A láthatatlan web tudományos részének feltárása

Bevezetés

A láthatatlan web definiálása után *Bergman* publikációjának [1] tévedéseit mutatjuk be, és helyesebb becslést adunk a méretre vonatkozóan. Végül a láthatatlan web láthatóságának megoldására következik javaslat: együttműködés a szereplők között.

Az elmúlt évek tapasztalatai is megerősítik a felhasználók igényeit a naprakész, komplett és integrált végfelhasználói keresőszolgáltatásokra; még az akadémiai szektorban is, ahol pedig számtalan eszköz, adatbázis áll rendelkezésre, követve az interneten megjelenő tudományos tartalmak növekvő tendenciáját. Egyre több információforrás alakult ki, a felhasználói igények és szokások pedig gyökeresen megváltoztak. Mindez a web láthatatlan részének létrejöttéhez vezetett. A web egy része ugyanis láthatatlan a keresőmotorok számára, nem érik el a tartalmait, vagy csak igen kevéssé és gyenge minőségben.

Könyvtári gyűjtemények és adatbázisok tartalmai maradnak rejtve az elterjedt keresőszolgáltatások előtt. Figyelembe véve az egyre több digitalizálási projektet, valamint a Z39.50, az OAI-PMH és hasonló szabványok alkalmazásának hiányát, kijelenthetjük, hogy a láthatatlan web mérete folyamatosan növekszik.

De mi is pontosan a láthatatlan web, és vajon mekkora lehet a mérete?

A láthatatlan web definiálása

Sherman és *Price* [2] szerint a láthatatlan webet olyan hiteles, nívós és interneten keresztül elérhető szöveges oldalak, fájlok alkotják, amelyeket az általános célú keresők technikai korlátaik vagy hiányzó akaratuk miatt nem tesznek kereshetővé. Ez a meghatározás elég tág, (pl. a hiányzó akarat miatt a spamoldalak is ide érthetők), ezért megpróbálták összeállítani a láthatatlan web típusait. Ilyenek például

- azok az oldalak, amelyeket a rájuk mutató linkek hiányában a keresőrobotok nem fedeznek fel;
- az indexálható szöveg nélküli, csak képeket vagy egyéb médiafájlokat tartalmazó oldalak, vagy flash oldalak;
- az adatbázisok tartalmai;
- a valós időben keletkező tartalmak, amelyek gyors változásuk miatt nem kereshetők;
- a dinamikusan előálló tartalmak.

Bergman meghatározásában [1] az adatbázisokra helyezi a hangsúlyt, szerinte ugyanis az a láthatatlan web, amelynek tartalmait a keresők addig nem láthatják, amíg azok egy specifikus keresés eredményeképpen nem állnak elő dinamikusan.

A szabad és a védett tartalmak közötti különbséget és a tudományos tartalmak sajátosságait szem előtt tartva, a tudományos láthatatlan webet így lehetne meghatározni: a tudományos élet számára releváns adatbázisok és gyűjtemények tartalmai, melyek elérhetetlenek az általános keresők számára.

A tudományos láthatatlan webet leginkább szöveges fájlok alkotják, méghozzá a legkülönbözőbb fajtájúak (PDF, DOC, PS, PPT stb.) és tartalmúak (szakirodalom, on-line tartalom stb.), ezért a tudományos láthatatlan web csak egy része a teljes láthatatlan webnek. Ennek a résznek az elérhetővé tétele egyedül nem lehetséges, csak összefogással valósítható meg. A tudományos élet következő szereplőinek kell együttműködnie:

- adatbázis-szolgáltatóknak a megfelelő metaadatok előállításával és ember általi indexeléssel,
- könyvtáraknak lehetővé téve és nyílt rendszerrel segítve az ember általi indexelést (pl. OPAC),
- üzleti szereplőknek még több szöveges tartalom biztosításával,
- különböző társasági, szabadon hozzáférhető és egyéb adattáraknak.

A láthatatlan web mérete

A láthatatlan web méretével kapcsolatban a szakirodalomban Bergman becslése [1] az uralkodó. A 60 legnagyobb ismert láthatatlan webes oldal adataiból kiindulva, és feltételezve, hogy 100 ezer láthatatlan weboldal létezik, Bergman szerint 400-szor, vagy akár 550-szer is nagyobb lehet a láthatatlan web, mint a látható.

Bergman a becslésnél az adatbázisok átlagos rekordszámát használta fel, ami óriási szám: 5,43 millió (a top 60 adatbázis összrekordszáma 85 milliárd). Ám azt már nem vette figyelembe, hogy az adatbázisok mérete aszimmetrikus, például csak az első kettő teszi ki a top 60−75%-át. Megvizsgálva adatbázisok listáját tartalmazó katalógusokat, például a DIALOG-nál látható, hogy az aszimmetrikus eloszlás tipikusnak tekinthető. Ezért helyesebb lenne a rekordszámok középértékével számolni a félrevezető átlag helyett (Bergman top 60-as listájánál ez csak 4950 rekordot jelentene).

Bergman a láthatatlan web méretét tárterületben is megbecsülte, szerinte az mintegy 7500 TB információt tartalmaz. Ez a hatalmas szám két tévedés eredménye lehet. Az első az átlaggal való számolás, a második pedig az adatbázisok méretéből való következtetés helytelensége az aszimmetria miatt.

A láthatatlan web mekkora része lehet tudományos vonatkozású? Bergman listájának 90%-a, de ezek többsége pusztán feldolgozatlan adatokat tartalmaz, mint például szatellit-felvételeket a földről. Ezeket kihagyva pusztán csak 4%-ot kapunk. Ennek a kisebb résznek a mérete tárterületben mérve nehezen becsülhető meg külön, mivel a szöveges adatbázisok mérete általában lényegesen kisebb a képeket tartalmazókétól.

A láthatatlan web méretének pontosabb becsléséhez az adatbázisok egy részletes és megbízható gyűjteményére lenne szükség. Mindenestre, 60-nál biztosan több adatbázist kell vizsgálni, például a Gale-gyűjteményt [3]. A tudományos vonatkozású adatbázisok többségét is magában foglaló, hozzávetőlegesen 13 000-es lista összesen 18,92 milliárd dokumentumot tartalmazhat, átlagosan 1,15 millió rekordot adatbázisonként. Az aszimmetria miatt a legnagyobb méretűeket kihagyva az átlag rekordszám 150 ezer. Ezzel az átlaggal számolva, és külön hozzáadva a legnagyobbakat, a tudományos vonatkozású láthatatlan web mérete 20 és 100 milliárd dokumentum közé tehető. A Gale-listán sajnos nem szerepel az összes adatbázis Bergman top 60-as listájáról, ezért egyrészt tág ez a becslés, másrészt nehezen mérhető össze Bergmanéval. Ha a feldolgozatlan adatokat nem számítjuk, akkor az előbb becsült érték nyilván sokkal kisebb.

A láthatatlan web láthatóvá tétele

Többféle modell létezik a probléma megoldására, de most csak négy kerül említésre, melyek különböző fajtájú tudományos tartalmakat tesznek elérhetővé.

A Google Scholar (*http://scholar.google.com/*) nemzetközi tudományos, műszaki és orvosi kiadók több millió dokumentumát teszi kereshetővé, valamint a Crossref.org-on keresztül csatlakozott kiadókét. Sajnos kevés információ áll rendelkezésünkre a Google Scholar működéséről, és a kereshetővé tett tartalmakról.

A Scirus (*http://www.scirus.com/*) a FAST technológiára épülő tudományos kereső, amely leginkább a látható web tudományos részét indexeli. Közel 250 millió rekorddal a Scirus messze a legnagyobb kereső a hozzá hasonlók között.

A BASE (*http://www.base-search.net/*) szintén a FAST technológiára épülő tudományos kereső, amely a *Bielefeldi Egyetem Könyvtárá*nak és 160 egyéb szabad hozzáférésű adattárnak összesen mintegy 2 millió rekordját teszi kereshetővé.

A Vascoda (*http://www.vascoda.de*) német könyvtárak és dokumentációs központok együttműködésével létrejött kereső, amely több tudományterülethez kapcsolódó könyvtári gyűjteményt, szakirodalmi adatbázist és egyéb tartalmakat tesz kereshetővé angol és német nyelven. FAST technológiára épülve a keresőfelület az alatta lévő rétegeket fogja össze, minden tudományterülethez tartozó réteg ugyanis saját, külön is elérhető doménnévvel és (kereső)felülettel rendelkezik.

A láthatatlan web fontosságából, méretéből és a fenti projektekből is látszik, hogy a tudományos tartalmak láthatóvá tétele csak összefogással lehetséges. Egyedi kezdeményezés, illetve az általános célú keresők alkalmazása nem elég hatékony ezen a területen, nem vezet, nem vezethet célra. A tudományos élet szereplőinek kell tehát együttműködniük a láthatatlan web (tudományos

tartalmainak) láthatóvá tételéhez; ebbe az üzleti világ szereplői is bevonhatók.

A láthatatlan web, illetve a tudományos vonatkozású része további vizsgálatokat igényel a pontosabb becslések, valamint a keresőmotorok hatékonyabb működése érdekében.

Irodalom

[1] BERGMAN, M. K.: The deep web: surfacing hidden value. = Journal of Electronic Publishing, 7. köt. 1. sz. 2001. http://www.press.umich.edu./jep/07-01/bergman.html

[2] SHERMAN, C.−PRICE, G.: The invisible web: Uncovering information sources search engines can't see. = Information Today, Medford, NJ. 2001.

[3] WILLIAMS, M. L.: The state of databases today: 2005., Gale Directory of Databases, 2. köt. Gale Group, Detroit, MI. 2005. p. XV-XXV.

/LEWANDOWSKI, Dirk–MAYR, Philipp: Exploring the academic invisible web. = Library Hi Tech, 24. köt. 4. sz. 2006. p. 529–539./

(Somogyi Tamás)