Tóth Erzsébet

Az internetes keresők tárgyköri fogalomrendszere

Az írás áttekinti a magyar nyelvű szakirodalom lényeges megállapításait az internetes keresők tárgykörében, rávilágítva annak gazdag és összetett fogalmi hálójára. Tárgyalja az internetes keresők fogalmi és terminológiai kérdéseit, rámutat az itt előforduló fogalmi kapcsolatokra. A szerző szerint egy ilyen jellegű szakirodalmi áttekintés hasznos kiindulópontként szolgál a tárgykör tanulmányozásához a felsőoktatási gyakorlatban.

Bevezetés

Korunk számos folyamata közül csak a legjelentősebbeket emelném ki, amelyek a következők: a politikai, katonai értelemben vett egyhatalmú világ kialakulása, a demokratizálódás, az európai integráció, a fokozódó ázsiai gazdasági együttműködés, a technológia forradalma, a globalizáció, az információs és a fogyasztói társadalom kialakulása. Az első és az utolsó három a világ minden részére közvetlenül, vagy közvetve ható, globális folyamatnak tekinthető. Napjainkban a „globális" jelzőt több, tartalmában eltérő jelenségre is rá lehet húzni, azonban helyesebb, ha csak a gazdaság, a tőke, az ipar, a kereskedelem, a szolgáltatás földrajzi és módszertani terjeszkedését nevezzük globalizációnak.

Az információs és kommunikációs technológia (IKT) rohamos fejlődésének és konvergenciájának eredményeként a társadalomban egy új életforma, újszerű működés és viselkedés alakult ki. Új értékrendek jöttek létre. Ezt a széles körben elterjedt új életmódot, magatartást, információs technológiára épülő gazdaságot nevezzük információs társadalomnak [2]. Az „információs társadalom" kifejezés az 1960-as évek második felében jelent meg, az '50-es '60-as évek fordulóján jöttek létre a „tudástársadalom", „tudásgazdaság" stb. összetételek. Az „információs társadalom" a világban fellelhető információk általános gazdagságát tükrözi, míg a „tudástársadalom" arra a gazdagságra hívja fel a figyelmet, amit a tudás teremt, és arra a szegénységre, amelyet a tudástársadalom viszonyai között a tudás hiánya okoz [6]. Az információs társadalom kialakulása országonként különböző időszakban és ritmusban zajlik. A társadalom tág értelemben vett fejlettségi szintje és a kultúra nagymértékben befolyásolja azt. Lényeges, hogy erre a folyamatra sokkal „könnyebben" lehet hatni, társadalmi szinten jó irányba terelni, mint a globalizációra. A fogyasztói társadalom megjelenése a globalizációhoz és az információs társadalomhoz kapcsolódó harmadik jelentős folyamat [2].

A könyvtárak viszonylag korán felismerték annak fontosságát, hogy meg kell felelniük az információs társadalom kihívásainak. Az *Európai Unió* könyvtárakkal kapcsolatos tevékenysége az *EU 3-4. Kutatás-Fejlesztés*i, azon belül *Telematikai Keretprogram*ja keretében ment végbe. 1991–1994 között a 3. keretprogramban három pályázati felhívás jelent meg, melyek eredményeképpen 81 ún. akcióterv kezdődött el. Ebből mintegy kétszáz intézmény részvételével 51 közösen finanszírozott projektre került sor. 1995–1998 között a 4. keretprogram nyitott volt a közép-kelet-európai országok számára is. Két pályázatot írtak ki, 15 kutatási projekt, 7 összehangolt közös nagy munka és 20 újabb, ún. horizontális támogatási program indult. A könyvtárakkal kapcsolatos európai uniós programok részletes ismertetését lásd [1]-ben.

A globalizáció gyorsuló és könyörtelen versennyel jár együtt. Csak úgy lehetünk versenyképesek, ha az információs társadalom legfőbb értékét, magát az információt részesítjük előnyben. Egy adott szakmai kérdés megválaszolásának a leggyakoribb kiindulási pontja lehet számunkra az internet, amely a minket körülvevő globális társadalomnak egyik fontos eszköze [7]. Azonban az internetről nem tételezhető fel, hogy az minden feltett kérdésünkre kielégítő választ fog nyújtani. A világhálón történő információkereséskor egyre nagyobb gondot jelent számunkra a minőségi, releváns információk felkutatása és kiválogatása a ránk zúduló információáradatból. Ebben támogatnak minket a rendelkezésre álló keresőszolgáltatások, bár azok

sem minden esetben nyújtanak tökéletes megoldást.

Az interneten megjelenő keresőeszközök köré külön iparág szerveződött, amelybe kisebb-nagyobb méretű cégek, vállalatok nagy pénzösszegeket fektetnek be. Mindezt pedig saját versenyképességük, sikerességük és hatékonyságuk javítása érdekében teszik. A számadatok érzékeltetésére a *Search Engine Marketing Professionals Organization* (*SEMPO*) 2005-ös felmérésének lényeges megállapításaira utalnék: az Egyesült Államokban és Kanadában 5,75 milliárd dollárt költöttek 2005-ben keresőmarketingre (*SEM* = *Search Engine Marketing*). Ez az összeg 44%-kal haladta meg a 2004-es költségeket. Az előrejelzések szerint a keresőmarketingbe fektetett pénzösszeg 2010-re elérheti a 11 milliárd dollárt Észak-Amerikában [12]. A SEMPO 2009-es felmérése szerint a keresőmarketing-ipar Észak-Amerikában 16,6 milliárd dollárra növekszik 2010-re [13]. *Safa Rashtchy* internet média- és marketingelemző szerint a keresőpiac fizetős része 2005-ben megközelítőleg 10 milliárd dollár hasznot termelt globálisan, ami 41%-kal fog növekedni 2006-ban. Előrejelzése szerint a keresőpiacnak ez a része globálisan 37%-os éves növekedésre számíthat 2010-ig, ami több mint 33 milliárd dollárnak felel meg. A 2005-ös felmérésből kiderült, hogy a keresőmarketing-kampányok elsődleges célja a „branding" (egy márka ismertté tétele) és az értékesítés volt. A kisebb cégek inkább a termékek eladására helyezték a hangsúlyt, míg a nagyobbak (500 alkalmazott felettiek) a weboldalukra érkező forgalom növelésére [12]. Mindkét felmérés adatai rávilágítottak arra, hogy a keresőszolgáltatások, valamint a különféle cégek, vállalatok abban érdekeltek, hogy minél több bevételre tegyenek szert. Ebben a kiélezett versenyhelyzetben a keresőeszközök folyamatosan törekszenek arra, hogy megújuljanak és minél több speciális, új szolgáltatással vonzzák a használókat maguk köré. Ezért a versenyben részt vevő szereplők számára rendkívül fontos, hogy az egyes keresőeszközök minőségét hogyan értékelik a kutatók.

Az információkeresésre irányuló kutatás több mint két évtizedes múltra tekint vissza. Ezen a területen a vizsgálódás egyik lehetséges iránya a kérdést feltevő felhasználók viselkedésének tanulmányozása, azaz milyen kérdést, hogyan, és milyen társadalmi rétegből, milyen tanultságúak tettek fel. A kutatók 1981-től számos modellt alkottak meg. A modellek kialakítását nagymértékben befolyásolta a kutatók világlátása, kutatási területe és jártassága. Ennek függvényében beszélhetünk kognitív perspektivikus, szociális, szociális-kognitív vagy szervezeti modellekről [7]. Mindez azt tükrözi számunkra, hogy a felhasználók weben történő keresése több szinten vizsgálható, beleértve a társadalmi és a szervezeti szintet, az információkeresés szintjét, az ember és a számítógép közötti kapcsolat szintjét, valamint a megfogalmazott keresőkérdés szintjét [15]. Csak jelzésszerűen hivatkoznék néhány jeles kutatóra, akik ezen a téren komoly eredményeket értek el: *Spink, Jansen, Saracevic, Ingwersen*. Az információkeresési viselkedésekkel, modellekkel kapcsolatos kutatási eredmények megtalálhatók [7]-ben.

A vizsgált témakör fogalmai

Fontosnak tartom, hogy meghatározzam ennek a komplex tárgykörnek az alapvető fogalmait és a közöttük lévő kapcsolatrendszert. Először a „metaadat" fogalmának meghatározásával kezdeném, mert a hozzá tartozó információknak a megléte szükséges az internetes keresők működéséhez. Metaadat kifejezésen a weblapok intellektuálisan vagy automatikusan létrehozott másodlagos adatait értjük, amelyek magát a dokumentumot jellemzik [24]. A keresőrendszerek a saját adatbázisukat csupán olyan technikai metaadatokkal látják el, mint a begyűjtött dokumentum URL címe, fájlformátuma, mérete, begyűjtési dátuma stb. Egy másik meghatározás szerint metaadat alatt mindazokat a többletinformációkat értjük, amelyeket a weboldalak készítői a weboldalakhoz kapcsolnak a keresőkérdés pontosabb megválaszolása reményében [19]. Ezen adatok körébe tartoznak: a bibliográfiai leírás szabványosított adatelemei, a dokumentum tartalmát leíró kulcsszavak, tárgyszavak, deszkriptorok és az osztályozási jelzetek. Elengedhetetlen követelmény volt a metaadatok egységes elektronikus kezelése, ami kiterjedt ezeknek az adatoknak az elsődleges dokumentumokból való kinyerésére és a dokumentumok számítógépes leírására [24]. Számos metaadatrendszer jött létre a hálózati információk feldolgozására, például az *OCLC InterCat*, a *DublinCore*, a *WWW Semantic Header*, a *TEI* (*Text Encoding Initiative*) fejléc stb. Ezek közül a metaadatrendszerek közül a DublinCore jelentőségét hangsúlyoznám, mivel napjainkban ez az egyik legáltalánosabban elterjedt metaadat-alkalmazás. A Dublin Core formátum 15 leíró elemet tartalmaz, és ez áll a legközelebb a könyvtári katalogizáláshoz. Elterjedését elősegítette, hogy adatelemeit az európai szabványosítási szervezet, a *European Committee*

for Standardization (*CEN*) is elfogadta [8]. A Dublin Core-ra vonatkozó magyar nyelvű szabvány letölthető a *mek.oszk.hu/dc* oldalról.

A keresőszolgáltatásoknak két típusát különböztethetjük meg: az indexelőszolgáltatásokat és az internetkatalógusokat. Az előbbieken belül különleges változatként fordulnak elő a gyűjtő- és a metakeresők. A metakeresők (*meta search engines, Meta-Suchmaschinen, métamoteur, métachercheur*) segítségével több indexelőszolgáltatásban kereshetünk párhuzamosan anélkül, hogy az egyes szolgáltatásokkal külön foglalkoznunk kellene. A rendszer mindegyik keresőszolgáltatás adatbázisában végrehajtja a keresést, megjelenítve a találatoknál, hogy melyik szolgáltatás adatbázisában találta meg a rekordot, valamint a duplumszűrésre is törekszik. A metakeresők előnye, hogy rövid idő alatt valószínűsíthetően több releváns találathoz juthatunk [23]. Továbbá, nehezebben csapják be őket azok az oldalak, amelyek mindenféle trükkös megoldásokkal a javukra befolyásolják a keresők találatrangsorolását, azonban ezeknek az oldalaknak nincs igazi, használható tartalmuk; „spam"-eknek hívjuk őket. A metakeresők azért képesek a „spam"-oldalak kiszűrésére, mert azok általában egy-egy keresőre szakosodnak és egyszerre több keresőt már nem tudnak becsapni [19]. A „spamdexing" kifejezés a „spamming" és az „indexing" szavak összeolvadásából született, amely a '90-es évek közepén jelent meg a keresőiparban. A *search spam*, *search engine spam*, illetve a *web spam* kifejezéseket szintén használjuk rá. Ez a folyamat számos módszert foglal magába, amelyeket azért alkalmaznak, hogy a kereső által indexelt oldalak relevanciáját vagy fontosságát növeljék. Használt módszerei azonban nincsenek összhangban a kereső indexelésének célkitűzésével. Néhányan úgy vélekednek, hogy a spamdexing a keresőoptimalizálás részét képezi. Több kereső ellenőrzi a spamdexing előfordulásait és eltávolítja a gyanús oldalakat indexéből [14]. Átmeneti típusnak tekinthető a gyűjtőszolgáltatás (*configurable unified search interface [CUSI], all-in-one formular, sample service, Sammeldienst*), amely több keresőszolgáltatást ajánl fel, de mindig csak egyet választhatunk ki a lekérdezésre [23].

Az indexelőszolgáltatások („keresőgépek"-nek is hívjuk őket), (*search engines, Suchmaschinen, moteur de recherche*) emberi munka nélkül, számítógépes programok segítségével végzik a keresést a hálózaton. Ezek a szolgáltatások két fő részből állnak: a keresőrobotból (*crawler, web spider, web robot, bot*) és az indexelőből (*indexer*). A robotok állandóan figyelemmel követik és begyűjtik a weboldalakat a világhálóról a keresőszolgáltatás adatbázisába. A webhelytulajdonosok adhatnak utasításokat a robotoknak begyűjtéskor, ekkor egy *robots.txt* állományt kell elhelyezniük a webhely gyökérkönyvtárában. A robotok úgy vannak kialakítva, hogy követniük kell az utasításokat, ezért megpróbálják megtalálni a robots.txt állományt és elolvasni az utasításokat belőle, mielőtt a webhelyről bármilyen állományt begyűjtenének. Ha ez az állomány nem található meg, akkor feltételezik, hogy a webtulajdonos nem kíván speciális utasításokat meghatározni számukra. A robots.txt állomány valójában egy olyan kérés a webhelyen, amely megszabja, hogy egyes robotok bizonyos állományokat vagy könyvtárakat figyelmen kívül hagyjanak begyűjtéskor. Ha a webhely több aldoménből áll, akkor azok mindegyikének rendelkeznie kell a saját robots.txt-jével [9]. Az indexelő elemzi a begyűjtött dokumentumokat, amelyekből előállítja az indexkifejezéseket. Létrehoz egy indexet, amely minden szóhoz – a stopword-öket kivéve – hozzárendeli az őt tartalmazó *Uniform Resource Locator*-ok (*URL*) listáját. A keresőszolgáltatás erre az indexre támaszkodik, amely révén elvégzi a keresést a felhasználó számára [19]. A keresőrobotot és az indexelőt integráló egységet „keresőgépnek", „keresőmotornak", „keresőműnek" (*search engine*), „keresőrendszernek" (*search system*) nevezik. Hibásan a teljes keresőszolgáltatást is „keresőgépnek", „keresőmotornak", „robotnak" hívják, ami a szolgáltató rendszernek csak az egyik részét jelenti. Ebbe beletartozik még a felhasználói felület és a szolgáltatott tartalom is [23]. Ezek a keresőszolgáltatások általában rendelkeznek egy egyszerű és egy összetett keresési lehetőséggel. Egyszerű kereséskor (*quick search*) rendkívül nagy lehet a visszakeresett, nem releváns dokumentumok száma, azaz a zaj. Ennek csökkentése érdekében tanácsos használnunk a részletes keresési lehetőséget (*advanced search*, *powered search*) [26].

Amikor egy vagy több releváns kulcsszót írunk be a keresőablakba, a kereső indexében megvizsgálja, hogy melyek a kérdésünkre legjobban illeszkedő találatok és azokat szolgáltatja számunkra. A találatlistában szereplő oldalakról általában egy rövid ismertetést kapunk, amely magába foglalja a forrás címét, valamint annak kiemelt szövegrészeit. Találati halmazaink mennyiségi viszonyait (a halmazok egymáshoz viszonyított terjedelmét, illetve helyzetét) logikai műveletekkel adhatjuk meg. Ezeket a műveleteket pedig logikai műveleti

jelekkel – ún. operátorokkal – fejezhetjük ki. A legtöbb kereső támogatja az ÉS, VAGY, NEM Boole-operátorok használatát, amelyekkel a keresés tovább finomítható. A keresők egy része megengedi a helyzeti operátorok (*proximity operators*) használatát is, amelyek lehetővé teszik számunkra, hogy meghatározzuk a kulcsszavak közötti távolságot (pl. NEAR, BETWEEN, WITH operátorok stb.). Kereséskor a találati halmaz terjedelmét úgy módosíthatjuk, hogy megengedjük, hogy a keresőszó elején, végén vagy meghatározott karakterpozícióin bármilyen karakter helyezkedjen el. Ehhez „jolly joker*"* jeleket (*wild card*) adhatunk meg a keresőszóban. Bővebb találati halmazokat nyerhetünk abban az esetben, ha a keresőszó elején („balról csonkolás") és/vagy végén („jobbról csonkolás") meghatározott karaktert használunk, amely minden megelőző és/vagy követő karaktert helyettesít. Ezt a műveletet csonkolásnak (*truncation*) nevezzük. A csonkoló jelek használata keresőrendszerenként eltérő [26]. A keresőknél létezik egy kifinomult keresési technika, a fogalomalapú keresés (*concept-based searching*). Ennél a technikánál statisztikai elemzéssel találjuk meg azokat az oldalakat is, amelyek nem tartalmazzák az általunk megadott kulcsszavakat. Ekkor azonban az oldalak olyan egyéb szavakat (pl. szinonimákat, tulajdonneveket, állandósult szókapcsolatokat) foglalnak magukba, amelyek ugyanabba a fogalomkörbe tartoznak, mint a beírt keresőszavak. Így a keresőrendszer akkor is relevánsnak minősíti az oldalakat, ha a megadott keresőszavak nem találhatók meg bennük. Egy másik kereső funkció a fuzzy megfeleltetés/illesztés (*fuzzy matching*), amelynek az a lényege, hogy a keresőszót a szótőre redukálják és minden lehetséges szóalakot ráillesztenek különböző algoritmusokkal. Ez nagymértékben megnöveli a találati halmazunkat, mert minden kapcsolódó szót visszakeres, még a kevésbé relevánsakat is. Néhány keresőnél alapértelmezett funkció a *stemming*, ami a keresőkérdés összes toldalékolt alakjának a visszakeresésére alkalmas. Ha ezt a funkciót használjuk a keresőkérdésre, akkor még bővebb találati halmazt kapunk a csonkoláshoz képest.

Megállapítható, hogy egy kereső hasznossága valójában a szolgáltatott találatlistája relevanciájától függ. A legtöbb kereső rangsorolja a találatokat fontosságuk szerint arra törekedve, hogy a legjobb oldalakat jelenítse meg a találatlista elején. Keresőnként változó, hogy milyen rangsorolási módszert alkalmaznak erre a feladatra. A *Google Page Rank* algoritmusa az egyik legismertebb rangsorolási módszer, amely az oldalak közötti linkstruktúrát veszi alapul és más egyéb tényezőket egyaránt figyelembe véve súlyozza a találatokat. Beszélhetünk olyan keresőkről is, amelyek nem egy egyszerű találatlistában jelenítik meg a találatokat, helyette inkább a keresőkérdéshez kapcsolódó tematikus kategóriákba rendezik azokat. Ezek a csoportok (klaszterek) abban segítenek bennünket, hogy könnyen áttekinthessük a keresett témát, és hogy kiválaszthassuk a megfelelő kategóriát. A találatok klaszterálása segítséget nyújt a keresés finomításában a korábbi keresés találati halmazára támaszkodva (pl. *clusty.com* kereső) [4]. Találkozunk olyan vizuális keresőeszközökkel is, amelyek a találatokat grafikusan jelenítik meg (*graphical visualization*) két- vagy háromdimenziós képekben (pl. *viewzi.com*, *eyeplorer.com*). Az internetkatalógusokat (*directories, annuaires internet, répertoires internet*) [26] „böngészőszolgáltatásnak" *(browsing service*, *browsing Dienste*) [24], „tárgyszótárnak", „tématárnak" (*subject directory, Themenverzeichniss, annuaire thématique*) [21], valamint „webes katalógusnak" (*annuaire Web, répertoire Web*) is nevezik [19]. Továbbá a „linkgyűjtemény" és a „tematikus katalógus" megnevezések is ismertek. Ezek a katalógusok hierarchikus osztályozási rendszert használnak. Adatbázisaik többnyire intellektuálisan feldolgozott weboldalak rekordjait foglalják magukba, valamint kapcsolatokat más adatbázisokhoz. Az osztályozást és a tartalmi kivonatok készítését szerkesztőségben végzik. Azonban sok linkgyűjtemény egyéni vagy közösségi munka eredménye és nincs mögötte szerkesztőség, lásd például a „Startlap" tematikus oldalait. Ezekben a katalógusokban osztályok alapján böngészhetünk, de lehetőségünk van arra is, hogy egy keresőkérdés megadásával, célzott kereséssel találjuk meg a kívánt osztályt. Vannak olyan katalógusok is, amelyek indexelőszolgáltatásként is működnek, ilyen például az *ok.hu/linktar*. Az internetkatalógusok adatbázisai sokkal kisebbek, mint az indexelőszolgáltatásokéi, azonban a keresés kevesebb zajt eredményez az intellektuális feldolgozásnak és a gondos osztályozásnak köszönhetően. A szakterületre specializálódott keresők nagy része internetkatalógusnak tekinthető. Egy részüket híres kutatóintézetek gondozzák, más részük kereskedelmi szolgáltatásnak minősül [26].

Kapcsolódó kutatási területek

Elsősorban a szemantikus webnek, mint perspektivikusan fejlődő területnek a jelentőségét hangsúlyozom, melynek feladata a jelentés megtalálása a webes tartalmakban. A szemantikus web kialakítá-

sára irányuló törekvések nyomán jelentek meg az ún. ontológiák. *Gruber* megfogalmazása szerint az „ontológia megegyezésen alapuló fogalmi rendszer formális, egyértelmű leírása” [3]. Ebben a meghatározásban a „megegyezésen alapuló” kitétel lényeges, hiszen azt a szemléletet tükrözi, hogy az ontológiák szemantikai szabályrendszerek, amelyek a dolgok rendezésére használhatók [25]. Az ontológiák lehetővé teszik, hogy tisztázzuk az alapvető fogalmakat és a közöttük lévő relációkat. Továbbá elősegítik, hogy az erre vonatkozó tudásunkat formálisan és gépi következtetésre alkalmasan fogalmazzuk meg [18].

Számos fejleménynek kellett ahhoz bekövetkeznie, hogy webes ontológiák jöhessenek létre. Ezek közül csak a legfontosabbakat emelném ki. 2000-ben közreadtak egy „tématérképnek” (*topic map*) nevezett hierarchikus fogalmi struktúrát kezelő szabványt. A weben jelenleg elérhető vizualizált fogalmi struktúrák többsége ezen vagy ehhez hasonló fejlesztéseken alapszik [27]. A W3C konzorcium irányítása alatt egy másik irányban kezdődött el a fejlesztés. Ennek egyik fontos eredménye, hogy 2000-ben a web metaadatainak leírására egy szabványt hoztak létre, az XML-en alapuló webforrás leíró nyelvet (*Resource Description Framework = RDF*). A weben található hierarchikus fogalmi struktúrák formális leírására is ezt a nyelvet használták fel. 2002-ben a W3C konzorcium kezdeményezésére hozzákezdtek az ontológiák szabványának tekinthető webontológia-nyelv (*Ontology Web Language = OWL*) kidolgozásához [25]. Az OWL 2-re vonatkozó szabványajánlást 2009-ben adta közre a W3C konzorcium [5]. Jelenleg elérhető és már létező általános ontológiáknak tekinthetők például a *Dublin Core*, a *Magyar Egységes Ontológia*. Szakterületi ontológiaként megemlíthető a *Galen*, amely orvostudományi szakterületen használatos [18]. A „Museo24” projektben kifejlesztett ontológiának érdekes felhasználási területe a virtuális múzeum, amely gondolatvilágában közel áll a könyvtárakéhoz. (Lásd a projekt leírását [17]-ben.) Jelenleg egyfajta közeledés figyelhető meg hazánkban a könyvtári és az informatikai szakmai közösségek között az ontológiák terén, amit a W3C konzorcium magyar irodája szakmai előadások szervezésével egyaránt támogat [18, 22].

A szakirodalomban az „invisible web” (láthatatlan web), „hidden web” (rejtett web), vagy „deep web” (mély web) angol kifejezéseket használják mindazon dokumentumok és adatok körének az összefoglalására, amelyek számos oknál fogva nem érhetők el a keresőszolgáltatások számára. A láthatatlan web csoportjába sorolhatók: a dinamikus weblapok (azaz pl. a kereshető adatbázisokból nyert oldalak), azok az oldalak, amelyek csak regisztráció után érhetők el, a nem szöveges dokumentumok, valamint a keresőrobotok elől elzárt oldalak. Fontos hangsúlyoznunk, hogy a web csak egy szolgáltatás az interneten, tehát az nem azonos vele. Egy olyan hipertext-struktúrára épül, amelyben szabadon böngészhetünk a szöveges formában megjelenített információk közötti kapcsolatok (linkek) alapján. Ha egy weblapra nem mutat egyetlen link sem, akkor az nem kerül bele a kereső adatbázisába. Azoknak a weboldalaknak az összességét, amelyeket a keresők keresőmotorjai megtalálnak „felszíni webnek” (*surface web*) vagy „statikus webnek” nevezzük. Ennek nagysága a teljes web méretének a 0,18%-ára becsülhető. Ezzel szemben a láthatatlan web információmennyisége 550-szer nagyobb, mint a felszínié és növekedése, gyarapodása is sokkal gyorsabb ütemű [10, 7]. Sokféle törekvéssel igyekeztek a rejtett webet „láthatóvá tenni”, például bizonyos metakeresőkkel, intelligens keresőprogramokkal (ágensek), témakatalógusok kialakításával, egyéb speciális keresőkkel. Mindezeket a lehetséges megoldási kísérleteket, eszközöket bővebben kifejtve lásd [7]-ben.

Ehhez a tárgykörhöz kapcsolódóan hivatkoznék a szövegbányászat és az adatbányászat ígéretes lehetőségeire, amelyek a rejtett tudás kinyerésére törekednek a weben található, nagy mennyiségű strukturálatlan vagy félig strukturált HTML és egyéb formátumú dokumentumokból. Fiatal kutatási területnek számít a „web mining”, amely kiterjed az adatbányászatra, az internettechnológiákra, valamint a szemantikus webre [11].

A weben találkozhatunk speciális keresőszolgáltatásokkal is, például képek, videoanyagok visszakeresésére alkalmas keresőkkel, amelyek nagy népszerűségnek örvendenek a használók körében. Megjelenésük azt jelzi, hogy a használók rendkívül nagymértékben igénylik a nem szöveges dokumentumok eredményes megtalálását is. Ezen az új kutatási területen a megfelelő információkereső nyelvek létrehozása és azok további fejlesztése elengedhetetlenül fontos feladat amellett, hogy a tartalomgazdák metaadatokat helyeznek el a kép- és videofájlokba, továbbá, hogy egyre fejlettebb kép- és beszédfelismerő eszközöket használnak a keresőgépek.

Egy másik, szerkezeti sajátosságaiból adódóan elkülönülő dokumentumcsoportot is meg kell említenünk a weben: a blogokat és a mikroblogokat (pl. Twitter). Számuk rohamosan növekszik, mert divatos véleménynyilvánítási forma a használók körében. Komoly kihívást jelent napjainkban a blogok hatékony visszakeresésének megoldása, amely a hagyományos információkereső rendszerek módszereitől eltérő mechanizmusokat követel meg. Ez abból is adódik, hogy a blogoknál rendkívül nagy szerepe van az aktualitásnak és a kapcsolódó linkeknek, azonban sokszor nehéz meghatároznunk a témájukat. Megjegyzem, hogy az időtényező fokozottabb kiaknázása a keresésekben új vonásnak számít, például a blogok és a hírek esetében [4, 16]. 2006-ban a korábbi passzív internetezők önszerveződő közösségek aktív tagjaivá váltak, ami főként a web 2.0 térhódításának volt köszönhető. Ezáltal a web rendkívül nyitott közösségi színtérré alakult át az innováció, a kibontakozás, valamint az értékteremtés számára. A Google és más üzleti cégek fokozatosan teret engednek a tömeges együttműködés kultúrájának, még pedig olyan formában, hogy nyíltan hozzáférhetővé teszik alkalmazásaik programozófelületét (az API-t) platformjaikon. Az API-k megnyitása után a fejlesztők (akik közül néhányan korábban „hekkerek" voltak…) gyors tempóban kezdték el gyártani az új alkalmazásokat [20]. A Google számos szolgáltatásával támogatja az egyedi felhasználók kényelmét kereséskor. Ezek például a következők: *Google Suggest, Custom Search, Google Alert, Desktop Search, Google Toolbar*. Kifejezetten a közösségi tartalmak visszakeresésére alkalmas keresőként megjegyezhető például a *grub.org*.

Irodalom

[1] Creating a European library space: Telematics for libraries programmes 1990–1998. = *http://cordis.europa.eu/libraries/en/intro.html* (2008.01.22.)

[2] FODOR, I.: Merre megy a világ gazdasága, merre mehetünk mi? = Az információs társadalom. Összeáll. Demetrovics J., Keviczky L. Budapest, MTA, 2000. p. 95–113.

[3] GRUBER, T.: A translation approach to portable ontology specifications. = Knowledge Acquisition, 5. köt. 2. sz. 1993. p. 199–220.

[4] LANGVILLE, A. N. – MEYER, C. D.: Google's PageRank and beyond. The science of search engine rankings. Princeton-Oxford, Princeton University Press, 2006.

[5] OWL 2 Web Ontology Language Document Review. (2009). = *http://www.w3.org/TR/2009/PR-owl2-overview-20090922/* (2010.04.09.)

[6] NYÍRI, K.: Globális társadalom, helyi kultúra. = Az információs társadalom. Összeáll. Demetrovics J., Keviczky L. Budapest, MTA, 2000. p. 43–64.

[7] PAJOR, E.: A láthatatlan/mély web felhasználása a könyvtári tájékoztatásban. [Doktori (PhD) értekezés]. Budapest, ELTE BTK, 2006. 214 p.

[8] RÁCZ Ágnes.: A kiadványok bibliográfiai számbavétele; leíró katalogizálás. = Könyvtárosok kézikönyve. 2. köt. Feltárás és visszakeresés. Szerk. Horváth Tibor – Papp István. Budapest, 2003, Osiris. p. 187–295.

[9] Robots Exclusion Standard (Robot Kizárási Szabvány) szócikk. = *http://en.wikipedia.org/wiki/Robots_Exclusion_Standard* (2010.04.10.)

[10] RUTKOVSZKY, E. – RUTKOVSZKY, Á: A láthatatlan web keresése. (2003). [Előadásanyag]. = *https://nws.niif.hu/ncd2003/docs/ehu/EHU-61.htm* (2008.01.15.)

[11] SCIME, A.: Web mining. Applications and techniques. 2005, Idea Group Inc.

[12] SHERMAN, C.: The state of search engine marketing. (2006). = *http://searchenginewatch.com/showPage.html?page=3575926* (2008.01.20.)

[13] SHERMAN, C.: The State Of Search Engine Marketing 2010. (2010). (A cím félrevezető lehet, mert a legfrissebb felmérés 2009-es, amit a SEMPO elvégzett. A SEMPO honlapján 2010. április közepén a 2009-es év felmérése érhető el csak tagoknak vö. *http://www.sempo.org/learning_center/research/*) *http://searchengineland.com/the-state-of-search-engine-marketing-2010-38826* (2010.04.10.)

[14] Spamdexing szócikk. = *http://en.wikipedia.org/wiki/Spamdexing* (2010.04.10.)

[15] SPINK, A. – JANSEN, B. J.: A study of web search trends. = Webology, 1. köt. 2. sz. 2004. *http://www.webology.ir/2004/v1n2/a4.html* (2008.01.27.)

[16] SULLIVAN, D.: What is real time search? Definitions & players. (2009). = *http://searchengineland.com/what-is-real-time-search-definitions-players-22172* (2010. 04.10.)

[17] SZÁSZ, B. – SARANIVA, A. – BOGNÁR, K. – UNZEITIG, M. – KARJALAINEN, M.: Cultural heritage on the semantic web – the Museum24 project. (2006). [Előadásanyag]. 10 p. *http://www.seco.tkk.fi/events/2006/2006-05-04-websemantique/presentations/articles/Szasz-museum24Paris.pdf* (2008.01.14.) *http://www.museo24.fi* („Museo24" portál honlapja) (2008.01.14.)

[18] SZEREDI, P.: Ontológiák – egy matematikus-informatikus szemével. = Ontosz. Előadássorozat a formális ontológiákról. Az ontológia fogalmának, felépítésének, alkalmazási lehetőségeinek különböző megközelítései. Budapest, W3C, 2007. ápr. 25. *http://www.w3c.hu/rendezvenyek/2007/ontologia/index.html* (2008.01.10.)

[19] SZEREDI P. [et al.]: A szemantikus világháló. = A szemantikus világháló elmélete és gyakorlata. Szerz. Szeredi P. [et al.]: Budapest, 2005, Typotex. p. 17–59.

[20] TAPSCOTT, D. – WILLIAMS, A. D.: Wikinómia. Hogyan változtat meg mindent a tömeges együttműködés. Szerk. Török Hilda; ford. Garamvölgyi Andrea. Budapest, HVG, 2007.

[21] UNGVÁRY Rudolf: Az információkeresés értékelése. = Osztályozás és információkeresés: kommentált szöveggyűjtemény. 2. köt. Az információkeresés és elmélete. Szerk. Ungváry Rudolf, Orbán Éva. Budapest, OSZK, 2001. *http://mek.oszk.hu/01600/01683/pdf/01683-2.pdf* (2007.11.17.)

[22] UNGVÁRY, Rudolf: Az ontológia fogalma, avagy az eltűnt tezaurusz. = Ontosz. Előadássorozat a formális ontológiákról. Az ontológia fogalmának, felépítésének, alkalmazási lehetőségeinek különböző megközelítései. Budapest, W3C, 2007. ápr. 25. *http://www.w3c.hu/rendezvenyek/2007/ontologia/index.html* (2008.01.10.)

[23] UNGVÁRY Rudolf: A tartalom szerinti információkeresés az interneten: I. indexelőszolgáltatások. = TMT, 47. köt. 1. sz. 2000. p. 3–17. *http://tmt.omikk.bme.hu/show_news.html?id=1624&issue_id=15* (2008.01.27.)

[24] UNGVÁRY Rudolf: A tartalom szerinti információkeresés az interneten: II. internetkatalógusok. = TMT, 47. köt. 2. sz. 2000. p. 55–67. *http://tmt.omikk.bme.hu/show_news.html?id=1625&issue_id=16* (2008.01.27.)

[25] UNGVÁRY Rudolf: Tezaurusz és ontológia, avagy a fogalmi ismertetőjegyek generikus öröklődésének formalizálása. = TMT, 51. köt. 5. sz. 2004. p. 175–191. *http://tmt.omikk.bme.hu/show_news.html?id=3615&issue_id=450* (2008.01.27.)

[26] UNGVÁRY Rudolf – VAJDA Erik: Könyvtári információkeresés. 2. jav. kiad. Budapest, Typotex, 2002.

[27] XML-Topic-Map (XTM) Standard, ISO/IEC 13250:2000. XTM TopicMaps Org. = *http://www.topicmaps.org/xtm* (2008.01.14.)

Beérkezett: 2010. V. 25.-én.

Tóth Erzsébet
a Nyíregyházi Főiskola Matematika és Informatika Intézetének docense.
E-mail: *tothe@nyf.hu*

Már 400 millióan használják a Firefoxot

A kontinensünkön egyetlen esztendő alatt 8,4 százalékkal csökkent az *Internet Explorer* (*IE*) piaci részesedése és jelenleg alig haladja meg az 50 százalékot. A legnagyobb konkurensének számító *Firefox* viszont köszöni szépen, jól van. Ugyanakkor *Tristan Nitot*, a *Mozilla Europe* elnöke tisztában van azzal, hogy nem ülhetnek a babérjaikon. Gőzerővel fejlődik ugyanis a *Google Chrome*, az IE 9-es verziója is sok tekintetben előrelépést jelent a korábbi változatokhoz képest, így a Firefoxnak is fejlődnie kell, ha lépést akar tartani a konkurenseivel.

„A Google rendkívül innovatív vállalat, amely tavaly felerősítette a böngészőpiacon zajló versenyt. Évek óta vágytunk erre a konkurenciaharcra, mert úgy véljük: sokat lendíthet a böngészők fejlesztésén. Ugyan a Mozillára nehezedő nyomás nem lett kisebb, de ez így van jól." – jelentette ki Tristan Nitot. Az elmúlt hónapokban számos kritika érte a Mozillát, hogy túlzottan elkényelmesedett és nem figyelt eléggé a Firefox fejlesztésére. Ez különösen annak tükrében érthető, hogy a Firefoxot több mint 400 millióan használják világszerte.

„Néhány országban, mint Lengyelország vagy Németország már közel 50 százalékos piaci részesedésre tettünk szert, ami szenzációs teljesítmény. Minél nagyobb azonban a tortából általunk kihasított szelet, annál nehezebb tovább növekednünk. A hiányzó innovációs képességünkre vonatkozó kritikákat viszont visszautasítom, hiszen a kiadási stratégiánkat az igényekhez igazítottuk és gyorsítottunk a fejlesztéseken is. Ráadásul a Firefox 4-essel egy olyan verzióugrás következhet be, amely pont a sebesség és a dizájn tekintetében hoz magával jelentős javulást." – hangsúlyozta a szakember.

Nitot hozzátette, hogy a jövőben szeretnék az online közösség kreatív potenciálját még jobban hasznosítani, ezért a céljuk az, hogy a Mozilla Labs kísérleti ötleteit közvetlenül integrálni lehessen a fejlesztési folyamatba. Emellett jelentősen leegyszerűsítenék a kiegészítők fejlesztési lehetőségeit is és a modulok a böngésző újraindítása nélkül, azonnal használhatók lesznek.

/SG.hu Hírlevél, 2010. július 26., *http://www.sg.hu/*

(SzP)