

FEJLETT INFORMÁCIÓKERESÉSI TECHNOLÓGIA A FELSŐOKTATÁSBAN

*Karácsony Gyöngyi, e-mail cím
Debreceni Egyetem*

*Jóföldi Endre, jofoldi.endre@polymeta.hu
K-prog Bt.*

Összefoglaló

Milyen problémákkal szembesülünk a keresés során? Az információk az Interneten nagyon különbözőek (heterogének): A tartalom szétszórtan helyezkedik el különböző szervereken, címeken, formátumokban és eltérő nyelveken, más-más hallgatóságot megcélozva. Még a nagyobb kereső programok, mint a Google is csak a lapok 45%-át indexelik és teszik kereshetővé. A "Rejtett Web" (Hidden Web) adatbázisai (pl. PubMed, Web of Science) becslések szerint ezerszer több adatot tárolnak, mint a "Nyitott Web" oldalai. A "Nyitott Web" és a "Rejtett Web" közös kérdései: probléma az információ lefedettség, a minőség, a túl sok felesleges adat, relevancia, aktualitás és teljesség, valamint a nyelvi félreérthetőségek és az eltérő felhasználói felület

Egy metakereső megoldás: A metakeresők képesek egyszerre több "Nyitott Web" és "Rejtett Web" forrás keresésére azért, hogy növeljék a keresési területet, a keresés pontosságát, fontosságát (relevanciát), eredményességét és hatékonyságát.

Milyen problémákkal szembesülünk a keresés során?

Heterogén információk

Az információk az Interneten nagyon különbözőek (heterogének): A tartalom szétszórtan helyezkedik el különböző szervereken, címeken, formátumokban és eltérő nyelveken, más-más hallgatóságot megcélozva.

Holott a legtöbb információ valóban elérhető mégis gondjaink vannak a keresésnél, ugyanis sokan sokféle formában írták le azt, amit felhasználni szeretnénk. Ezen a területen egyértelműen fejlődés tapasztalható, hiszen a fejlettebb keresők már nem csak a html fájlokat indexelik, hanem például a pdf vagy doc fájlokat is, amelyek igen sok korábban közvetlenül hozzá nem férhető információt tettek kereshetővé. De azért érzékelhető, hogy a rengeteg különböző fájltypus nehezíti ezt a helyzetet.

A heterogenitás nem csak a már megtalált információknál jelentkezik, hanem a keresés során is. Különböző keresők és adatbázisok különböző felületekkel rendelkeznek, különböző címeken érhetőek el. A kényelem nagy úr. Nem véletlen, ha a legtöbb ember számára a keresés egyetlen kerső használatát jelenti, hiszen a legtöbb esetben valóban lehet valamilyen információt kapni bármilyen témáról.

Az indexelés nem teljes

Még a legnagyobb kereső programok, mint a Google is csak a lapok 45%-át indexelik¹ és teszik kereshetővé. Miért van ez? A Google adatbázisa jelenleg több mint 8 milliárd oldal indexét tartalmazza. Hogy hogy nem elég ez? Nem túlzás ez a 45%? Érdemes végignézni, hogyan növekedett az internetes tartalmat szolgáltató szerverek száma².

Nagyon fontosnak tartjuk felhívni a figyelmet, hogy ebben az esetben még nem az úgynevezett Hidden (rejtett) vagy Deep (mély) jelzőkkel illetett technikailag nem elérhető információkról beszélünk. Ezek hagyományos statikus html oldalak valamiért mégsem kerülnek indexelésre.

¹ A google által indexelt lapok száma 2005. június 9-én: 8,058,044,651 lap.

² Forrás: <http://www.isc.org/index.pl?ops/ds/host-count-history.php>

Dátum	Internet hosztok száma
08/1981	213
05/1982	235
08/1983	562
10/1984	1,024
10/1985	1,961
02/1986	2,308
12/1987	28,174
07/1988	33,000
01/1989	80,000
10/1990	313,000
01/1991	376,000
01/1992	727,000
10/1992	1,136,000

01/1993	1,313,000
01/1994	2,217,000
01/1995	4,852,000
01/1996	9,472,000
01/1997	16,146,000
01/1998	29,670,000
01/1999	43,230,000
01/2000	72,398,092
01/2001	109,574,429
01/2002	147,344,723
07/2002	162,128,493
01/2003	171,638,297
01/2004	233,101,481
01/2005	317,646,084

Itt egyszerre két problémával kell szembenézni. Látható egy hihetetlen gyors növekedés, ami nagyságrendileg is komoly növekedést takar. Úgyis megfogalmazhatnánk, hogy a tavalyi évben napi 230.000-el nőtt a szerverek száma. Egyrészt tehát, van egy nagyon gyors növekedés, ami nehézséget jelent.

A probléma másik oldalát úgy világítanánk meg, hogy pl. 2005. június 8-án ³ 53 millió domain név létezett. 24 óra alatt több mint 700 ezer új domain nevet jegyeztek be, és ugyanakkor 680 ezret töröltek. Vagyis napi szinten a domain nevek több mint 1 százaléka változott. Hogy egy analógiát használjak, ahhoz lenne ez hasonlítható, mintha az Országos Széchényi Könyvtár 7,5 milliós gyűjteményébe naponta 75 ezer új könyvet kellene felvenni, és mondjuk 70 ezret pedig leselejtezni. Azt hiszem jól látjuk mindannyian, hogy hamarosan komoly nyilvántartási problémáik lennének.

Rejtett web⁴

A "Rejtett Web" (Hidden Web) adatbázisai (pl. PubMed, Web of Science) becslések szerint több százszor több adatot tárolnak, mint a "Nyitott Web" oldalai. Rengeteg olyan adatbázis, adatforrás található meg interneten keresztül elérhető formában, amely a tartalmát csak meghatározott kérésekre tárja fel. Vagyis nem indexelhető a hagyományos módszerekkel. Egyszerűen nincsenek a kereső crawler kutató robotjai által elérhető fájlok, amiket kereshetővé lehetne tenni a hagyományos módon. Néhány évvel ezelőttig a hidden web körébe sorolták még a különböző nem html fájlformátumban levő tartalmakat is, de ahogy ezt már korábban jeleztem mára a fejlettebb keresési technológiák erre problémára

³ <http://www.whois.sc/internet-statistics/>

⁴ A következő címen egy nagyon részletes és hasznos gyűjteményt találunk a témával foglalkozó cikkekről: <http://www.deepwebresearch.info/>

megoldást kínálnak, és a legfontosabb dokumentum típusok ma már tartalmukban is kereshetőkké váltak. A probléma és a kérdéskör mégis fent maradt. Miért?

Talán nem is gondolunk bele, de nagyon sok olyan forrás létezik, ahol az információk háttér adatbázisokban találhatóak és csak kérésre kerülnek elő azokból egy dinamikusan legenerált html oldalon való megjelenítésre. Például publikációs adatbázisok, amelyek ma már a legtöbbször teljes szövegükben tartalmazzák az adott publikációt, telefonkönyvek, enciklopédiák, szótárak, könyvtári katalógusok, törvények szövegei, szabványok, szabadalmak, hirdetések, hírek – amelyek sokszor a legfontosabb információkat osztják meg egy témával kapcsolatban. Ezt hallva talán már nem tűnik túlzásnak a fent említett arány.

Természetesen ezeket a forrásokat is kereshetővé lehet tenni, mint ahogy például a NIH PubMed adatbázisa kereshetővé lett téve a google-lal, azon az áron, hogy az NIH meghatározott időközönként átadja az adatbázisainak tartalmát indexelésre. Ennek azonban nyilvánvaló korlátai vannak. Egyrészt az együttműködés oldaláról, hiszen a keresőknek sorra meg kellene állapodniuk ezekkel a forrásokkal, másrészt a legfrissebb – ilyen módon legrelevánsabb cikkek – csak az eredeti adatbázisban érhetőek el, hiszen nem lehet naponta átadni ezeket az információkat.

A "Nyitott Web" és a "Rejtett Web" közös kérdései:

....

A minőség:

Probléma az információ lefedettség

Túl sok felesleges adat, áttekinthetőség:

Relevancia: például az úrkutatás keresőszóra a google az ötödik a yahoo a negyedik helyen egy viccgyűjteményt hoz, ahol az úrkutatáshoz kapcsolódó vicceket találunk. Ez a példa jól illusztrálja, hogy egy-egy forrás relevanciája a kérdésben nem feltétlenül könnyen meghatározható.

Aktualitás: a keresők indexelő programjai, csak meghatározott időközönként képesek végiglátogatni az internetet, holott nagyon sok tartalom nagyon gyakran változik. Természeteseb ezek az algoritmusok is folyamatosan fejlődnek de nem várható minden tekintetben kielégítő megoldások

Teljesség

Nyelvi félreérthetőségek

Eltérő felhasználói felület

Speciális problémák az oktatási, könyvtári információ keresés területén

....

Egy metakereső megoldás: PolyMeta⁵

A metakeresők képesek egyszerre több "Nyitott Web" és "Rejtett Web" forrás keresésére azért, hogy növeljék a keresési területet, a keresés pontosságát, fontosságát (relevanciát), eredményességét és hatékonyságát. Milyen módon teszik ezt?

A rejtett webhez tartozó forrásokat kapcsolhatunk össze az indexelhető webbel a keresések során ezzel a *keresési területet* lehet növelni. Adott tématerületeken minőségi és releváns információt tartalmazó oldalakat emelhetünk ki. Tetszőleges számú egymástól teljesen eltérő adatokat tartalmazó adatforrás egyidejű keresésére (keresők, híroldalak, könyvtári katalógusok, publikációk oldalak stb.) van lehetőség. A keresett adatbázisok köre felhasználó szinten is szabályozható. Igény szerint kihagyhatunk vagy hozzáadhatjuk az általunk kívánt forrást a kereséshez. Adott esetben ugyanis a megoldáshoz inkább a keresési terület célirányos szűkítése vezethet bennünket, kiválasztva a témában releváns információt tartalmazó forrásokat a kereséshez, kevesebb de sokkal jobb minőségű információt találhatunk.

Az *áttekinthetőség*, túl sok információ, illetve hatékonyság (information overload) problémájára a szoftverünkben a dinamikus generált tartalomjegyzék jelent megoldást. Nagyon sok esetben a keresés nem a túl kevés, hanem a túl sok információ miatt mondható sikertelennek. A keresők nagy része csak az első néhány találatot nézi meg, és ha ott nem talál valami érdekeset, akkor egy másik kérdéssel, vagy keresővel próbálkozik. Hogyan tehetnénk elérhetővé a sokadik oldalon megbújó esetleg mégis értékes találatokat? Könnyen lehetséges, hogy egy téma megsemmisül az első oldalon illetve, a témához kapcsolódó találatok mindegyike szinte biztosan nem látható egyszerre. Erre jelent megoldást a tartalomjegyzék vagy index. Azonosítjuk és nyelvi csoportokba rendezzük a találatokat, amelyek így jobban és gyorsabban áttekinthetővé válnak.

The screenshot shows a search interface with a content index on the left and search results on the right. The index lists categories like 'baleset...', 'vontatbaleset...', 'történet...', 'főúton...', 'balesetben...', 'értesült az OBJEKTÍV Hírogynökség...', 'Szemtanúkat keresnek a 47-es úti...', 'harmincheten meghaltak', 'halt meg a szombaton délután...', 'súlyos...', 'jelentések szerint 37-en veszítették életüket a Japán nyugati...', and 'Többet'. The search results on the right show multiple entries for 'Baleset miatt teljes útzár a 35-ösön' and 'Baleset miatt teljes útzár a 35-ös főúton', each with a date, time, and source (HírTV, MNO, HírTV, MNO, HírTV, MNO). The results also include 'Nagy vontatbaleset Japánban' and 'Szemtanúkat keresnek a 47-es úti háromhalás balesetben'.

A tartalomjegyzék fejlett nyelvi elemző technológiák felhasználásával készült, amiben nyelvi elemzéssel kapcsolat kutatók⁶ évtizedes tapasztalait használtuk fel a lexikai elemzéshez. Felhasználunk szinonimaadatbázis speciális területeken (pl MESH), a magyar nyelvi elemzés területén pedig a Morphologic-kal működünk együtt.

⁵ Részletesebb információk a www.polymeta.hu oldalon érhetőek el.

⁶ Doszkoes Tamás, a National Library of Medicine kutatója <http://tamas.nlm.nih.gov>

A keresés *eredményességét* egy jelenleg fejlesztés alatt álló módon is növelni kívánjuk, amikor a forrásokhoz mintegy kívülről szeretnénk hozzáadni nyelvi tudást az ún. query expansion segítségével. Például ha a google-ben az „információkeresés” szóra keresünk 4930 találatot kapunk, ha az „információkeresési” szóra keresünk 337 találatot kapunk. Ha a két keresést kombináljuk az OR szóval. Ezt szeretnénk automatikussá tenni a magyar nyelv esetében. Mondhatnánk persze, hogy miért nem teszik meg ezt maguk a felhasználók? Hadd idézzek egy tanulmányból, amit történetesen a testvérem írt ☺: „Általános benyomásként elmondható, hogy a hallgatók kevés ismerettel rendelkeznek a webes keresési módszerekről, mint pl.: szavak összekapcsolása, logikai operátorok... Néhány hallgató használta ugyan keresései során az idézőjeles keresést, amivel konkrét kifejezésekre kereshetünk... A másik ilyen problémát a logikai operátorok okozták. A Google-ben nem szükséges használatuk, mégis többen kapcsolták össze "AND"-del, a beírt szavakat. Minden ilyen esetben megjelent egy tájékoztató szöveg, miszerint nem szükséges ezt használni, mert a Google alapértelmezetten így kapcsolja össze a szavakat, de ezt egyik résztvevő sem vette észre.”⁷. Azzal folytatnám, hogy sok esetben nem is egyértelmű, hogy mivel lenne érdemes kibővíteni a keresési kifejezést. Szeretnénk ezt a terhet amennyire csak lehet levenni a keresők válláról.

A keresési funkciók háttérét egy jól kidolgozott, széleskörűen paraméterezhető adminisztrációs rendszer adja meg, amely lehetőséget ad arra, hogy egészen eltérő igényekhez is testre szabható legyen a rendszer az egyszerűtől a legösszetettebb kutatói funkciókig.

⁷ Jóföldi Hajnalka: Kultúrák hatása a weben való információkeresési szokásokra, 2003