TÁJÉKOZÓDÁS A WEBEN

K. Princz Mária
pmaria@delfin.unideb.hu
Debreceni Egyetem Műszaki Főiskolai Kar

1. Bevezetés

Az Internet használata egyre inkább tért hódít a mindennapi életben, s különösen igaz ez az oktatás területén. Hatalmas mennyiségű információ érhető el a weben keresztül, de vajon megtaláljuk-e mindig a számunkra éppen szükségeset?

A Debreceni Egyetem Műszaki Főiskolai Karán az első éves hallgatók körében vizsgáltuk, mennyire képesek tájékozódni a weben, hogyan, milyen hatékonysággal keresnek. Tapasztalataink megegyeztek a CyberAtlas honlapján közzétett iProspect tanulmánnyal:

- Az Internetet használók mintegy ¾-e használ kereső szoftvert.
- A felhasználók 52 %-a ugyanazt a keresőszoftvert vagy tematikus keresőt használja az információk keresése esetén és csak 35% használ alternatív kereső eszközt. 13% esetében igaz, hogy különböző típusú kereséseknél különböző keresőszoftvert használ.
- A keresőszoftvert használók 16%-a csak az első egy-két eredményt nézi meg, 32% végignézi az első eredményoldalt, 23 % eljut a második eredményoldalig, 10,3 % a harmadikig, 8,7% még a harmadik után néz, és 10 % végignézi az összes eredményt, ha az nem oldalak tucatja.
- Az első keresés sikertelensége esetén a felhasználók csupán 7,5%-a finomítja tovább a kereső kérdést, 27,2% pedig más keresőszoftverrel próbálkozik.

A hiányokat érzékelve a webes keresés tanítását felvettük az általános gyakorlati tananyagba, s erről a hallgatói visszajelzések pozitívak.

2. A web megoszlása

2.1. A látható web (Surface Web vagy Visible Web)

A web része, amely a weben át elérhető, az általános keresőszoftverek által begyűjtött és indexelt dokumentumokat tartalmazza.

2.2. A láthatatlan web (Invisible Web)

A webnek azon része, amely a keresőszoftverek számára láthatatlan. Ezen dokumentumokat a keresőszoftverek nem tudják (technikai korlátok) vagy nem akarják (elvi döntések) adatbázisukba beemelni.

Különböző becslések találhatók a weben a láthatatlan web nagyságáról. A legtöbb becslés a látható webnek 400-550–szeresére értékeli, de van, aki ettől is nagyobbnak tartja.

A láthatatlan web részei:
Nem átlátható web (Opaque Web)

Azon dokumentumok tartoznak ide, amelyek könnyen elérhetőek lennének a begyűjtő robotprogramok számára, de valamely okból mégis kimaradnak a keresőszoftverek adatbázisaiból.
Ennek oka lehet a begyűjtés mélységének korlátja, az egy domainről begyűjthető dokumentumok számának korlátja, a frissítési periódus alatt megváltozott, megjelent oldalak, a dokumentumszigetek.

Magán web (Private Web)

Az Interneten át elérhető, de különböző megoldásokkal a nyilvánosan látható oldalak közül kizárt anyagok, az intranet hálózatokon található dokumentumok tartoznak ide.
A kizárás történhet a robot.txt fájl alkalmazásával vagy a noindex meta tag használatával, de a tűzfallal, jelszóval való védelem is lehetséges.

Szabadalmazott web (Proprietary Web)

Olyan adatbázisok, dokumentumok tartoznak ide, amelyeket tartalomszolgáltatók állítanak elő, és ezért regisztráció után előfizetéssel vagy külön díjért tekinthetők csak meg

Mély web (Deep Web)

A láthatatlan web legnagyobb részét teszi ki. A mély web a weben át szabadon elérhető adatbázisokat tartalmaz (pl. menetrendek, sárga oldalak, stb.). Ezen adatbázisok tartalma technikai korlát miatt láthatatlan a begyűjtő robotok számára, hiszen a lekérdezés megadására (begépelés, kiválasztás) e programok képtelenek.

Az igazán láthatatlan web (Truly Invisible Web)

A nem HTML formátumú dokumentumokat (audio, video, képek) nehezen vagy sehogy sem értelmezik a keresőszoftverek, ezért nem is építik be az adatbázisukba. Számos egyéb formátumra (pdf, flash, office fájlok, stb.) is igaz, hogy időigényes indexelni őket, ezért számos keresőszoftver nem foglalkozik ezen típusú dokumentumokkal.
Elvi okból kerülnek kizárásra a dinamikusan generált web oldalak, valamint a valós idejű tartalmak.

3. A webes keresés eszközei

A weben lévő információ keresésekor néhány jól bevált stratégiát követhetünk: a jónak vélt URL cím beírása, keresőszoftverek (search engines), témakatalógusok (subject directories), metakeresők (meta-search engines), webes adatbázisok alkalmazása.

3.1. Keresőszoftverek

A weben lévő dokumentumok keresésében a keresőszoftverek szerepe elsődleges. A keresőszoftverek tulajdonságainak ismerete segít a lekérdezések minél hatékonyabb megfogalmazásánál, de hasznos a jól kereshető web oldalak írásánál is.

A keresőszoftverek három elkülönülő részből épülnek fel:

Begyűjtő rész

Minden keresőszoftvernek megvan a saját robotprogramja, amely a begyűjtést végzi. A különböző robotok a weben lévő dokumentumok csak egy részét indexelik. Különböznek abban, hogy mely szervereket tekintenek kiindulási pontnak, egy adott domainről hány dokumentumot gyűjtenek be, milyen frissítési periódust használnak.

Indexelő rész

Az indexelő részben a begyűjtött dokumentumok különböző elemeiből (pl. a dokumentum neve <TITLE>, fejléc információk, címsorok, horgony elemek, kiemelt szövegrészek) a keresőszoftver saját adatbázist épít fel vagy tovább bővíti azt.

Lekérdező rész

A lekérdező rész további három részből áll:

- a lekérdezési interfész (egyszerű, összetett, részletes keresés lehetősége)
- a lekérdezést megvalósító egység (a keresőszoftver adatbázisából veszi elő a dokumentumokat),
- az eredményeket rangsoroló rész.

Valamennyi kereső úgy rendezi a keresés eredményét, hogy az eredménylista elejére az általa legfontosabbnak tartott dokumentumok kerüljenek. A rangsorolási algoritmusok keresőnként különböznek (szövegek statisztikai analízise, hivatkozások analízise, klaszterezés).

A web át elérhető információk döntő hányadánál a keresőszoftverek nem segítenek minket az információk megtalálásában.
A különböző keresőszoftvereken feltett kérdések találati listája nagyon különböző, és kevés az átfedés közöttük. Ez azt jelenti, hogy célszerű minél több keresőt használni, ha valamely témában alaposan át szeretnénk nézni a weben tárolt dokumentumok tömegét.

3.2. Metakeresők

A metakeresők használatával számos weben keresztül elérhető adatbázist kérdezhetünk le egy időben, egységes lekérdező interfészen keresztül, így a web nagyobb részéből kapunk találatokat.

3.3. Témakatalógusok

A témakatalógusok az Interneten található dokumentumokra mutató hiperhivatkozások sokszor hierarchikus gyűjteménye, ahol tartalom szerint felépülő könyvtárakban kereshetünk. A témakatalógusok a webnek szűkebb részét fedik le, mint amit a keresőszoftverek adatbázisai tartalmaznak.

3.4. Webes adatbázisok

Számos hasznos információ különböző adatbázisokból nyerhető. (pl. telefonszámok, menetrend, stb.)
A láthatatlan web részét képező források, adatbázisok elérését segítik a rendszerezett hivatkozások gyűjteményei, a témakatalógusok (pl. www.invisible-web.net, de egy-egy adatbázis kezdőlapjának megtalálásához a keresőszoftverek is jól használhatók. Ekkor a keresés megfogalmazásakor célszerű a kulcsszavak megadása mellett a lekérdezést bővíteni az *adatbázisok OR archívumok OR gyűjtemények* megadásával, majd a kapott nyitóoldalon keresni az adatbázis saját lekérdező eszközeivel.

4. A webes keresés tanításának tapasztalatai

Hatékonyabban, erőteljesebben kereshetünk, ha ismerjük a rendelkezésre álló lehetőségeket, éppen ezért fontos a webes információkeresést tanítani, növelni a felhasználók tudatosságát a keresőeszközök használatakor.

A tapasztalat azt mutatja, hogy a hallgatók jelentős része nem ismeri a lehetőségeket egy-egy információ keresésekor. A többség keresőszoftvert használ minden esetben, de legfeljebb egy-két keresőszoftver URL címét ismeri. Nem ismerik a különbözőségeket e szoftverek

működésénél (pl. milyen típusú információt gyűjtenek), így választani sem tudnak, mikor melyiket érdemes használni.

A sikeres kereséshez a megfelelő keresőszoftver kiválasztásán túl fontos annak tulajdonságainak ismerete is: nem mindegy, hogyan fogalmazzuk meg a lekérdezést, hiszen az az eredményhalmazt döntően befolyásolja.

Érdemes tudatosítani, hogy az összetett lekérdezések könnyítésére egyre több keresőszoftver lehetővé teszi a részletes keresést is, így nem kell megjegyeznünk a pontos beírási szabályokat, de bonyolultabb összetett kereséseket nem lehet megfogalmazni a részletes keresés oldalon.

Számos portál lehetővé teszi nyitólapján a weben való keresést, de csak egy mezőt kínál fel a keresés megfogalmazására, ami nehezíti a lekérdezés megfogalmazását.

Általános hibák:

- Sokszor előforduló hiba, hogy a hallgatók magyar nyelvű szövegek keresésekor begépeléskor nem használják az ékezetes betűket (pl. *epiteszet építészet* helyett). Ez esetben számos jelentős honlap kimarad az eredménylistából, s csak azok az oldalak jelennek meg, amelyeknél az URL címben szerepel a keresési kifejezés (pl. epiteszet.html), esetleg a szöveg létrehozója meta adatként keresési kulcsként hibásan is szerepelteti (gondolva az ékezet nélküli begépelésekre), vagy a dokumentum szövegében is helytelenül írva jelenik meg a keresett szó.
- Szintén kizárunk releváns oldalakat, hibásan szűkítjük a találati listát rossz helyesírással megadott keresési kulcsszóval.(Pl. *szinész* – 1200 találat, *színész* – 37600 találat a Goggle esetében.)
- Sok esetben túl messziről közelítenek a hallgatók egy-egy keresésnél. Ha a gótikus építészet stílusjegyeire vagyunk kíváncsiak, akkor kevés az *építészet* keresési kulcs megadása.
- Érdemes tudatosítani, hogy keresőnként különböző eredményeket kapunk, ha a szavakat egyes vagy többes számban szerepeltetjük. A * helyettesítő karaktert sem ismeri minden kereső (pl. Google).
- Sokan nem tesznek különbséget szavak és kifejezések keresése között. Fölöslegesen bővítjük az eredményhalmazt, ha kifejezéseket keresési kulcsszavanként adunk meg (osztott rendszerek “osztott rendszerek” helyett). Szókapcsolatok keresése esetén érdemes az idézőjelet kitenni vagy a részletes lekérdezés ablakánál a megfelelő mezőbe írni a szavakat, mert így az eredménylistát lényegesen szűkíthetjük.
- Megszorítások adhatók a dokumentum előfordulási helyére domainenként, site-onként. Szűkíthetjük az eredménylistát, ha megadjuk a dokumentum nyelvét, típusát, méretét, a dokumentumok létrehozására vonatkozó időkorlátot, stb. Természetesen keresőszoftverenként változik, hogy milyen szűkítést enged meg, s milyen módon fogadja el a megadást.
- Legyünk körültekintők az összetett keresések megfogalmazásánál! Ha egy keresőszoftver nem ismer egy logikai operátort, ami a lekérdezésben szerepel, akkor azt a megadott keresési kulcsszavak mellett egy újabb kulcsszónak veszi, ezáltal alapértelmezésének megfelelően hibásan tovább szűkíti (pl. AllTheWeb) vagy fölöslegesen bővíti az eredményhalmazt.

A keresések megfogalmazásától lényegesen függ a találatok száma. Az alábbi táblázat erre mutat néhány példát:[1]

Kereső kifejezés	*AltaVista*	*AllTheWeb*	*Google*	*Heuréka*	*Vizsla*
osztott rendszerek	687	1571	4980	1189	114545
osztott rendszer	1373	3321	4110	1775	109047
"osztott rendszerek"	66	93	142	-	142
"osztott rendszer"	31	49	75	-	120
osztott AND rendszerek	687	428	4980	1189	115923
osztott OR rendszerek	28493	140	69100	46022	2332276
tanszék	14356	95526	78100	19455	112993
tanszék site:klte.hu	-	924	904	615	1145
tanszék host:delfin.klte.hu	2	141	145	39	193

2. ábra: A lekérdezések eredményei[2]

Mára a legtöbb keresőszoftver nyitólapjáról biztosítja valamely válogatott, rendszerezett hivatkozásgyűjtemény elérését, illetve a tematikus keresőként induló szolgáltatók is lehetővé teszik nyitólapjukon a weben lévő keresést. A tanórák alatt lényeges tudatosítani a hallgatókban, hogy mikor melyik eszközt célszerű használniuk, s érdemes rávilágítani a gyakori hibákra.

Irodalomjegyzék:

[1] Search Engine Watch *http://searchenginewatch.com*
[2] WebReference *http://www.webreference.com*
[3] Search Engine Showdown *http://www.searchengineshowdown.com*
[4] S.Brin, L.Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine WWW7 / Computer Networks 30(1-7), pp. 107-117,1998 *http://www.stanford.edu*
[5] C. Sherman, G. Price: The Invisible Web: Uncovering Information Sources Search Engines Can't See. CyberAge Books, 2001
[6] CyberAtlas http://cyberatlas.internet.com/markets/advertising/article/0,,5941_1500821,00.html
[7] J. Nielsen, Search: Visible and Simple Alertbox, May 13, 2001 *http://useit.com*
[8] K. Princz Mária: Systems to access information in the Web MicroCAD'2000 International Computer Science Conference, Miskolc
[9] K. Princz Mária – Rutkovszky Edéné: Content Discovery of Invisible Web 6th International Conference on Applied Informatics, Eger 2004
[10] K. Princz Mária: A webes keresők tanításának tapasztalatai E-learning alkalmazások a hazai felsőoktatásban, Budapest, 2003
[11] K. Princz Mária: A weben lévő információk hozzáférhetősége NetworkShop, Pécs 2003
[12] K. Princz Mária – Rutkovszky Edéné: Ismeret reprezentáció a weben NetworkShop, Eger 2002

[1] 2002-es adatok
[2] "- " tartalmú cella: a keresőgép nem támogatja az adott műveletet