatkozások analízisét minden kereső felhasználja valamely mértékben az eredmények rangsorolásakor.

Az oldal népszerűségének elismerése

A második generációs szabadszavas keresőkre jellemző, hogy egy oldal helyezését a rangsorban előnyösen befolyásolja az oldal népszerűsége. A hivatkozási népszerűség mellett (pl. Google) egyes keresők a kattintási népszerűséget is figyelembe veszik (pl. DirectHit). Az ilyen kereső egy webhelyt azon elv alapján rangsorol, hogy egy egyszerű keresés eredménylistájából hányan választják az adott webhelyet (click popularity), illetve mennyi időt töltenek el a webhely oldalain.

A rangsort módosító technikák

A rangsorolást javító technikákat két csoportra lehet osztani: Azon technikákra, amelyeket a szabadszavas keresők a jól tervezett oldalak ismérvének tekintenek (weboldalak optimalizálása), s így elfogadják azokat, valamint azon technikákra, amelyekre a keresők úgy tekintenek, mint amelyek manipulálni akarják a keresési eredmények rangsorát (spam), s ezért büntetik az ezen technikákat alkalmazó oldalakat.

Weboldalak optimalizálása

A weboldalak optimalizálása egy olyan eljárás, amely az oldal számára a legjobb helyezést kívánja biztosítani a szabadszavas keresők eredménylistáin a weboldal tartalmát leíró legrelevánsabb keresési szavak, kifejezések esetén. Az optimalizálás célja a dokumentum belső elemeinek (pl. név, meta elemek, stb.) és külső elemeinek (pl. hivatkozási népszerűség) módosítása a jobb rangsorbeli eredmény érdekében.

A sikeres optimalizációhoz vezető lépések:

A megfelelő kulcsszavak megválasztása
Egy dokumentum az általa tartalmazott szavak, kifejezések alapján kereshető vissza a szabadszavas keresők indexéből, ezért lényeges, hogy milyen szavakat használunk egy-egy honlap kialakításakor. Fontos, hogy minél széleskörűbben írjuk le a képviselni kívánt tartalmat (pl. szinonimák használata, a hasonló témakörök megemlítése, stb.) Ne használjunk általános kulcsszavakat (pl. Internet), a specifikus szavak használatával könnyebben megtalálják a specifikus tartalmat a felhasználók. Tegyük könnyen kereshetővé az oldalainkat az idegen szavak, kis és nagybetű, többes szám átgondolt használatával! Figyeljünk a helyesírásra! Használjuk a kulcsszavak hosszabb és rövidebb formáját is (pl. TV, televízió)!

A tartalom optimalizálása
Fontos, hogy a nyitólap minél több kulcsszót tartalmazzon. A kulcsszavakat lehetőleg emeljük ki, azaz helyezzük a mondatok, illetve a szöveg elejére! Alkalmazzuk a kulcsszavakat minél közelebb egymáshoz, hiszen számos kereső (pl. a Google) a rangsorolásnál ezt figyelembe veszi!

Az oldal jelölőinek optimalizálása
A <TITLE> elemmel közbezárt rész a dokumentum legfontosabb jelölője, amelyet a legtöbb szabadszavas kereső indexel. A dokumentum nevének fel kell kelteni az emberek érdeklődését is, ezért célszerű a semmitmondó Honlap név helyett a legfontosabb kulcsszavakat ezen a helyen szerepeltetni. (pl. Cipők, csizmák, szandálok, táskák, övek és egyéb bőráruk forgalmazója: Kényelem Nagykereskedelmi Kft.)

A description, keywords, ALT meta elemeket a legtöbb kereső különböző mértékben használja fel, de miután jelenlétüket egyetlen kereső sem bünteti, így használatuk javasolt. A description elem segítségével a dokumentum tartalma adható meg, s számos kereső az eredménylistában – ha a dokumentum tartalmaz ilyen meta elemet – ezt az összegzést adja vissza. A keywords elem alkalmazásával a dokumentum tartalmára vonatkozó kulcsszavak adhatók meg, míg a képek tartalmának leírására az ALT jelölő elem szolgál.

A kiemelésre szolgáló jelölők (pl. fejlécek - H1, H2,…-, vastag, dőlt kiemelések, stb.) használata szintén javasolt.

Hivatkozások
A hivatkozások szövege fontos tényező rangsorolásnál, ezért érdemes rá odafigyelni. Image map helyett célszerű hivatkozásokat használni.. A keretes oldalszerkezetet helyett az oldal könnyű gyors bejárásához jól felépített belső link struktúra kialakítása célszerű.

URL címek
Miután az URL címeket számos kereső indexeli, így lényeges, hogy az URL kulcsszavakban releváns legyen. A dinamikusan változó tartalmú oldalakhoz állandó

URL használata javasolt.

Regisztrálás

A weboldal optimalizálása után az oldalt célszerű a legfőbb szabadszavas keresőknél és tematikus keresőknél regisztrálni. A regisztrálás célja a keresők figyelmét egy adott webhelyre irányítani, hogy indexeljék azt, vagy jelenítsék meg a könyvtárukban. Ezáltal az oldal hivatkozási népszerűsége növelhető.

A rangsor befolyásolása

A rangsort emelő tényezők mellett elengedhetetlenül fontos tudni azt is, hogy mit nem szeretnek a szabadszavas keresők. Valamennyi szabadszavas kereső küzd azon taktikák ellen, amelyek jobb rangsorolást adnak egy kevéssé releváns webhelynek. Ezen taktikák összességét spamnek nevezzük.

Spamnek minősülő technikák:

- A kulcsszavak túlzásba vitt ismétlése, ami a felhasználó számára láthatatlan, ha a háttérszín és a betűszín azonos, vagy ha a betűméret elég kicsinek van megválasztva, ugyanakkor a kereső számára a kulcsszavak láthatók.
- A TITLE elem többszörözése, amelyek közül csak az elsőt jelenítik meg a böngészők, de a robot valamennyit indexeli.
- A tartalomhoz nem releváns kulcsszó szerepeltetése a dokumentum megnevezésében, illetve a meta elemekben.
- A tartalom duplikálása ugyanazon, vagy közel ugyanazon oldalak más hoston való elhelyezése által.
- A nyitólapra mutató „mesterséges linkek” elhelyezése.
- A dokumentum frissítését kérő http-equiv="refresh" meta elem jelenléte a fejlécben, amely a felhasználó böngészőjét az adott oldal néhány másodperces időközökben történő újbóli letöltésére utasítja.
- Túl sok oldal előterjesztése egy rövid idő alatt, vagy tematikus keresők esetén a nem megfelelő fogalomkörbe történő előterjesztés.

Az említett próbálkozások eredménye számos keresőnél az, hogy az érintett webhelyet alacsonyabb rangsorolással büntetik, vagy automatikusan kizárják az adatbázisukból.

ÖSSZEFOGLALÁS

A szabadszavas keresőknél lekérdezéskor az alapmodellen túl a rangsorolást javító egyéb megoldások eredményezik egy-egy weboldal rangsorbeli helyét az adott keresőn. A találatok rangsorolásakor az adott weboldalon a keresési kifejezések előfordulásának helyét, gyakoriságát, az oldalra mutató hivatkozásokat, s ezáltal az oldal népszerűségét a legtöbb kereső értékeli, de keresőnként a rangsorolás szempontjai változók. Egy-egy weboldal jobb rangsorbeli helyezése az oldalnak nagyobb látogatottságot biztosít, ami a profitorientált vállalkozásoknál jobb üzletmenetet, nagyobb hasznot is jelent. Ezért a weboldalak készítésekor célszerű

figyelmet fordítani arra, hogy az oldal mely tulajdonságait értékelik pozitívan a keresők, s melyek azok, amelyeket büntetnek a találatok rangsorolásakor.

FELHASZNÁLT IRODALOM

[1] K.Princz Mária, Információkeresés a weben és tanítása, PhD értekezés, Debreceni Egyetem, Természettudományi Kar, 2007
[2] Baeza-Yates,R., Ribeiro-Neto,B., Modern Information Retrieval, Addison Wesley, 1999
[3] Shapiro,Y.,Lehoczky,E. Factors that influence search engines rankings http://www.searchengines.com
[4] SearchEngines http://www.searchengines.com/
[5] SearchEngineShowdown http://www.searchengineshowdown.com
[6] SearchEngineWatch http://searchenginewatch.com
[7] Wikipedia http://www.wikipedia.org/
[8] WebReference http://www.webreference.com

SEARCH ENGINE RANKING

Ranking is a very important part of information retrieval systems. There are some ranking techniques, but search engine ranking algorithms are closely guarded secrets. This paper deals with the mathematical basis of ranking, and details some principles that determine the relevancy of a web page. There are techniques to improve ranking. These techniques include two broad categories: techniques that search engines recommend as part of good design, and those techniques that search engines do not approve of because they try to manipulate search engine results.