Keresőmotorok a Hálón

Dong és *Su* tanulmánya vizsgálat tárgyává teszi a WWW-alapú adatbázisokat, összeveti őket a hagyományos (online és CD-ROM) adatbázisokkal, és kiértékeli a *keresőmotornak* (search engine) nevezett web-segédeszközöket. A korábbi vizsgálódásokkal szemben az értékelés valóságos (real-life) használók valóságos keresőkifejezéseivel készült. A kutatók kiindulópontja az a megállapítás, hogy a releváns információk visszakeresése a világhálón egyre nehezebb és bonyolultabb az Internet-források óriási mérvű szaporodása és a szolgáltatások keveredése folytán.

A keresőmotor

A hatékonyabb böngészés igénye hozta létre a *keresőmotor* és a tudásrobot (knobot) nevű eszközöket. A keresőmotor program, amely adatbázisokat keres végig, és a weben a robot által gyűjtött HTML dokumentumokat pásztázza. A keresőmotor három összetevője a *robot*, az *adatbázis* és a *közvetítő* (agent).

A robot

A robot vagy web-vándor olyan program, amely a WWW információs térben járkál. A web-oldalakon beágyazott hipercsatolókat (hyperlinks) kihasználva mozog egyik web-dokumentumtól a másikig. Visszakeresésre a HTTP protokollt használja. Felkutatja a web új forrásait, a kulcsszavas kereséshez indexeli a web-lapokat, s kiszűri az elhalt csatolókat. A különböző robotok különböző stratégiát használnak utazásuk során. A *Lycos* pl. minden nap végigpásztázza a WWW-t, a Gophert és az FTP szervereket. Az *Alta Vista* web-oldalakat és *hírcsoportokat* (news group) néz végig. Ez a stratégia nagyban meghatározza az adatbázisokból visszakeresett és elérhetővé tett információk minőségét és mennyiségét.

Az adatbázis

A robot a felkutatott információkat indexeli és adatbázisba rendezi. Az adatbázis tartalma lehet web-cím, cím, fejléc, szavak, első sorok, szövegkivonat, de akár teljes szöveg is. Az adatbázisok képesek több millió web-oldal tárolására. Van olyan keresőmotor, amely több adatbázist is készít (pl. *Lycos*), s természetesen a bennük lévő csatolók nagysága megszabja a találatok mennyiségét. Az adatbázisban tárolt információ frissítése kumulatív vagy reprodukáló lehet. A *Lycos* pl. az új URL-ek (web-címek) feltérképezését kumulatív módon végzi, hozzáadja azokat a már meglévő adathalmazhoz. Az *Excite* ezzel szemben hetente küld ki gyűjtőprogramot az újdonságokért, és az összegyűlt adatokkal a teljes adatbázist újraépíti. A *WebCrawler* egyesíti a kétféle keresést: az újdonságokat hetente hozzáragasztja az adatbázishoz, havonta pedig teljes adatcserét végez.

A közvetítő

Amikor a felhasználó keresni kíván, a közvetítő keresési felületet ad, majd kilistázza a találatokat. A lista rendezett: a legrelevánsabb találatokkal kezdődik.

Keresőmotor-típusok

Önálló, tárgyszavas és *meta*-keresőmotort különböztetünk meg. Más osztályozás szerint a keresőmotorokat adatgyűjtési elveik különböztetik meg egymástól.

Önálló keresőmotorok

Teljes szövegű és nem teljes szövegű adatbázist is tartalmazhatnak. Az automata robotot használók végigpásztázzák a web-teret, ahová csak beengedik őket. A dedikált robotot alkalmazók viszont a webnek csak egy bizonyos szeletét kutatják.

Meta-keresőmotorok

Kombinált csoportos vagy szimultán keresőmotorok, amelyekkel a felhasználók egy időben több keresőmotort is igénybe vehetnek a visszakereséshez. (A *MetaCrawler* pl. egyszerre nyolc keresőmotorral dolgozik.) A típus problémája, hogy nem kapnak teljes hozzáférést a használt keresőmotorok valamennyi eszközéhez, így a találatok kevésbé pontosak lesznek, a keresési idő pedig megnő.

Tématárak (Subject directories)

Az Internet-források kereshető, böngészhető, hierarchikus indexei. Az információkeresést tárgyszavakkal segítik; több közülük (pl. *Yaho!* és *Infoseek*) kulcsszavas keresést is lehetővé tesz. A téma szerinti keresés a web rendezetlensége miatt nagy segítség.

A web-alapú adatbázisok speciális tulajdonságai a hagyományos adatbázisokkal szemben

Adatbázis-tartalom

- az információ szelekciója és feldolgozása miatt az online adatbázisok és a CD-ROM-ok jobb minőségű és jobban strukturált információt tárolnak;
- az Internetről származó információk véletlenszerűek, esetlegesek, minőségük és érvényességük nem meghatározható.

Indexelt mezők

- online/CD-ROM adatbázisoknál az indexelés ellenőrzött szótárak, tezauruszok alapján történik, emberi szelekció révén;
- web-dokumentumoknál az indexelés automatikus; bármely mező indexelésre kerül, és némelyik teljes szöveget indexel (pl. *Altavista*).

A kivonatolás módszerei

- fontos szempont, mert ennek alapján dönti el a felhasználó, hogy az adott találat megfelel-e információs szükségletének;
- az *Excite* pl. automatikus technikát alkalmaz, teljes mondatokkal, de nem jelöli a méretet, a dokumentum dátumát stb.

Keresési technikák

A web-keresés hátterében bonyolult hipertext kapcsolatokon való „ugrálás" folyik, a használó rengeteg opció közül választhat folytatást. Csaknem lehetetlen kétszer ugyanazt a bonyolult keresést végrehajtani. A hagyományos adatbázisok kereséséhez képest probléma a keresési formula is. Az *Excite* pl. teljes, leíró mondatokat is lefogad, a *Magellan*nál viszont végletesen le kell egyszerűsíteni a kérést.

Megjelenítés, rendezés

A web-keresés előnye a találatok „súlyozása", mérlegelése: a keresett szó előfordulási száma, illetve a szó pozíciója a dokumentumban megszabja, hogy mennyire releváns a találat (minél gyakoribb az adott szó a dokumentumban, annál kevésbé releváns a találat).

Keresőmotorok összevetése

Lycos

Kifejlesztője: *Carnegie Mellon University.* Alapja robot alapú C program, amely naponta átlag 10 000 dokumentum átnézésére képes. A legnagyobb és legerősebb adatbázisokkal rendelkezik, kiválóan alkalmas szokatlan és homályos témák keresésére. A keresőkifejezéseket automatikusan csonkolja.

WebCrawler

Kifejlesztője: *University of Washington, Seattle.* A teljes világhálót átvizsgálja, és a népszerű helyekről vett dokumentumokat tárolja. Felhasználóbarát interfésze van, gyors válaszidő jellemzi. Új felhasználók számára kitűnő segédeszköz a weben való kereséshez.

WWW Worm

A WWW-t kereső eszközök egyik úttörője, de nehéz hozzáférni. Amellett már elavult és alacsony relevanciájú eredményt produkál.

Alta Vista

Kifejlesztője: *Digital Research Lab.* Scooter nevű robotja naponta mintegy 2,5 millió web-lapot néz át. A legátfogóbb eredményeket produkálja a legnagyobb precizitással.

Excite

Tematikusan osztályozott témakörökként és kulcsszavakkal kereshető (16 témakör). Hátránya, hogy nincs benne lehetőség a Boole-operátorok alkalmazására, s hogy megjelenítéskor nem mutatja az URL-címeket. Egyik különlegessége az Excite Reviews, amely kb. 60 000 web-hely értékelését néhány mondatban megadja, ugyancsak témakör szerinti bontásban.

Infoseek

Kifejlesztője: *Infoseek Corporation, California.* A web-oldalak átfogó indexe, ingyenes hozzáférést nyújt újságokhoz, folyóiratokhoz: Számos, weben nem elérhető adatbázishoz eljuthatunk vele.

Yahoo!

Ugyan nem önálló keresőrendszer (az Alta Vistára épül), de az egyik legismertebb és legnépszerűbb tematikus kereső a Hálón. Erőssége az indexelés, és a források hierarchikus elrendezése témakörök szerint. A generikus kategórián belül alkategóriák és kulcsszavak alapján is lehet benne keresni. Mivel a Yahoo!-ban csak beküldött lelőhelyek szerepelnek, a helyek minősége bizonytalan. A több szóból álló keresésnél automatikus az ÉS-kapcsolat és a csonkolás. E-mail címek keresését is végzi.

/NOTESS, G. R.: Comparing net directories. = Database, 20. köt . 1. sz. 1997. p. 61–64.
DONG, X.–SU, L. T.: Search engines on the World Wide Web and information retrieval from the Internet: a review and evaluation. = Online & CD-ROM Review, 21. köt. 2. sz. 1997. p. 67–81./

(Koreny Ágnes)