Krauszné Princz Mária

A WEBEN LÉVŐ INFORMÁCIÓK HOZZÁFÉRHETŐSÉGE

E cikk a weben lévő óriási mennyiségű információ elérhetőségének lehetőségeit elemzi. Ezen belül rendszerezi az eligazodás legjelentősebb eszközeiről, a keresőszoftverekről szóló információkat. A keresők tulajdonságainak ismerete segít a lekérdezések minél hatékonyabb megfogalmazásánál, valamint a jól kereshető web oldalak írásánál is.
Kulcsszavak: keresőszoftverek, tematikus keresők, a láthatatlan web, adatbázisok a weben

1. BEVEZETÉS

Az Internetnek nevezett világháló használata a World Wide Web megjelenésével (1992) rendkívül felgyorsult. A webet kezdetben az akadémiai közösségekben használták kizárólagosan, de gyorsan és fokozatosan terjedt el az élet szinte minden területén.
Hatalmas tömegű információ érhető el az Internet dokumentumhálóján keresztül. Ezen óriási mennyiségből a szükséges információt tartalmazó dokumentum előkeresésére néhány jól bevált keresési stratégiát követhetünk: tematikus keresők, keresőszoftverek segítségével képesek lehetünk megtalálni a keresett anyagot. A weben át elérhető dokumentumok egy jelentős része viszont ”láthatatlan”, azaz a keresőszoftverek nem segítenek minket ezen dokumentumokban lévő információk megtalálásában.
E cikk a következők szerint épül fel: A 2. fejezet a lehetséges keresési stratégiákat ismerteti, a 3. fejezet pedig ezek egyikéről, a keresőszoftverekről szóló, a felhasználó számára fontos ismereteket rendszerezi. Ezen belül külön alfejezetek foglalkoznak a keresőszoftverek három főbb funkcionális részével: a begyűjtő, az indexelő, valamint a kereső és rangsoroló egységekkel. A 4. fejezet néhány szempontot ad meg, amire érdemes odafigyelni, ha jól kereshető web oldalt szeretnénk írni. Az 5. fejezet a láthatatlan webbel foglalkozik, a 6. fejezetben pedig a cikket konklúzió zárja.

2. KERESÉSI STRATÉGIÁK

A weben lévő információk közül a szükséges ismeretek megtalálásában több stratégiát követhetünk:

2.1. A jónak vélt URL cím beírása

Számos esetben jó taktika a nem hivatalos www.társaság_neve.com általános címet kipróbálni (pl. www.oracle.com). Ekkor a társaság nyitólapjára kerülve, onnan a hiperhivatkozásokat, elágazásokat követve megtalálható a céggel kapcsolatos információ. A .com mellett használatosak még az .edu, .org, .net, .int, valamint a .gov, .mil általános domainek, egyéb esetekben a cím legtöbbször az ország kódjára végződik.

2.2. Tematikus keresők alkalmazása

A tematikus keresők a weben található dokumentumokra mutató hiperhivatkozások sokszor hierarchikus gyűjteménye, ahol tartalom szerint felépülő könyvtárakban kereshetünk. Az összegyűjtött anyag jellege szerint lehet akadémiai vagy egyéb szakgyűjtemény, illetve a közönség széles rétegeinek szánt, minél több szolgáltatást nyújtó üzleti, közszolgálati portál.
A tematikus keresők használata számos előnnyel jár:

- A kiválogatott dokumentumok témakörönként rendezve, csoportosítva találhatók meg.
- A témák széles köre szerepel az ilyen gyűjteményekben.
- A szakgyűjteményekbe a dokumentumokat a témák szakértői válogatják be, így biztosított, hogy minőségileg kontrolált anyagot találunk.
- Az összegyűjtött anyag sokszor kiértékelve, magyarázatokkal ellátva érhető el.
- A legfrissebb dokumentumok (pl. hírek) itt találhatók meg, hiszen a keresőprogramoknak idő szükséges, míg adatbázisukba beemelik, s így kereshetővé teszik ezen oldalakat.

A tematikus keresők a webnek szűkebb részét fedik le, mint amit a keresőszoftverek adatbázisai tartalmaznak.

2.3. Keresőszoftverek alkalmazása

A keresőprogramokat a weben lévő óriási mennyiségű információ lekérdezésére hozták létre.
Működésükre általánosan jellemző, hogy:

- a weben lévő dokumentumok valamely halmazát kiválasztják és összegyűjtik,
- a kiválasztott és begyűjtött dokumentumok egyes részeiből adatbázist építenek fel (indexelő rész),
- majd a felhasználók által megfogalmazott lekérdezéseket ezen adatbázisok alapján kísérlik meg megválaszolni.

Mikor célszerű keresőszoftvert használni?

- Ha nagyon speciális téma iránt keresünk.
- Ha a web oldalak millióinak teljes szövegében akarunk keresni.
- Ha nagy számú, érdeklődési körünknek megfelelő web oldalt szeretnénk visszakapni.
- Ha keresett dokumentumok típusára, forrására, nyelvére, keletkezésének időpontjára megszorításokat akarunk tenni.

A keresőszoftverek egy része saját adatbázisából keresi vissza a lekérdezésnél megadott szót, szavakat, kifejezést, de vannak olyan keresési szolgáltatók, amelyek egyszerre több keresőszoftver adatbázisában is keresnek (meta-search engines). Van arra is példa, amikor egy szolgáltató saját indexállományát több más keresési szolgáltatónak adja át, akik abból keresnek, vagy saját indexállományukat erősítik vele. (Pl. az Inktomi indexét használja többek között a LookSmart, HotBot, iWon, MSN).

Általában elmondható, hogy 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.

3. KERESŐSZOFTVEREK FELÉPÍTÉSE

Napjainkban a keresőszoftverek többsége üzleti jellegű, amely szabadalmazott technológiát is jelent egy erős versenypiacon. Ebből következik, hogy pontos működésükről annyit tudunk, amennyit a fejlesztők biztonságosnak tartanak közzétenni, s ez általában a technikai részletekről kevés publikációt jelent. Az egyik legrészletesebb nyilvános leírás a Google-ról érhető el [5].

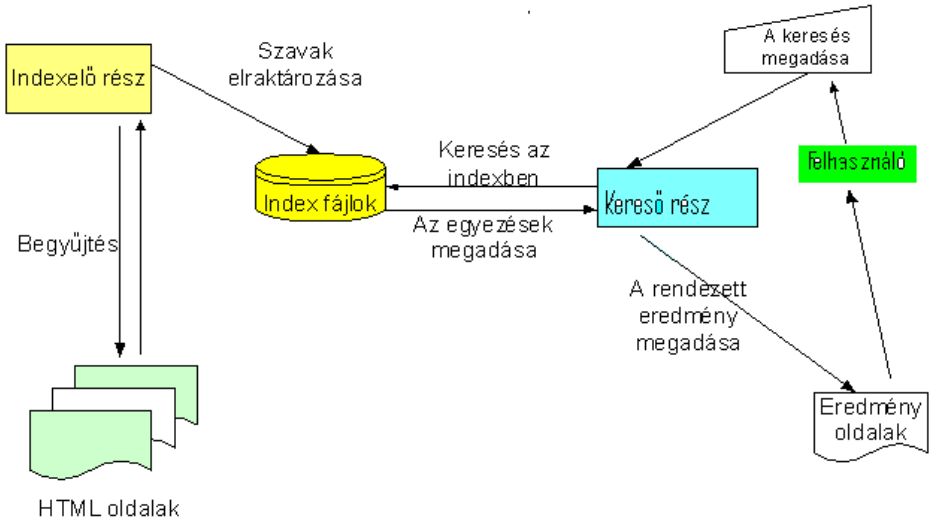
A keresőszoftverek működési elve: Adatbázisban tárolt indexeken alapul a keresés, amely adatbázisok a web robotok által begyűjtött dokumentumok alapján épülnek fel.

A *web robotok* olyan számítógépes programok, melyeknek célja a World Wide Web oldalain történő adatgyűjtés vagy keresés.

Jellemzőik:

- Önállóság: bizonyos előre meghatározott korlátok között a működésüket emberi beavatkozás és irányítás nélkül végzik.
- Rekurzivitás: egy adott pontról kiindulva képesek bejárni az összes olyan oldalt, amely a kiindulási ponttal közvetlen kapcsolatban áll, majd a kapcsolt oldalakat, mint kiindulási pontot tekinteni és a folyamatot egészen addig folytatni, amíg valamilyen kilépési feltétel érvényre nem jut.

A keresőszoftverek működését szemlélteti az 1. ábra.



1.ábra
A keresőszoftverek működésének általános folyamata

A keresőszoftverek a következő főbb funkcionális részekből állnak: a begyűjtő robot, az indexelő eljárás valamint a kereső és rangsoroló rendszer.

3.1. Begyűjtés

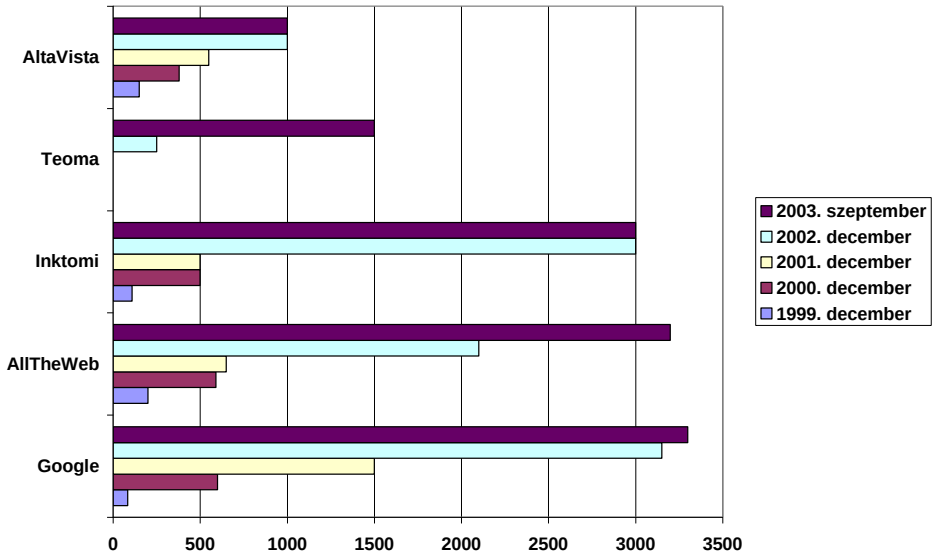
A begyűjtő részben a keresőszoftverek robot programjai az Interneten át elérhető dokumentumokat gyűjtenek, hogy azokból a keresők saját adatbázist építsenek fel. Minden keresőszoftvernek megvan a saját robotprogramja, amelyek különböznek abban, hogy a begyűjtésnél mely szervereket tekintenek kiindulási pontnak, az információ kigyűjtésének mely módszerét alkalmazzák, milyen frissítési periódust használnak.

A keresőszoftverek egy része csak HTML oldalakat gyűjt indexelésre, míg mások más típusú információk között is keresnek. Pl. gopher, WAIS, ftp, telnet (OPACs), UseNet News, IRC, különböző adatbázisok, multimédia termékek (kép, film, hang), egyéb típusok (pl. e-mail címek).
Különböznek a keresők abban is, hogy egy domainről kiindulva mélységi vagy szélességi bejárást alkalmaznak a domainről induló hivatkozások felkeresése során.
Számos kereső felső korlátot ad meg, hogy egy domainről hány dokumentumot indexel (pl. Alta Vista), így oly hatalmas dokumentum mennyiséget képviselő site-ok, mint pl. geocities.com vagy microsoft.com ezen keresők adatbázisában különösen alulreprezentáltak.
Fontos szempont, hogy milyen sűrűn írja felül saját adatbázisát egy kereső, újralátogatva azokat a helyeket, amelyeket egyszer már indexelt. Ez a paraméter jelentősen eltér a különböző keresőknél: egy héttől akár több hónapig is terjedhet, így ha közben módosulnak, megszűnnek az indexelt oldalak, a keresőszoftver adatbázisában még a régi adatok szerepelnek (halott linkek). Bizonyos keresők a gyakrabban módosuló fontosabb oldalakat sűrűbben látogatják és indexelik újra.
Valamennyi kereső lehetővé teszi, hogy a felhasználók saját nyitólapjukat regisztráltathassák, azaz web szerverük, nyitólapjaik URL címét átadják a kereső robotprogramjának, bár ez nem jelent automatikus bekerülést a kereső adatbázisába. A legtöbb keresőnél azonban különböző nagyságú regisztrációs összegek befizetése mellett biztosítható, illetve gyorsítható az adatbázisba való bekerülés. [1]
1998-ban publikálták az első eredményeket arról, hogy az egyes keresők a web hány százalékát indexelik [7]. Az adatok szerint az akkori 6 legnépszerűbb és legjobb eredményt felmutató keresőszoftver együttesen a webnek mintegy 60%-át fedte le, egyenkénti eredményük a következő volt:

Keresőszoftver	*A web hány százalékát indexeli?*
HotBot	34%
AltaVista	28%
Northern Light	20%
Excite	14%
Infoseek	10%
Lycos	3%

2.ábra
A web lefedettsége a legjobb eredményt mutató keresőszoftverek által (1998-as adat)

A következő évben a legjobb eredményt a Northern Light mutatta fel, de a webnek már csak 16%-át indexelte, s az akkori első 10 keresőszoftver együttvéve is a webnek csupán 42%-át fedte le.
A web növekedési ütemével a keresőket üzemeltető szolgáltatók nem tudnak lépést tartani. A legfrissebb adatok szerint [1] a jelenlegi legnagyobb indexállománya a következő keresőknek van:



3. ábra
A legtöbb oldalt indexelő keresőszoftverek mérete (millió web oldal)

3.2. Indexelés

Az indexelő részben a begyűjtött dokumentumok bizonyos elemeiből, szövegéből a keresőszoftver saját adatbázist épít fel, vagy tovább bővíti azt. A keresők különböznek azon szavak gyűjteményének nagyságában (lexikon, thesaurus), amelyeket ismernek, s amelyekből adatbázisukat felépítik. A Google esetében ez több mint 14 millió szót jelent. A keresők adatbázisában minden egyes szót (index) a koordináták egy halmaza reprezentál, amely leírja, hogy a keresett szó hol található (mely dokumentumban, bekezdésben, mondatban, címsorban, stb.).
Részletes leírások az indexelés folyamatáról nem érhetők el, de az általában ismert, hogy egyes keresők mit indexelnek.
A keresőszoftverek a HTML dokumentumok különböző részeiből építik fel adatbázisukat:

- A dokumentum nevét <TITLE> valamennyi kereső indexeli.
- A fejléc információkat, azaz a <HEAD> és </HEAD> közötti részt (pl. file információk, meta adatok) sok kereső beépíti az adatbázisába.
- Számos kereső egyéb HTML elemeket is indexel: például címsorokat, horgony elemeket (<A> és </A> közötti rész), kiemelt szövegrészeket.
- Míg kezdetben a keresők csak a dokumentum elejét, esetleg az első néhány bekezdést indexelték, ma már a legjobb keresők képesek a dokumentumok teljes szövegét indexelni.

Vannak olyan sűrűn előforduló szavak (pl. névelők, elöljárók, számok), amelyeket néhány kereső nem épít be adatbázisába (stop words). Ezen szavak listája keresőnként eltérő lehet.
Az URL címeket (host, könyvtár, file név) tartalmazzák a keresők adatbázisai, s néhány keresőnél keresési szempontként külön is megadható.

3.3. Kereső és rangsoroló rendszer

A keresőrendszer – a tényleges keresést megvalósító egység – a keresési kulcsszavak, kifejezések alapján a keresőszoftver adatbázisából veszi elő a dokumentumokat. A legtöbb kereső a "legjobb egyezés" elvét használja a visszakereséseknél, néhány csak a pontos egyezéseket tekinti eredménynek. Egy-két szolgáltató néhány pontosságot növelő tulajdonságot is használ: például a tárgyhoz tartozás visszajelzése, hasonlók keresése, találatok csoportosítása.

3.3.1. A keresés interfész

A kereső résznél lényeges, hogy a keresőszoftver milyen lehetőségeket nyújt adatbázisának lekérdezéséhez. Ma már a legtöbb keresőszoftver az egyszerű lekérdezéseken túl az összetett lekérdezések lehetőségét is biztosítja. A keresés megadásakor használható operátorok, kiegészítő feltételek keresőszoftverenként változnak, így egy-egy új alkalmazás használata előtt célszerű megismerni annak lehetőségeit.
Az összetett lekérdezéseknél használható operátorok, kiegészítő feltételek:

- Logikai operátorokkal (AND, OR, NOT) köthetjük össze a keresési kulcsszavakat. Elhagyásuk esetén több keresési kulcsszó megadásakor egyes keresők AND, míg mások OR műveletet vesznek alapértelmezettnek. A szó előtti + jel használatával előírhatjuk, a – jellel kizárhatjuk a szó jelenlétét az eredményhalmazban.
- A kulcsszóegyezéses kereséseken kívül kifejezések keresésére is lehetőség van a legtöbb keresőnél. Ekkor a szavak sorrendje lényeges és a kifejezést idézőjelek közé kell zárni.

- Néhány keresőnél a NEAR operátor használatával előírhatjuk, hogy egy adott szó környezetében kell előfordulnia a keresett szónak. A szavak közötti távolság keresőnként változhat: 2-től (WebCrawler) akár 25 szóig (Lycos) is terjedhet.
- Keresési megszorítások is megadhatók néhány kereső szolgáltatásnál. Előírhatunk a dokumentumok létrehozására vonatkozó időkorlátot, megadhatjuk a dokumentum nyelvét, típusát, méretét, előfordulási helyét domainenként, site-onként.
- Számos keresőnél kereshetünk csak a dokumentum nevében, a hiperhivatkozások között, illetve az URL alapján.
- Egy-két keresőnél lehetőség van a szavak végének levágására. Ekkor a *rádió** keresés eredménye egyaránt lehet a *rádió, rádiózás, rádióhullám* szavak valamelyikét tartalmazó dokumentum.
- Kereséskor érdemes arra is figyelni, hogy néhány kereső kis- és nagybetű érzékeny.

A felhasználók keresési szokásait vizsgálva megállapították, hogy a felhasználók túlnyomó többsége átlag 2 szavas egyszerű kereséseket használ, és egyáltalán nem tudja használni az összetett keresési lehetőségeket, logikai operátorokat. A keresések sikeressége az első lekérdezésnél 51%, a második alkalommal 32%, harmadjára 18%, azaz ha az első alkalommal nem találta meg a felhasználó azt, amit akart, összetett keresések alkalmazásával sem lesz nagyobb esélye sikeres találatokra. Talán éppen ezért a felhasználók mintegy fele az első lekérdezés sikertelensége esetén feladja a további próbálkozásokat. [14]

A lekérdezések megfogalmazásának könnyítésére egyre több keresőszoftver az operátorok explicit megadása helyett (vagy mellett) a megfelelő beviteli mező kiválasztását teszi lehetővé, tehermentesítve így a felhasználókat a lekérdezéseknél használt operátorok és használatuk szintaktikai szabályainak ismeretétől. A részletes keresést biztosító oldal sokszor a nyitólapon található hivatkozáson keresztül érhető csak el. Ennek az az oka, hogy sok felhasználó első ránézésre elbizonytalanodik a felkínált s számára bonyolult lehetőségektől.

3.3.2. Sorrend

Valamennyi kereső úgy rendezi a keresés eredményét, hogy az eredmény lista elejére az általa legfontosabbnak tartott dokumentumok kerüljenek. A rangsorolási algoritmusok keresőnként különböznek.

Egy oldal fontosságának meghatározásakor figyelembe veszik a kulcsszó előfordulásának helyét (a dokumentum nevében, címsorokban, az első bekezdésekben), valamint az előfordulás gyakoriságát. E feltételek meglétét különböző súlyozással tekintik a keresők.

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

A második generációs keresőszoftverek újításai a keresési eredmények megjavítására:

- A keresés irányultságának fogalmi felismerése.

Ide tartozó területek az alkalmazott jelentéstan, a természetes nyelvi feldolgozás. E technikákat alkalmazó két legjelentősebb kereső az Ask Jeeves és a Northern Light.

Az Ask Jeeves adattárában számos, különböző tárgyterületek szakértői által előre megfogalmazott kérdés van, amelyekhez hozzárendeli a legrelevánsabb válaszoldalakat. A kereső elemzi a felhasználó által feltett kérdést, majd a hozzá legjobban hasonlító előredefiniált kérdés alapján szolgáltatja a válaszokat.

Northern Light a keresés eredményét a természetes nyelvi feldolgozó alkalmazásával csoportosítja site és/vagy tartalom alapján, s a felhasználó e csoportok közül választhat.

- Hivatkozások analízise

A Google elsősorban a dokumentumra mutató külső hivatkozásokkal számol, s a népszerűbb oldalról történő külső hivatkozásokat súlyozottan veszi figyelembe. Algoritmusa annyira hatékony, hogy általánosan elfogadottan a Google-t tartják a legjobb keresési eredményeket nyújtó szolgáltatásnak.

- Az oldal népszerűségének elismerése

Számos kereső az eredménylista sorrendjének meghatározásánál figyelembe veszi az eredményoldalak népszerűségét. Pl. Google, DirectHit.

A DirectHit egy oldalt azon elv alapján rangsorol, hogy egy egyszerű keresés eredménylistájából hányan választják az adott oldalt.

A második generációs keresőszoftverekre jellemző, hogy egy oldal helyezését a rangsorban előnyösen befolyásolja az oldal népszerűsége. Ha ehhez hozzávesszük, hogy a felhasználók ritkán böngésznek a keresési eredménylista második oldalán túl, akkor az a várható trend, hogy az eddig is népszerű oldalak még népszerűbbek lesznek, az új oldalaknak pedig egyre nehezebb lesz jó helyezést elérni.

A rangsor befolyásolására tett kísérletek büntetése

Valamennyi keresőszoftver küzd a webmesterek különböző mesterkedései ellen, amellyel megpróbálják oldalaik fontosságát megemelni. Ez számos módon történhet, az ötletek kifogyhatatlanok:

- A kulcsszavak vég nélküli ismétlésével, ami a böngésző 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 többszörözésével, amelyek közül csak az elsőt jelenítik meg a böngészők, de a robot valamennyit indexeli.
- A tartalom duplikálásával vagy ugyanazon oldal többszöri előterjesztése, vagy ugyanazon tartalom más hoston való elhelyezése által.
- A nyitólapra mutató „mesterséges linkek” elhelyezésével. Ilyenek az olyan oldalak, amelyek egyetlen tartalma egy link egy másik oldalra, vagy amelyek elsődleges szándéka a felhasználót egy másik oldalra átirányítani.

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

3.3.3. Eredmények

Rendszerint a keresési eredmények listái tartalmazzák a dokumentum címét (TITLE), helyét (URL), összegzést róla, néha a dokumentum létrejöttének az időpontját (néhány keresőszoftver esetében ez az adatbázisukba való bekerülés időpontja), vagy a dokumentum méretét.
A dokumentumok összegzésénél különböző szemléletek figyelhetők meg: Számos keresőszoftver a szerzők által megadott meta adatokat használja az összegzésnél, vagy a dokumentum azon mondatát jeleníti meg, amely a lekérdezéssel egyezést mutat, esetleg az oldal szövegének első 100-200 karakterét írja ki. 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.

4. JÓL KERESHETŐ WEB OLDALAK

A keresőmotorok tulajdonságainak ismerete nemcsak a lekérdezések minél hatékonyabb megfogalmazásánál, s ezáltal a találati eredmények leszűkítésénél segít, de hasznos a webre szánt dokumentumok elkészítésénél is. Ha tudjuk, hogy a kereső szoftverek mely tulajdonságok alapján szerepeltetik az eredménylistán előkelőbb helyen az egyes oldalakat, akkor a webre szánt oldalak optimalizálásával, kereshetőbbé tételével több emberhez juttathatjuk el az információt, s ez az üzletmenetben keményen éreztetheti a hatását. Nem véletlen

tehát, hogy számos cég foglalkozik a honlapok minél inkább kereshetővé tételével.
A weben át elérhető információ egy része strukturált formában van tárolva. (pl. relációs adatbázisok), amelyek a megszokott eszközökkel (pl. SQL alkalmazásával) lekérdezhetők.
A HTML–ben írt szöveges dokumentumok félig strukturáltak, ahol a strukturálást, s evvel a lekérdezhetőséget a különböző meta elemek, jelölő tagek alkalmazásával érjük el. A HTML-t követő XML–ben különböző névtereket alkalmazva fokozhatjuk a szöveges állományaink strukturáltságát, s ezáltal kereshetővé tételét, ma még azonban a keresőszoftverek döntő része nem indexeli az XML tageket.

A dokumentum neve

Érdemes figyelmet fordítani a webre szánt dokumentumunk nevére. A <TITLE> elemmel közbezárt részt minden keresőgép indexeli, és visszakereséseknél is számos keresőgép előrébb rangsorolja azokat a dokumentumokat, ahol a keresési kifejezés a dokumentum nevében előfordul. Célszerű tehát a semmitmondó *Honlap* név helyett például *Kényelem Nagykereskedelmi Kft: cipők, csizmák, szandálok, táskák, övek és egyéb bőráruk forgalmazója* nevet használni. Az eredménylistán számos kereső az azonos fontosságú dokumentumokat név szerint abc sorrendben jeleníti meg, érdemes tehát erre is figyelni.

Meta jelölő elemek

HTML dokumentumok leírásánál használatosak a HTML meta jelölő elemek, amelyek segítségével különböző információt adhatunk át az indexelő résznek. Ezen elemeket a böngésző programok nem jelenítik meg, s megadásuk mindig a dokumentumok fejlécében történik. A legtöbb keresőgép különböző mértékben használja fel ezen elemeket. A két leggyakrabban használt meta elem a *description* és a *keywords*.
A description segítségével a dokumentum tartalma írható le. Számos keresőgép az eredménylistában – ha a dokumentum tartalmaz ilyen meta elemet – ezt az összegzést adja vissza, így mi magunk adhatjuk meg dokumentumunk lényegi leírását, s evvel kelthetjük fel a kereső felhasználó figyelmét.
A keywords alkalmazásával a dokumentum tartalmára vonatkozó kulcsszavak adhatók meg. A kulcsszavak megadásánál érdemes minél több, a dokumentumra illő kulcsszót feltüntetni, gondolva a szinonimákra, egyes-többes számra, általánosításokra, stb. Ezen elem különböző visszaélésekre ad lehetőséget, így indexelését számos szolgáltató nem támogatja (Nem támogatja: Google, AllTheWeb, AltaVista, támogatja: HotBot, Inktomi, Teoma.)
Ahhoz, hogy egy meta jelölő nyelv elterjedjen és sikeres legyen, fontos a keresők támogatása. Ezt példázza a Dublin Core esete is, amelyet a könyvtárosok igényeinek megfelelően hoztak létre 1995-ben. A Dublin Core 15

meta jelölőelemet tartalmazó készlettel (pl. kulcsszó, leírás, szerző, kiadó, forrás, azonosító, szerzői jogok, stb.) adja meg valamely dokumentum bibliografikus leírását és ezáltal kísérli meg feljavítani a weben való kereshetőségét. A gyakorlatban azonban alig használják ezeket a jelölő elemeket, s a főbb keresők egyike sem indexeli őket.[18]

5. A LÁTHATATLAN WEB

A láthatatlan web részét képezik azok a weben át elérhető dokumentumok, információk, amelyek a keresőprogramok számára láthatatlanok, azaz nem épülnek be a keresők adatbázisába.
Alapvetően két oka van, hogy a keresők nem indexelnek egy web oldalt: egyrészt technikai korlát, amely gátolja az oldal elérését, másrészt a szolgáltató tudatos döntése vagy választása, amely kizárja az oldalt a kereső adatbázisából. Ideiglenesen a láthatatlan web részét képezik a legfrissebb dokumentumok (pl. hírek, újdonságok leírásai), amelyeket a keresőprogramok még nem indexeltek.

5.1. Technikai korlátok

Nagyon sok adatbázis érhető el a weben keresztül, de tartalmuk a keresőprogramok számára láthatatlan. Az adatbázisokban tárolt adatok lekérdezését valamely lekérdező nyelv (SQL, OQL) teszi lehetővé. A lekérdezés megadásakor valamely karaktersorozatot be kell gépelni vagy tulajdonságok egy sorozatát ki kell választani, amire a begyűjtő robotok képtelenek.
Lekérdezhető webes adatbázisok tartalma, azaz a dinamikusan generált web oldalak teljesen vagy részlegesen láthatók vagy láthatatlanok egy keresőmotor számára attól függően, hogy mennyit tartalmaz abból egy hiperhivatkozás által elérhető statikus oldal. Ha nem mutat egy hiperhivatkozás a web oldalra, akkor az a robot számára láthatatlan.
A fentiekből következik az is, hogy valamennyi webes adatbázis tartalma, amely felhasználói nevet és jelszót igényel a használótól, a keresők számára elérhetetlen.
A láthatatlan web részét képezik az intranet hálózatokon tárolt dokumentumok is.

5.2. Elvi okok

Léteznek olyan oldalak, amelyeknek kizárása a keresőszoftvert üzemeltető szolgáltató irányelveinek alapján történik. Ezen oldalak elérésének nincs

technikai akadálya, de tartalmuk, minőségük, formátumuk miatt mégis kizárásra kerülhetnek.

A formátum miatti kizárás különböző okai:

A keresőprogramok és a robotok HTML nyelvű programok olvasására vannak optimalizálva, hisz ez a web alapnyelve. A szinte teljes egészében képeket tartalmazó oldalak gyakran kerülnek kizárásra, mert az oldalak nem tartalmaznak szöveget, leírást, ami alapján a keresők indexelhetnék az oldal tartalmát. A képek leírására használható ALT elemet, valamint az image mapben megadott dokumentumokat sem indexeli minden kereső. Néhány kereső nem támogatja a frame szerkezetet, ezért a framekben megadott dokumentumokat sem indexeli. (pl. Excite, Fast).

A különböző felhasználói programok által készített dokumentumok (pl. PDF, Word, PowerPoint fájlok) tartalmának indexelésére a legtöbb keresőnek hosszabb időre volna szükség, ezért eleve kizárják ezen típusokat az adatbázisukból.

A strukturálatlan állományok, azaz a programok, képfájlok, animációs állományok, hangfájlok esetén is nehéz vagy egyelőre lehetetlen relációt megadni a fájl tartalma és belső struktúrája között, hacsak valamely magyarázó szöveggel nem adjuk meg a fájlok tartalmának leírását. Sehogy vagy nehezen - valamilyen konverziót igénylő - indexelhetőségük miatt kerülnek kizárásra ma még a legtöbb esetben ezek a fájlok.

Az egyes szolgáltatók döntése alapján kerülnek teljes vagy részleges kizárásra a különböző scripteket tartalmazó oldalak, amelyeknek URL-je tartalmazza a kérdőjelet. (pl. http://www.cre8asiteforums.com/viewtopic.php?t=1130) Ennek oka az, hogy a kereső robot nem tudhatja biztosan, nem egy végtelen ciklust eredményező csapda-e a scriptet tartalmazó oldal, amely a fenntartónak idő- és pénzveszteséget eredményezne, így megelőzés végett kerül sor az oldal kizárására.

A láthatatlan web részét képező források, adatbázisok elérését segítik a rendszerezett hivatkozások gyűjteményei. Ilyen található például a *www.invisible-web.net* címen.

6. ÖSSZEFOGLALÓ

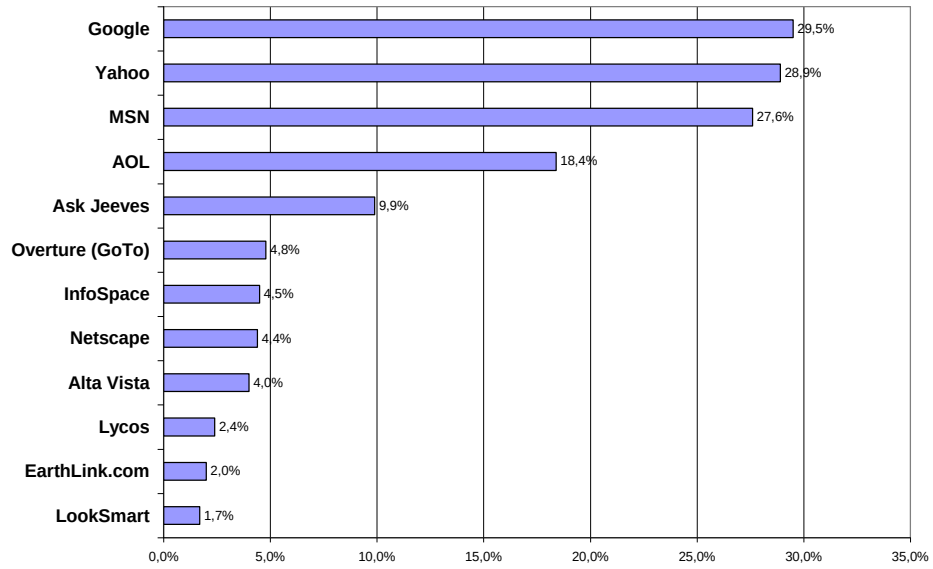
A keresőszoftverek tulajdonságainak ismerete segít a döntésben, mely keresőszoftvert célszerű használnunk a kívánt információ felleléséhez, illetve hogyan ajánlatos megfogalmazni a lekérdezésünket, hogy minél jelentősebb, értékesebb eredményoldalakat kapjunk.

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, így a hajdan volt különbségek napjainkban a felhasználók számára már nem érzékelhetők.
A kereséskor használt szolgáltatók népszerűsége időről-időre változik. A kezdeti időkben kedvelt szolgáltatók ma kevésbé népszerűek (HotBot, AltaVista), míg később létrejött kis cégek mára kiemelkedő eredményekkel bírnak (Google, Teoma).

A jelenlegi trend:

A 4. ábra azt mutatja, hogy 2003. januárjában az USA-ban melyek voltak a legnépszerűbb keresők, azaz a felhasználók hány százaléka látogatta meg a különböző keresőszoftverek illetve tematikus keresők oldalait egy-egy információ fellelése érdekében [1]. (Mivel egy felhasználó több oldalt is meglátogatott, a részeredmények összege meghaladja a 100%-ot.)



4. ábra
A legnépszerűbb keresési szolgáltatók az USA-ban (2003. januári adat)

Egy-egy szolgáltató népszerűségét erősen befolyásolja, hogy mely témájú, nyelvű, típusú dokumentumok vannak nagyobb számban reprezentálva indexállományában vagy rendszerező könyvtárában, meghatározó lehet a megszokás, s a 3. és 4. ábra adatait összevetve kevésbé lényeges az indexállományának mérete. Természetesen az egyes országokban a különböző keresési szolgáltatók népszerűsége eltérő**.**

A weben át elérhető dokumentumok jóval nagyobb részét, becslések szerint 400-500-szorosát alkotják azok a dokumentumok [13], amelyek megtalálásában a keresőszoftverek nem segítenek minket. A láthatatlan web nagyon gyors mértékben növekszik, így tartalmának lekérdezhetővé tétele sürgető feladat.

7. IRODALOMJEGYZÉK

[1] Search Engine Watch *http://searchenginewatch.com*
[2] WebReference *http://www.webreference.com*
[3] Search Engine Showdown *http://www.searchengineshowdown.com*
[4] SearchEngines.com *http://www.searchengines.com/*
[5] 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*
[6] S. Lawrence, L. Giles,: Accessibility and Distribution of Information on the Web, *Nature, Vol. 400, pp. 107-109, 1999*
[7] S. Lawrence, L. Giles: Searching the World Wide Web, *Science, April 3, 1998*
[8] Development of a European Service for Information on Research and Education *http://www.lub.lu.se/desire*
[9] Recommended Sites and Search Techniques *http://library.albany.edu/internet/search.html*
[10] Hu, Chen, Schmaly, Ritter: An Overview of World Wide Web Search Technologies, 2001 *http://www.eng.auburn.edu/users/wenchen/publication/overview.ps*
[11] T.Perinotti, How Search Engines Work *Microsoft Interactive developer, January 1997*
[12] Invisible Web *http://www.lib.berkeley.edu/TeachingLib/Guides/Internet*
[13] BrightPlanet *http://www.brightplanet.com*
[14] J. Nielsen, Search: Visible and Simple *Alertbox, May 13, 2001* http://useit.com
[15] What Is Natural-Language Searching? *http://www.nwc.com/1120/1120f1side2.html*
[16] W3C Metadata Area *http://www.w3.org/Metadata/*
[17] A Dictionary of HTML META Tags *http://vancouver-webpages.com/META*
[18] Dublin Core Metadata *http://dublincore.org*
[19] Platform for Internet Content Selection (PICS) *http://www.w3.org/PICS/*
[20] Breadth first search and depth first search ICS 161: Design and Analysis

of Algorithms, Lecture notes for February 15, 1996
http://www1.ics.uci.edu
[21] C. Morris, So What's the Problem?
http://www.wdvl.com/Internet/Dead_SearchEngines
[22] K. Princz Mária: Systems to access information in the Web
MicroCAD'2000 International Computer Science Conference, pp 169-173
[23] K. Princz Mária – Rutkovszky Edéné: Ismeret reprezentáció a weben
Networksho, 2002
[24] K. Princz Mária - Husi Géza: A webes keresők használatának tanítása
Informatika a felsőoktatásban, 2002
[25] K. Princz Mária: A weben lévő információk hozzáférhetősége
NetworkShop, 2003
[26] K. Princz Mária: A webes keresők tanításának tapasztalatai
e-Learning alkalmazások a hazai felsőoktatásban, 2003