A webkeresés fejlődése

A web meghatározása

A világhálónak, a World Wide Webnek két jól elkülöníthető eleme van: a *látható* és a *láthatatlan* web. Ahhoz, hogy ezek jelentőségét az információkeresésben megértsük, tudnunk kell, hogy kétféle weboldal létezik. A *statikus* weboldalak manuálisan készülnek, és bárki elérheti őket. A *dinamikus* oldalakat a számítógép hozza létre scriptek (gyakran CGI, Java, Perl) segítségével. Ezek a scriptek a statikus oldalon információt kérő vagy közlő felhasználó és egy, az adott információt szolgáltató vagy feldolgozó adatbázis között közvetítenek. A statikus weboldalak ugyanazt a generikus információt, míg a dinamikus oldalak egyedi, a felhasználó specifikus követelményeinek megfelelő információt nyújtják mindenkinek. A mindenki számára hozzáférhető, a webkeresők által indexelt, statikus weboldalak összessége alkotja a látható webet. A láthatatlan web az azonosítást igénylő, az indexelésből a robotokat kizáró metacímkével (robot exclusion meta tag) kizárt weboldalakból és azokból az adatbázisokból áll, amelyek csak időlegesen vannak jelen a weben dinamikus weboldalak formájában.

A látható web nagyságára vonatkozóan nincsenek tudományosan megalapozott adataink. Egy 1999. februári becslés 800 millió oldalról szól, ami 6 terabájt adatot jelent. A látható webnél jobban strukturált és indexelt, nagy értéket képviselő adatbázisok is a weben érhetők el. A professzionális információ tehát sokkal inkább a web útján, mintsem a weben érhető el, bár a látható web is jelentős mértékben hozzájárul az emberi tudás terjesztéséhez.

Webkalauzok és webkeresők

A WWW-n történő keresésben a webkeresők (keresőgépek, keresőmotorok, search engines) mellett a webkalauzok (directoryk, webkatalógusok) is fontos szerepet játszhatnak. Ezek weboldalak előre meghatározott listái, amelyeket emberek szerkesztenek (nem gépi úton készülnek), téma (tárgy) szerint vannak rendezve, és szelektívek. Ennek megfelelően minőségi forrásokat tartalmaznak, és népszerűek is a felhasználók körében. Egy-egy ilyen kalauz használatakor a felhasználó az adott szakrend szerint navigálhat, vagy kereshet is a kalauz tételei között. Nagyobb kalauzokhoz keresők is kapcsolódnak, hogy segítségükkel másodlagos adatokat nyerhessünk, ha a felhasználó a kalauzban nem talált megfelelő dokumentumot.

Tágabb értelemben véve minden olyan weboldal, amely kapcsolatok (ugrópontok, linkek) valamilyen rendezett listáját tartalmazza, kalauznak tekinthető. Egy személyes kezdeményezésből született kalauz, az *Open Directory* projekt a piacvezető Yahoo! jelentős versenytársává nőtte ki magát. Az üzleti információs kalauzok közül kiemelkedik a *Business Information Sources on the Internet (http://www.dis.strath.ac.uk/business)*, valamint a *Business Researcher Interests (www.brint.com/interest.html)*.

Hogyan működnek a keresők?

Amikor egy keresőt használunk, indexelt weboldalak adatbázisában keresünk. Minden keresőnek három fő eleme van:

- a webhelyeket vizsgáló spider,
- a megvizsgált weboldalak indexe/adatbázisa,
- a keresőszoftver.

A spiderek olyan programok, amelyek automatikusan és nagy gyakorisággal indulnak el, hogy új oldalakat keressenek a weben, indexeljék az ezeken az oldalakon található szavakat és csatolókat, és összekapcsolják az indexelt szavakat az adott oldal URL-jével. A kereső legfontosabb része, amellyel a felhasználó is kapcsolatba kerül, az index. Ezeket régebben hasonló elvek alapján szerkesztették, vagyis elsődlegesen a szavak helye és gyakorisága határozta meg, mely szavakat tekintenek relevánsnak. 1998-tól azonban, az új keresők megjelenésével ez megváltozott. Az új keresők közül a Direct Hit az adott oldalak „népszerűsége", a Google az oldalak és helyek közötti kapcsolatok száma alapján indexel, míg a *Real Names* térítéses szolgáltatás, amely lehetővé teszi, hogy cégek kulcsszavakat regisztráltassanak márkáik és identitásuk védelmében.

Minden keresőnek van saját, testre szabott szoftverje, amellyel az adatbázisban keres. Alapvetően azonban működésük hasonló elveken alapszik. A találati listában szerepelni fog bármely webhely, amely tartalmazza a felhasználói kérdésben megfogalmazott szót, kifejezést (szavakat, kifejezéseket). A találatok relevancia szerinti rendezése olyan algoritmusokon alapszik, amelyek a szavak

helyét és gyakoriságát elemzik. Árnyalatnyi eltérések vannak e tekintetben a keresők között, és ezek okozzák, hogy ugyanazt a kérdést feltéve más és más eredményeket kapunk a különböző keresőkben. A különbségeket azonban még inkább az okozza, hogy a keresők közötti tartalmi átfedés viszonylag kicsi.

Portálok

A portálok először azoknál a szolgáltatóknál jelentek meg, amelyek vagy csak internetkapcsolatot nyújtottak, vagy a böngészők alapértelmezésben beállított honlapjai (a Netscape-nél ez a Netcenter, az Explorernél az MSN). Az utóbbiak esetében a felhasználók sokszor nem tudták vagy nem akarták átállítani az alapbeállítást. Hasonló helyzet volt a csak keresést kínáló webhelyeken. A keresett dokumentum megtalálása után már nem maradt tovább az olvasó az adott oldalon. Ha viszont a felhasználó számára más, további értékes információkat kínálnak, kialakul a portál, amely csábító lehet a felhasználók számára. Közben a keresők és a kalauzok közötti különbségek némileg elmosódtak.

A keresési módszerek fejlődése

A keresők 1998 táján megjelent új generációjához tartoznak a metakeresők, amelyek lehetővé teszik, hogy a felhasználó több kereső segítségével keressen párhuzamosan. A legnépszerűbb metakeresők: a *Dogpile* (*www.dogpile.com*) 14 különböző keresőben és kalauzban keres, de nem szűri ki a duplikált találatokat. A *Mamma* (*www.mamma.com*) hét keresőben keres, a találatokat saját relevancia-szempontjai szerint rendezi, és a duplumokat is kiszűri.

Az adott hely népszerűségének szempontja a *Direct Hit* (*www.directhit.com*) keresővel jelent meg. A gyakran látogatott webhelyeknek nincs külön keresőjük, saját indexszel, amelyet közvetlenül el lehetne érni. Itt a keresési eredmények másodlagos elemezéséről van szó, amelyet a meglevő keresőkhöz kapcsolnak.

A keresett karaktersor betű szerinti megfeleltetése mellett megjelentek a természetes nyelvi keresést alkalmazó, az emberi nyelv szemantikáját is figyelembe vevő keresők is. Ezek vagy korábbi kérdések adatbázisával vetik össze az adott kérést, és kínálják fel választásra a legmegfelelőbbnek talált formát, vagy szintaktikai elemzést végeznek.

Az *Ask Jeeves* (*www.askjeeves.com*) a felhasználói kérdéseket 7 millió mintakérdéssel veti össze. Ha nem talál egyezést, a felhasználónak a legközelebbi alternatívát kínálja fel. Egyúttal metakeresést is végez több kereső indexében.

Az *Electric Monk* (*www.electricmonk.com*) szintaktikai elemzést végez. Az első generációs keresők nem vették figyelembe az egyes webhelyek közötti kapcsolatokat, pedig az azonos alakú szavak, a rokonértelműség és a mondatszerkezetek okozta problémák úgy is leküzdhetők, hogy a bibliometriai módszerekhez hasonlóan figyeljük a kapcsolatokat. Így releváns, minőségi találatokat kaphat a felhasználó, hiszen a kereső a weboldalak szerkesztőinek értékítéletét veszi alapul. A kapcsolatok elemzésével jóval több weboldal tárható fel, mint az emberi munkával összeállított kalauzokban.

A keresők első generációja egy-egy oldal tartalmára koncentrált, és nem nagyon vette figyelembe a különböző oldalak közötti kapcsolatokat. Emellett a természetes nyelvi jellemzők, így a mondatszerkezetek, a szinonimák és az azonos alakú (de eltérő jelentésű) szavak kezelését nem sikerült megoldani. Ezért kínált új lehetőségeket mindezen problémák kiküszöbölésére a weboldalak közötti kapcsolatok elemzése. Ennek segítségével azonosíthatjuk egy-egy szakterület legfontosabb, a felhasználói igényeknek leginkább megfelelő, releváns információforrásait.

A *Google* (*www.google.com*) éppen arra épít, hogy ezeket a kapcsolatokat figyelembe vegye. Ha ugyanis egy-egy oldal szerzője más, általa fontosnak tartott oldalakra mutat, azokat valamilyen szempontból jónak, fontosnak tartja. Ennek köszönhetően sokkal több weboldalt tudnak elemezni, mint a Yahoo! típusú, a forrásokat emberek közreműködésével összegyűjtő kalauzok. Sőt, a Google szakemberei azt állítják, hogy nagyméretű indexeken még a szokásosnál is jobb eredményeket tud a szoftver produkálni. A Google a kapcsolatok körül található szövegeket is feldolgozza. A kapcsolatok elemzése néhány olyan kereső relevanciamegállapítási algoritmusában is szerepet játszik, mint az Excite és a HotBot.

A *Clever* projekt *(www.almaden.ibm.com/cs/k53/clever.html)* is a weboldalak hivatkozásait veszi alapul, és elemzi a kapcsolatok közelében található szövegeket. A kérést azonban egy keresőszolgálatnak, például az AltaVistának küldi el, majd a tőle kapott eredmények (tipikusan mintegy 200 oldalnyi) halmazán végzi el az elemzést. Két kate-

góriát használ: egységes fogalmi listák (authorities), amelyekre sokan mutatnak rá, tehát mérvadónak tekinthetők, és gyűjtőoldalak (hubs), amelyek – a portálokhoz hasonlóan – egy-egy területen fontosnak tartott oldalakra hivatkoznak, azok listái, forráskalauzai. Míg a Google az adott felhasználói keresőkérdéstől függetlenül megtartja az oldalak rangsorát, a Clever az adott keresés kontextusát figyelembe véve mindig új rangsort állít fel. A *Focused Crawler*, amely nincs még úgy kidolgozva, mint a Clever, már a begyűjtéskor megpróbálja a legrelevánsabb weboldalakat figyelembe venni.

A *hírcsoportok* (news groups) anyaga implicit tudást, azaz tapasztalatot, kreativitást és ötleteket képvisel, ami a tudásalapú társadalom egyre fontosabb összetevője. Ennek megfelelően a hírcsoportok anyagában való keresést szolgáló eszközök is egyre fontosabbak lesznek. A legismertebb ilyen kereső valószínűleg a *Deja News* (*www.dejanews.com*). Egy sor válogatott hírcsoport anyagát böngészhetjük segítségével, vagy kereshetünk egy meghatározott hírcsoportra, témára vagy üzenetre; szerzőre, dátumra, nyelvre, és az eredmények megjelenítése is beállítható.

A *Reference.com (www.reference.com)* emellett webes hirdetőtáblák és (e-mailen alapuló) vitafórumok (levelezőlisták) anyagában is lehetővé teszi a keresést. A *Liszt's Newsgroup Directory (http://liszt.com/news)* a Deja Newst használja, de nagy, saját listája van a levelezőlistákról és az IRC-csatornákról.

Egyre nagyobb szerepet kapnak a tematikus, azaz egy-egy szűkebb-tágabb szakterület anyagát feltáró keresőszolgálatok is. A keresők hatékonyabb használatát a kérdéseket interpretáló kiegészítő szoftverek segítik. Ezek közül az *1jump* (*www.1jump.com*) céginformációk és céghírek részletes keresését szolgálja.

A weben található információ közel 70%-a nem szöveges. Ezért egyre fontosabbak a multimédiakeresők. Ilyen keresők a *Ditto* (*www.ditto.com*), a *Scour* (*www.scour.net*), valamint az *AltaVista PhotoFinder* (*www.altavista.com*).

A szerverek képességeinek kihasználása mellett a felhasználók gépein is futtathatunk keresőprogramokat, amelyek többségükben ingyenesen letölthetők, de többet tudó változataikért fizetnünk kell. A *Mata Hari* (*www.thewebtools.com*) megtanulja a keresőkérdést, és azt az egyes keresők nyelvére lefordítja. A *BullsEye Pro* (*www.intelliseek.com*) 11 intelligens ágenst (robotot) foglal magában. A különböző ágensek 450 forrásban keresnek a látható és a láthatatlan weben. Automatikusan lefuttatja a kereséseket, és lehetővé teszi, hogy kereséseket importáljunk vagy exportáljunk más felhasználók gépeire, illetve gépeiről. A *Copernic* (*www.copernic.com*) különböző keresők nyelvére fordítja kérdésünket, és azokon szimultán le is futtatja a kereséseket keresőkön, forráskaluzokban és adatbázisokban. Lehetőség van arra is, hogy 20 kategóriában előre meghatározott forrásokat keressünk vele.

Az ilyen és hasonló eszközök előnyeit felismerve néhány keresőszolgálat letölthető programokat kínál, amelyek segítségével többet tud nyújtani, mintha csak böngészőnkkel, ezek nélkül érnénk el a keresőt. Ilyen szolgáltatást nyújt az Infoseek (Infoseek Express), az AltaVista (AltaVista Discovery) és a Lycos (See More).

Az *XML* megjelenésével nagy lehetőségek nyílnak meg a keresés fejlettebb formái előtt. Ehhez azonban megegyezés szükséges, hogy egy-egy szakterületen milyen egységes címkézést használjanak. Ezt a címkézést aztán a keresőszolgálatoknak is használniuk kell, és lehetőséget kell nyújtaniuk, hogy választhassanak a szöveges keresés és a mezőkben való keresés között.

A jövőben elképzelhető lesz, hogy a minőségi információk szolgáltatásáért a keresők kisebb (mikro-) összegeket kérnek majd. Várható az is, hogy a keresés kiterjed majd a mobil kommunikáció különböző formáira is.

A keresőszolgálatok profiljai

Az *AltaVista (www.altavista.com)* 1995 decemberében indult. Az egyik legnagyobb keresőszolgálat.

Az *Ask Jeeves* 1998 júniusa óta működik mint az első természetes nyelvi kereső. Ennek megvalósításához a felhasználó kérését egy 7 millió mintakérdést tartalmazó adatbázis tartalmával hasonlítja össze. Emellett metakereséseket végez több más keresőben is.

A *Direct Hit* 1998 áprilisa óta üzemel, és a kérdések népszerűsége alapján rangsorolja a forrásokat. Több keresőszolgálat is használja.

Az *Excite* (*www.excite.com*) 1995 végén indult. Nagy indexébe nemcsak webes, hanem más, így céginformációs anyagokat is integrál. Felvásárolta Magellan és WebCrawler nevű két korábbi versenytársát.

A *Fast* (*www.altheweb.com*) 1999 májusában jött létre, és – az URL-jében jelzetteknek megfelelően – az az ambíciója, hogy az egész webet indexelje. Létrejöttekor a legnagyobb keresőszolgálat volt, amely 200 millió weboldalt indexelt. Ez a szám 1999 végére 300 millióra emelkedett. A keresőszolgálatok többségével szemben nem nagyszámítógépen, hanem néhány száz összekapcsolt PC-n fut.

A *Go* (*www.go.com*) az 1995-ben indult Infoseek változata, új névvel. A keresés mellett emberi erőfeszítéssel összeállított tematikus listát is nyújt.

A *Google* a cég becslése szerint 70–100 millió oldalt indexel, ami a felhasználóknak mintegy 300 millió oldal elérését teszi lehetővé. Ezzel a Google a legnagyobb kereső.

A *HotBot* (*www.hotbot.com*) 1996-ban indult. 1998-ban felvásárolta a Lycos, de önállóan működik. Nem készít saját indexet, hanem az Inktomi indexét használja, az elsődleges eredményeket a Direct Hitből veszi.

Az *Inktomi* (*www.inktomi.com*) nem kereshető közvetlenül, de több keresőszolgálat használja.

A *LookSmart* (*www.looksmart.com*) emberek összeállította forráskalauz, amelyet az AltaVista és az Excite is használ.

A *Lycos* (*www.lycos.com*) 1994-ben indult. 1999-ben áttért a tematikus listaszolgáltatásra. Elsődleges eredményeit az Open Directoryból, a másodlagosakat saját indexéből veszi.

A *Northern Light* (*www.northernlight.com*) 1997-ben indult, és egyike a legnagyobb indexeknek. Népszerű a kutatók körében.

Az *Open Directory* (*http://dmoz.org*) 1998-ban indult, majd hamarosan felvásárolta a Netscape. Emberek katalogizálta források kalauza.

A *RealNames* (*www.realnames.com*) díj ellenében regisztrál cégeket, hogy neveik és termékeik, márkáik nevei kereshetők legyenek a RealNamest szolgáltató keresőkben (AltaVista, Go).

A Yahoo! (*www.yahoo.com*) 1994-es indulása után egyike a legnépszerűbb keresőszolgálatoknak. A weben a legnagyobb, emberek által összeállított kalauzt tartalmazza, mintegy 1 millió oldalt feldolgozva.

/GREEN, David: The evolution of Web searching. = Online Information Review, 24. köt. 2. sz. 2000. p. 124–137./

(Koltay Tibor)

Portálok, keresők és a Math-Net

Az 1990-es évek közepétől a matematika több szakterületén létesültek saját webszerverek, amelyek az intézményekről, munkatársaikról, a kutató- és oktatómunkáról tartalmaznak információkat. Fokozódó mértékben használták a webet a tudományos eredmények publikálására, mindenekelőtt preprintek formájában. Ennek egyik példája az *MPRESS* preprintindex *(http://www.mathnet.preprints.org)*, amely a kilencvenes évek közepe óta a *Harvest* szoftverrel *(http://www.math-net.de/project/tools/harvest/index.html)* gyűjtötte a preprinteket. A *Math-Net* a matematika egyik információs és kommunikációs rendszere a weben. Tartalma személyek és intézmények önkéntes inputján alapul.

A Math-Net webszolgáltatás olyan előlapot alakított ki, amely a matematikával foglalkozó intézmények változó struktúrájú weboldalakon megjelenő kínálatát egységes formában mutatja be a következő (további alfejezetekre oszló) fejezetekben: General (Általános információk), People (Személyi információk), News (Hírek), Research (Kutatás), Teaching (oktatás), Information Services (Információs szolgáltatások). A Math-Net oldal alkalmas a begyűjtött adatok automatikus kiértékelésére is. A Math-Net oldalak a Math-Net Page Creator *(http://www.math-net.de/project/tools/pagecreator/index.en.html)* segítségével könnyen létrehozhatók. Ezek csak kiegészítik az intézmények helyi weblapjait, de nem korlátozzák szerkezetüket és megjelenési formájukat.