Hároméves kutatás a webes keresőgépek adatbázisainak frissességéről

Bevezetés

A weben működő keresőgépek minőségének meghatározása összetett feladat. A legtöbb figyelem a visszakeresés hatékonyságára irányul, de valójában négyféle szempontból lehet értékelni a keresőket:

- Az indexállomány minősége: az indexelés a teljes web mekkora részére terjed ki, milyen az országok szerinti megoszlás, és mennyire naprakész az adatbázis?
- A találatok minősége: ez a klasszikus relevancia-tesztekkel vizsgálható; de kérdés, hogy az egyes keresőgépek és felhasználóik speciális jellemzői milyen módszerekkel mérhetők?
- A keresési funkciók minősége: milyen opciók állnak rendelkezésre (pl. „összetett keresés"), mennyire kifinomult a keresőnyelv, mennyire megbízhatóan dolgozik a keresőprogram?
- A keresőgép használhatósága: ez a felhasználók viselkedésének vizsgálatával mérhető, és kérdőíves felmérésekkel, előre gyártott tesztek végrehajtásával, vagy a naplóállományok utólagos elemzésével valósítható meg a gyakorlatban.

A *Hamburgi Egyetem Design, Média és Információ Szak*án dolgozó szerző jelen tanulmánya kizárólag az indexállomány minőségével és azon belül is annak a frissességével foglalkozik. Ez egy fontos szempont, hiszen a web folyamatosan és gyorsan változik: új oldalak jelennek meg, a régiek tartalma megváltozik vagy törlődik, és módosulnak a hiperhivatkozások is. Bár bizonyos esetekben a már régóta nem változott weblapok tartalma is hasznos lehet, de a felhasználók gyakran inkább az új vagy a mindig frissen tartott oldalakat preferálják, ezért a keresőgépeknek képeseknek kell lenniük arra, hogy ezt a fajta „up-to-date" igényt is kielégítsék.

2004-es kutatásuk alapján *Ntoulas*, *Cho* és *Olston* [1] úgy becsülték, hogy mintegy 320 millió új weblap jelenik meg minden héten. Az adott pillanatban létező weboldalak kb. 20 százaléka eltűnik egy éven belül, és nagyjából felének módosul a tartalma ugyanezen időszak alatt. A hiperstruktúra még ennél is gyorsabban változik: az URL címek legalább 80 százalékban megváltoznak vagy újak lesznek egy év alatt. Ebből is láthatjuk, hogy milyen fontos, hogy a keresőgépek frissen tartsák az adatbázisukat.

Feltehetően mindenki találkozott már 404-es hibákkal a keresése során – vagyis amikor a találati listában levő ugrópontok elérhetetlen oldalakra mutatnak. Az elmúlt évek vizsgálatai szerint ezek aránya viszonylag alacsony: az első 20 találatnál 2,2 és 6,5 százalék között van – keresőgéptől függően –, de még így is elég sok bosszúságot okoznak. Ez a probléma is az index frissességével függ össze.

Az világos, hogy gazdaságossági és műszaki okok miatt egyik keresőgép sem képes a teljes indexállományát naponta felfrissíteni. A szakemberek szerint egyhavi teljes körű frissítési ciklus lenne elvárható a keresőgépektől. A szerző egy korábbi vizsgálata szerint ez még az olyan nagy keresőknél sem teljesül, mint a Google és a Yahoo. Kérdés persze, hogy kell-e ragaszkodni ehhez az ajánlott periódushoz, vagy a régi oldalak hosszabb ideig is újraindexelés nélkül maradhatnak?

A problémakör áttekintése

A weblapok frissességének megállapítása nem egyszerű, pedig ez fontos szempont a találatok súlyozásánál és olyan kereséseknél, ahol a felhasználó csak az új információkra kíváncsi. Hogy egy weblap mennyire friss, az megbecsülhető többek között a készítési dátumából, a tartalmának változási gyakoriságából, a benne levő ugrópontok elavultságából. Pusztán a változás gyakoriságának figyelése félrevezető lehet olyankor, amikor például csak az oldalon feltüntetett napi dátum vagy a külalak változik, de a lap tartalma érdemben nem.

Úgyhogy a módosítási gyakoriság mellett a módosítás mértékét és jellegét is figyelembe kell/kellene vennie a keresőrobotnak, amikor eldönti, hogy milyen gyakran látogat meg egy weboldalt.

Az indexállományok felépítése kétféle módon történhet: kötegelt (batch) vagy növekményes (incremental) módszerrel. Előbbinél minden frissítéskor újra felépítik a teljes adatbázist, utóbbinál csak a változó és az új oldalak adatait adják hozzá folyamatosan a korábbi állapothoz. A kötegelt üzemmódnál ezért nincs nagy jelentősége annak, hogy milyen gyakran változik egy oldal, mert a következő ciklusban mindenképpen újraindexelésre kerül. Az inkrementális elven működő rendszereknél viszont nagyon fontos kiszámolni azt az optimális időtartamot, ami után egy adott oldalt ismét be kell gyűjtenie a keresőrobotnak. A gyakorlatban az újraindexelés gyakoriságánál még az oldal népszerűségét is figyelembe szokták venni: a gyakran hivatkozott, sok ember által látogatott webhelyeket sűrűbben járják be a robottal, mint a kevéssé népszerűeket.

Greg Notess [2] egy 2003-ban publikált kutatása során hat keresőkérdést futtatott le több keresőgépen (MSN, HotBot, Google, AlltheWeb, AltaVista, Gigablast, Teoma és Wisenut). A találati listákból azokat a weblapokat választotta ki, amelyeket naponta frissítettek, és amelyeken fel volt tüntetve az utolsó módosítás dátuma. Mindegyik nagy rendszernél (MSN, HotBot, Google, AlltheWeb és AltaVista) voltak aznapi vagy előző napi tételek is a találati listában. A legrégebbi találatok kora erősen megoszlott: például az MSN és a HotBot esetében ez 51 nap volt, az AlltheWeb listájában viszont előfordult 599 napja nem indexelt oldal is. Vagyis az optimálisnak tartott 30 napos teljes indexfrissítést még a nagy keresőgépek sem tudtákmegoldani, de azért a megvizsgált találatok dátumának átlaga a legnagyobbaknál nagyjából egy hónap volt, kivéve az AltaVistát, ahol ez az átlag három hónapnál is nagyobbnak adódott.

A jelen kutatás ismertetése

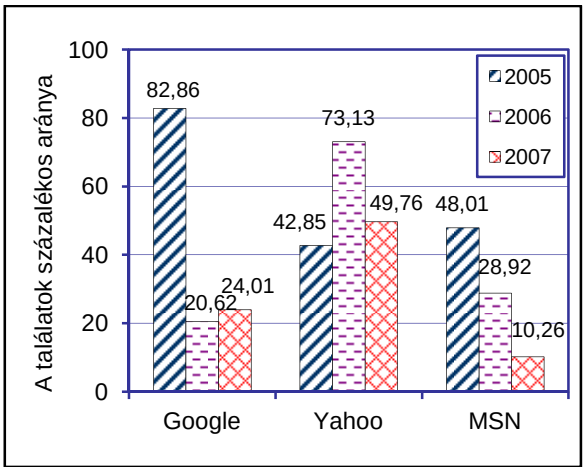
Ezen tanulmány szerzője és társai [3] 2005-ben folytattak le egy hasonló vizsgálatot, amikor is 38 német nyelvű, naponta frissített webhelyet választottak ki, és megnézték, hogy a nagy keresőgépek (Google, Yahoo és MSN) által tárolt weblapok dátuma milyen régi és hogyan változik egy hathetes időszak alatt. Mivel a Yahoo nem mutatja meg a cache tárolójában levő oldal lementési időpontját, ezért csak olyan weblapokat vizsgáltak, amelyeken fel van tüntetve az utolsó módosítás dátuma. Összesen 1558 oldalt ellenőriztek mindegyik keresővel minden nap, és összeszámolták, hogy ezek között mennyi az egy napnál nem régebbi. A legjobb eredményt akkor a Google adta 82,86%-kal, az MSN esetében ez az arány 48,01% volt, a Yahoo-nál pedig 41,85%. Egyszerű számtani átlagot számolva a Google adatbázisában a vizsgált weblapok 3,1 naposak voltak, ugyanez az átlag az MSN esetében 3,5 nap, a Yahoo-nál pedig 9,8 nap volt. A középérték (medián) így alakult: Google és MSN: egy nap, Yahoo: négy nap. A három keresőgép közül egyedül az MSN frissítette valamennyi vizsgált oldalt 20 vagy annál kevesebb napon belül, a másik kettőnél előfordultak egészen régi találatok is.

Ezt a 2005-ös kutatást a következő két évben megismételték, hogy arra is választ kapjanak: vajon változik-e a keresőgépek frissítési stratégiája az időben? 2006–2007-ben egy kontrollcsoportot is összeállítottak 30 olyan weblapból, amelyek nem naponta, hanem csak hetente, havonta, vagy rendszertelenül változnak, így azt is tesztelni tudták, hogy van-e különbség a napi frissítésű és a ritkábban módosuló webhelyek újraindexelési stratégiájában? Emellett arra is választ szerettek volna kapni, hogy milyen gyorsan válnak kereshetővé a begyűjtött oldalak az egyes keresőgépekben?

Az első kérdéssel kapcsolatos vizsgálat eredményét a mellékelt grafikon mutatja *(1. ábra)*. Jól látható, hogy 2005-höz képest a Google erősen visszaesett; az MSN is egyre rosszabbul teljesít ebből a szempontból; a 2005-ben még utolsó Yahoo pedig a következő két évben a legjobbnak bizonyult, igaz, 2007-ben már jóval gyengébb eredménnyel. Ezek elég meglepő trendek, mert az ember azt várná, hogy a nagy keresőgépek ragaszkodnak a frissítési stratégiájukhoz, és nem cserélgetik évenként. Feltehetően az indexállomány gyors növekedése lehet a fő ok ezen látványos változások mögött: a web exponenciális ütemben gyarapodik és a keresőgépek nem tudnak ezzel a tempóval megbirkózni. A Google esetében az is hozzájárul a 2005 óta történt komoly visszaeséshez, hogy bár még most is igen nagy arányban indexeli le naponta a vizsgált oldalakat, de ezek csak kétnapos késéssel válnak kereshetővé.

A legfrissebbek helyett a legrégebbi oldalakat kikeresve a vizsgált halmazban, a következő eredmények adódtak: a Google ebből a szempontból fo-

kozatosan javult (2005-ben 54, 2007-ben pedig csak 10 napos volt a legrégebben lementett weblap); a Yahoo-nál is csökkent ez az érték (62-ről 26 napra); míg az MSN-nél romlott a helyzet (2005–2006-ban 17-16 nap volt, 2007-ben pedig 30 nap). A nem naponta változó oldalakból álló kontrollcsoport eredményei azt mutatják, hogy bár ezek között is vannak olyanok, amelyeket naponta újraindexelnek a keresőgépek, de a Google esetében 2006-ban lehetett találni 253 napos, 2007-ben pedig 175 napos másolatokat is a cache-ben. Ugyanakkor az MSN-nél 19 illetve 30 nap volt a legrégibb állapot az elmúlt két évben. Tehát bár a



1. ábra **Az egy napnál nem régebbi találatok arányának változása három év alatt**

Google a gyakran módosuló lapokat a másik két keresőgépnél sűrűbben indexeli újra, a ritkábban aktualizált oldalak esetében egész nagy elmaradások is lehetnek. Az MSN viszont képes a teljes indexállományát megújítani egyhónapos ciklusokban. Ez két eltérő stratégiát jelez: a Google-nál a fontos és gyakran módosuló lapok újraindexelése élvez prioritást, míg az MSN-nél az egész indexállomány rendszeres felújítása is lényeges cél. A felhasználók szempontjából ez azt jelenti, hogy a Google-val nagyobb valószínűséggel lehet a legújabb információkat megtalálni, az MSN-nél viszont nem fordulhat elő az, hogy egy teljesen elavult verziójú weblapba botlunk. Bár az is igaz, hogy a Google a „kevéssé fontos”-nak minősített és ezért ritkán leindexelt oldalakat a találati listáiban hátrább rangsorolja, így ilyenekre feltehetően amúgy sem kattint rá az emberek többsége.

Frissítési ütemezés és indexelési késedelem

Jogosan gondolhatnánk, hogy egy adott keresőgépnél egységes frissítési ciklusok vannak: ugyanazt az oldalt ugyanolyan gyakorisággal látogatják meg. De a tapasztalatok azt mutatják, hogy ez nem egészen így van. Példaként a német nyelvű Wikipédia honlapjának indexelését vizsgálták meg a kutatók. Ezt a Google robotja minden nap begyűjti, de többnyire csak két nap múlva válik kereshetővé a legutóbb lementett állapot. A Yahoo rendszertelenül viselkedett a vizsgálat idején: a cache-ben levő példány hetekig nem változott, majd az utolsó két hétben gyakran – volt amikor naponta – újraindexelte a robot a kezdőoldalt. Az MSN szisztematikusabb volt, de itt is változtak az időintervallumok: az első frissítés 30 nap után történt, míg a következő 16 nappal később.

A Google esetében a kétnapos késlekedés a Wikipédia mellett más oldalakra is igaz; az esetek 68 százalékában nem található meg aznap vagy másnap a begyűjtött weblap, hanem csak később válik kereshetővé az indexelés lassúsága miatt. Ez a sávszélesség pocsékolása, mert fölösleges minden nap letölteni olyan oldalakat, amelyek azután csak kétnapos késéssel lesznek visszakereshetők. A Yahoo a begyűjtött oldalak több mint 50 százalékát még aznap képes kereshetővé tenni, míg az MSN-nek egy vagy két nap szükséges ehhez.

Összefoglalva megállapítható, hogy egyik nagy keresőgép sem működik ideálisan a felhasználók szempontjából, vagyis nem képesek a web jelentős részét lefedve olyan gyakran újraindexelni és visszakereshetővé tenni a lapokat, amilyen gyakran azok változnak. Nem egyszerű azt sem megállapítani, hogy egy-egy weboldalnál mi lenne az ideális begyűjtési gyakoriság, mert ezt a tényleges tartalom változásának gyakorisága és mértéke, az oldal népszerűsége, továbbá gazdaságossági szempontok és technikai paraméterek – például sávszélesség – is befolyásolják. Hogy melyik megközelítés a jobb: a Google módszere (csak a fontos és változó oldalak gyakori indexelése) vagy az MSN-é (minden oldal újralátogatása egy adott időintervallumon belül), az még további vizsgálatokat igényel.

Irodalom

[1] NTOULAS, A. – CHO, J. – C. OLSTON, C.: What’s New on the Web? The Evolution of the Web from a Search Engine Perspective. = Proceedings of the

Thirteenth WWW Conference, New York, USA, 2004.

[2] NOTESS, G. R.: Search Engine Statistics: Freshness Showdown, 2003. *http://www.searchengineshowdown.com/statistics/freshness.shtml*

[3] LEWANDOWSKI, D. – WAHLIG, H. – MEYER-BAUTOR, G.: The Freshness of Web search engine databases. = Journal of Information Science, 32. köt. 2. sz. 2008. p. 133–150.

/LEWANDOWSKI, Dirk: A three-year study on the freshness of web search engine databases. = Journal of Information Science, 34. köt. 6. sz. 2008. p. 817–831./

(Drótos László)