WORLD WIDE WEB KERESŐRENDSZEREK FEJLŐDÉSE

Johanyák Zsolt Csaba

Keresés egy WWW szerveren belül

A felhasználók életének megkönnyítése érdekében az első lépést a WWW szerver programok készítői tették meg beépítve rendszereikbe a kulcsszó alapján történő keresés lehetőségét. A keresetőség megteremtéséhez az alapot a WWW lapok készítésénél használt HTML nyelv <ISINDEX> mezője biztosítja. A fenti mezővel megjelölt szavak kulcsszavakká (indexekké) válnak. A böngésző program címsorában az URL cím után egy kulcsszavat és kérdőjelet megadva letölthetjük az adott szót tartalmazó lapot a szerverről. A módszer hátránya, hogy csak teljes kulcsszó egyezés esetén működik és csak azokat a szavakat kereshetjük, melyeket a lap készítője előre indexként kijelölt.

A fejlődés második fokozatát a WAIS protokoll WWW-be történő integrálása jelentette, ami lehetővé teszi az ún. teljes szöveges indexelést, vagyis azt, hogy a lapokon található minden szó keresési kulcsként felhasználható legyen.

Nem kellett sokáig várakozni a kezelés egyszerűsítésére sem. A Common Gateway Interface (CGI) kifejlesztése az űrlapok megjelenését eredményezte, melyek kitöltésével a felhasználó WWW böngészője segítségével adatokat küldhet a távoli szerveren a háttérben futó programok számára. Az interfész jelentősége túlmutat az egy szerveren belüli keresés kérdéskörén. Segítségével lehetővé vált egyszerű és könnyen kezelhető felület készítése adatbázisok és egyéb szolgáltatások igénybevételéhez.

Szolgáltatók - szerver alapú keresés

Az előző módszer nagy hátránya, hogy a felhasználó részéről előismereteket feltételez, azaz legalább azt tudni kell, hogy a keresett dokumentum melyik szerveren található. A problémára több megoldás is született. Egyrészt az újdonságokról beszámoló levelezési listák (pl. wagnerur), melyre feliratkozva az érdeklődő minden, a lista karbantartójának bejelentett, új WWW lapról értesül. Másrészt megjelentek a címetek tartalmazó adatbázisokat üzemeltető szolgáltatók. A felhasználó ezen információ-gyűjteményekben kutatva tudhatja meg, hogy hol található, hogyan érheti el az őt érdeklő dokumentumot vagy szolgáltatást.

A tárolt adatok nagy mennyisége miatt és a gyors visszakeresés érdekében ezeket általában szakterületek, témakörök szerint egy hierarchikus rendszerbe csoportosítják. A legtöbb keresőrendszer szolgáltatásaiért nem kell fizetni, vagy csak egy-két különleges szolgáltatás van díjhoz kötve. A fenntartást egyrészt szponzorok, másrészt reklámbevételek útján oldják meg. Az adatbázisok létrehozásának módja szerint két fő csoportot különböztethetünk meg, a regisztrációs bejegyzésen és a robotokon alapuló szervereket.

Regisztráción alapuló kereső-adatbázisok

Ezek képezik a kereső szerverek első generációját. Mint ahogy azt az elnevezés is sugallja egy ilyen adatbázisba jelentkezés alapján lehet bekerülni, azaz aki azt szeretné, hogy WWW lapja, szolgáltatása szerepeljen a nyilvántartásban az rendszerint egy űrlap kitöltésével megküldi a szükséges információkat az adatbank üzemeltetőjének. Ezt regisztrációnak nevezik. Néhány általános és speciális keresőrendszer neve és címe szerepel az 1. táblázatban.

1. táblázat

Regisztráción alapuló adatbázisok

Általános kereső szolgálatok	Speciális kereső szolgálatok
Clearinghouse http://www.clearinghouse.net	Apollo http://apollo.co.uk
CUI Catalog http://cuiwww.unige.ch/w3catalog	Ariadne http://ariadne.inf.fu-berlin.de:8000
Dino http://www.dino-online.de	BizWiz ! http://www.bizwiz.com
Galaxy http://www.einet.net	Britannica Online http://www.eb.com
Interlinks http://www.nova.edu/Inter-Links	DIB http://www.branchenbuch.de
IntIndex http://www.silverplotter.com/intindex/intro.htm	Four 11 http://www.four11.com
Magellan http://www.mckinley.com	Gamelan http://www.gamelan.com
Planeth Earth http://planetearth.net/SanDiego/index.html	Internet Address Finder http://www.iaf.net
SDSC http://www.sdsc.edu/sdsc/Geninfo/Internet	InterNic http://www.internic.net
Starting Point http://www.spt.com	Linkstar http://www.linkstar.com
The Whole Internet Catalog http://www.gnn.com/gnn/wic/index.html	Logos http://www.logos.it
WEB.DE vroom.web.de	Mesh http://www.ip.net/cgi-bin/mesh
WWW Virtual Library http://www.w3.org/hybertext/DataSources/bySubject/Overview.html	SFB Glossary http://wings.buffalo.edu/SBF/thumbtabs.html

Robot alapú keresőrendszerek

A robotok olyan programok, melyek a Web szövevényén végighaladva automatikusan feldolgoznak minden dokumentumot. Egy URL-höz érve letöltik az általa jelölt lapot, majd azt feldolgozva a benne szereplő kapcsokon haladnak rekurzívan tovább. A robot feladatai közé tartozik a halott, már nem létező dokumentumra mutató kapcsok felkutatása, a tükrözések felismerése, a szerveren belüli, relatív címzés átalakítása abszolúttá, az indexek, azaz bejegyzések előállítás az adatbázis számára. A rendszer egy óriási gráfnak tekinti a Web-et, ahol a lapok képezik a csomópontokat, míg a kapcsok az éleket. A letöltést párhuzamosan futó ügynök programok hajtják végre. Ezek átadják a kereső gépnek a lapokat vagy

hibaüzenetet halott kapocs, illetve átviteli hiba esetén. A kereső gép minden letöltött dokumentumon egy lexikális elemzést hajt végre, melynek célja a tartalomra nézve jellegzetes, fontos kifejezések kiemelése és elhelyezése az adatbázisban. Az elemzés kiterjedhet a teljes szövegre, címre, kivonatra, a dokumentumot tartalmazó gép INTERNET címére, vagy csak ezen adatok egy részére.

3. táblázat

Robot alapú keresőrendszerek

Alta Vista http://altavista.digital.com	Lycos http://lycos.cs.cmu.edu
DejaNews http://www.dejanews.com/forms.dnq.html	Open Text http://www.opentext.com:8080
Excite http://www.excite.com	RBSE http://rbse.jsc.nasa.gov/eichmann/urlsearch.html
Flipper http://frp.cs.tu-berlin.de/flipper/index.html	WebCrawler http://wc3.webcrawler.com
InfoSeek Guide http://www2.infoseek.com	WWW Worm http://www.cs.colorado.edu/WWW/
Inktomi http://inktomi.berkeley.edu	Yahoo! http://www.yahoo.com

A szolgáltatás igénybevevője egy felhasználói felületen keresztül kapcsolódik az adatbázishoz. A Háló rendszerezett végigbongészésének eredményeképpen idővel szinte minden lap jellemzői automatikusan bekerülnek a nyilvántartásba. Ez olyan mértékű gép és tárolási kapacitást igényel, hogy már most is a legtöbb keresőrendszer több, egymással összekapcsolt gépen fut. A mindent átható Szövet csomópontjainak száma azonban exponenciálisan növekszik, és evvel a tempóval nem igazán tudnak lépést tartani a kereső robotok, ezért, ha valaki mindenképpen be szeretné juttatni a lapjára vonatkozó információkat, akkor nem árt, ha közvetlenül felhívja rá a szolgáltató figyelmét. Egyes rendszereknél ez történhet e-mail útján, míg mások letölthető űrlapokat bocsátanak az érdeklődők rendelkezésére.

A robot alapú keresők megjelenése a szolgáltatások körének bővülését eredményezte. A kereső kérdések megfogalmazásakor lehetőség van logikai és közelség (proximity) operátorok, ellenőrzött mezők felhasználására, valamint egyes esetekben a szavakat súlyozni is lehet. A találatlista relevancia alapján sorba rendezve jelenik meg. Természetesen vannak gyenge pontok is. A keresés során a szavak közti szemantikai kapcsolat nincs figyelembe véve, ami a nagyméretű adatbázisok következtében magas találatszámot eredményez, és ebből a felhasználónak kell nem kis energiával és időráfordítással az értelmes eredményeket kiválogatnia. A fontosabb robot alapú keresőrendszerek címei és jellemzői megtalálhatóak a 2. táblázatban

A robotok megjelenése jelentős fejlődést eredményezett a keresőrendszerek által nyújtott szolgáltatásokban, de emellett negatív hatásai is vannak. Az ügynökök folyamatosan újabb és újabb lapokat töltenek le, ami egyrészt erősen leterheli a hálózatot, másrészt csökkenti a felkeresett WWW szerverek teljesítményét. Ez utóbbi kivédésére létrejött egy egyezmény (Standard for Robot Exclusion), ami azon alapszik, hogy a szerver könyvtárrendszerében elhelyeznek egy speciális állományt, amelyből a robotok megtudhatják, hogy mely dokumentumokhoz, vagy a könyvtárrendszer mely részéhez férhetnek hozzá. Természetesen ez csak addig jelent védelmet, míg olyan robot programok készülnek, amelyek betartják ezt az előírást.

A Web fejlődési üteme és a hálózat leterhelése megkérdőjelezi a robotok jövőjét. A keresőrendszerek következő generációinak a találatok jelenlegi magas számából eredő problémákat is meg kell majd oldania. Olyan keresési eljárások megjelenése várható, melyek figyelembe veszik a kulcsszavak közti szemantikus és hierarchikus kapcsolatot is.

Aliweb - egy alternatív megoldás

A robot alapú keresőrendszerek hátrányainak felismerése különböző alternatív megoldások megjelenéséhez vezetett. Az Aliweb (<http://web.nexor.co.uk/aliweb/doc>) az Archie ötletén alapszik, innen ered elnevezése is (Achie Like Indexing the Web). A rendszerhez csatlakozott WWW szerverek helyileg hozzák létre, kezelik és tárolják az indexadatokat. Meghatározott időközönként a keresőszolgálatok átveszik ezeket az információkat. A hálózat terhelése csökken, mivel nem a teljes dokumentumokat, hanem csak a strukturált indexállományokat kell átvinni. A rendszer hátránya, hogy a szerzőknek vagy a helyi adminisztrátoroknak kell létrehozni és karbantartani az index adatokat.

Köszönetnyilvánítás

A témához kapcsolódó kutatásban jelentős segítséget nyújtott a Collegium Hungaricum által biztosított Bécs-i kutatói ösztöndíj.

Irodalomjegyzék

- [1] Bekavac, B.: Suchverfahren und Suchdienste des World Wide Web, Nachrichten für Dokumentation, 1996/4, pp. 195-213.
- [2] Koster, M.: WWW Robots, Wanderers and Spiders, <http://info.webcrawler.com/mak/projects/robots/robots.html>
- [3] Koster, M.: Robots in the Web: threat or treat?, <http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html>
- [4] Koster, M.: ALIWEB Archie-Like Indexing The Web, Proceedings of the first International World Wide Web Conference, Geneva, May 1994.
- [5] Egyre gyorsabbak, de nem mindig jobbak a webes keresők, PCWorld, 1996 október, pp. 38-43.

Johanyák Zsolt Csaba, főiskolai adjunktus, Gépipari és Automatizálási Műszaki Főiskola,
Informatika Tanszék, H6001 Kecskemét Pf. 91.

Tel.: -36-76-481 291, Fax: -36-76-481 304, e-mail: csaba@gandalf.gamf.hu