Tóth Erzsébet

Az internetes keresők működésének technikai háttere

Jelen írás azok számára lehet érdekes, akik mélyebben érdeklődnek egy komplex kereső működésének technikai háttere iránt. Összefoglalja számunkra egy keresőrendszer működési elveit és az alkotóelemeinek feladatait. A találatok rangsorolása a keresők egyik legmeghatározóbb sajátossága, ehhez kapcsolódóan bemutatja a Google kereső PageRank eljárását. Tárgyalja a PageRank algoritmus alapgondolatát és rávilágít arra, hogy hogyan modellezi a felhasználói viselkedést egy keresés során.

A világháló új helyzetet teremtett a hagyományos információkeresés területén, hiszen a rendszerezettség, a homogenitás és a rend helyett azt láthatjuk, hogy bárki létrehozhat rajta tartalmat, amelynek minőségét és megbízhatóságát senki nem vizsgálja. Tehát a világháló heterogén szintaktikájú és szemantikájú, továbbá többnyire nem ellenőrzött tartalmú dokumentumok halmazát reprezentálja. Ebből adódóan az internetes keresés alapvetően eltér egy lassan változó, kontrollált dokumentumgyűjteményben való kereséstől. Ez a különbség többek között abban is megnyilvánul, hogy a keresőknek meg kell találniuk a releváns webes tartalmaknak azokat a halmazait, amelyek jól hasznosíthatók a felhasználók számára, nem pedig egy hagyományos gyűjteményből kell kiválogatniuk a keresőkérdésre pontosan illeszkedő dokumentumokat. Kereséskor a legjobb találatoknak egyéb jellemzőik is vannak (frissítési gyakoriság, minőség, hivatkozások száma, népszerűség stb.), amit a keresőknek szintén figyelembe kell venniük és nem elegendő csupán a keresésnek pontosan megfelelő dokumentumokat szolgáltatniuk. Egy-egy keresésre különböző válaszok adhatók, ezért nagyon lényeges, hogy mely találatok jelennek meg elsőként a felhasználóknak [10].

Még mielőtt rátérünk a téma részletes tárgyalására, pontosan meghatározzuk az internetes keresők fogalmát. Internetes keresők alatt a programoknak egy olyan általános csoportját értjük, amely lehetővé teszi a weben történő információkeresést a felhasználók számára. Ezek a programok dokumentumokat indexelnek és arra törekednek, hogy megtalálják a feltett keresőkérdésre a releváns találatokat.

Keresőszolgáltatások előretörése az interneten

Lényeges változásnak tekinthetjük, hogy az ezredfordulóra az országok döntő többsége információs társadalomként jelent meg a világtérképen. Végbement egy technológiai forradalom, a digitális eszközök mindennapjaink részévé váltak, azokat nemcsak eszközöknek tekintjük, hanem érzelmileg is kötődünk hozzájuk. Mindez kihatott a médiafogyasztásra, megváltoztak a kulturális objektumok átvételi csatornái. Átrajzolódott a gazdaság, a kormányzás, a tartalomipar, átalakult a fogyasztó és az előállító viszonya. Legfontosabb jelenségként tapasztalhattuk az informatika, az internet és a számítógép hétköznapivá válását, bár ezek a folyamatok jóval korábban (10-20-30 éve) kezdődtek el. Az internet egy szűk kör számára hozzáférhető újdonságból a világ lakosságának nagyjából hatoda által használt eszközzé vált. 1998–2008 között az internetezők száma több mint tízszeresére növekedett világszerte [11].

Az 1990-es évek elején néhány webszerverről lista készült, amelyet *Tim Berners-Lee* állított össze és a *CERN* (*Conseil Européenne pour la Recherche Nucléaire***)** szerverén helyezett el. Miután egyre több webszerver jött létre, ez a központi lista már nem volt elégséges. Később az *NCSA* (*National Center for Supercomputing Applications*) webhelyén jelentették be az új webszervereket „What's new!" megnevezés alatt.

Az első keresőeszköz az *Archie* volt, amelyet 1990-ben *Alan Emtage* hozott létre. A program az anonim FTP szervereken lévő állományok könyv-

tár struktúráját töltötte le, azonban nem indexelte ezeknek a szervereknek a tartalmát. Ezáltal megszületett az állománynevek első kereshető adatbázisa.

1991-ben a *Gopher* megjelenése két új keresőprogramhoz vezetett: a *Veronica*-hoz és a *Jughead*-hez. Mindkét program az Archie-hoz hasonlóan állománynevekre és címekre keresett a Gopher indexekben. A Veronica program (*V*ery *E*asy *R*odent-*O*riented *N*et-wide *I*ndex to *C*omputerized *A*rchives) megoldotta a menücímekre történő kulcsszavas keresést a Gopher világban. A Jughead (*J*onzy's *U*niversal *G*opher *H*ierarchy *E*xcavation *A*nd *D*isplay) keresővel adott Gopher szerverekről szóló menüvel információkat lehetett visszakeresni, amelyekre feltelepítették.

A web első kezdetleges keresője a W3Catalogus volt, amit 1993. szeptember 2-án indítottak útjára. 1993 nyarán *Matthew Gray* létrehozta az első keresőrobot-programot, amely *Perl* programozási nyelvre épült. Ezzel a keresőrobottal a „Wandex" nevű indexet állította elő. Keresőrobotját a web akkori méretének meghatározására használták 1995-ig. A web második keresője az *Aliweb* volt, amely 1993 novemberében jelent meg. Ez a keresőgép nem használt robotot az oldalak begyűjtésére, hanem kizárólag a webhely adminisztrátorok visszajelzésére épült azzal kapcsolatban, hogy létezik-e valamilyen indexállomány az adott webhelyen.

1993-ban kezdte el működését a *JumpStation*, amely már keresőrobotot alkalmazott az oldalak meglátogatására és az indexének létrehozására. Ez volt az első olyan kereső, amely egy keresőgép mindhárom alapvető jellemzőjét tartalmazta (a begyűjtést, az indexelést és a keresést). A szűkös erőforrások miatt a JumpStation indexelése és keresése a begyűjtött weboldalak címére korlátozódott.

Az első olyan kereső, amelynek a keresőrobotja már a begyűjtött weboldalak teljes szövegét vette figyelembe a *WebCrawler* volt (1994-ben jelent meg). Elődjeitől eltérően megengedte használóinak, hogy a weboldalakon lévő bármelyik szóra keressenek, ami ettől fogva alapvető elvárásként fogalmazódott meg a keresőknél. Szintén 1994-ben indult a *Lycos* a *Carnegie Mellon Egyetem*en, ami jelentős kereskedelmi sikerré vált. Hamarosan számos kereső jelent meg a piacon, amely egyre nagyobb lett. Ezek közé tartoztak a következők: *Magellan*, *Excite*, *Infoseek*, *Inktomi*, *Northern Light*, *AltaVista* és a *Yahoo!* Azonban a Yahoo keresőfunkciói a webes katalógusra épültek, nem pedig az oldalak teljes szövegére. Használói böngészhettek benne a kulcsszavas keresések alkalmazása helyett [12]. Ezeket a keresőszolgáltatásokat főként vállalati tőkéből, reklámokból, illetve kutatási költségvetésekből finanszírozták.

1996-ra már a különböző folyóiratok, üzleti és napilapok is komoly figyelmet szenteltek a keresőknek. Megnövekedett a keresésre specializálódó szoftvertermékek száma, például webes katalógusok, metakeresők, szakterületi szolgáltatások, kereső ágensek és „push" szolgáltatások jelentek meg [8]. Ugyanebben az évben a *Netscape* öt nagyobb keresővel kötött megállapodást, mely szerint azok évenként felváltva kerültek fel a keresőoldalára, meghatározott pénzösszeg fejében. Az öt kiválasztott kereső közé tartozott: a Yahoo!, a Magellan, a Lycos, az Infoseek és az Excite.

A '90-es évek végén a keresők fejlesztésébe jelentős pénzösszegeket fektettek be. 1997-ben kezdett el növekedni a „dot-com" névre keresztelt gazdasági buborék. Az e-szektor részvényeinek árai gyorsan emelkedtek, a külső tőke is meghatározóvá vált. Számos cég bukkant fel a piacon, néhány közülük felhagyott a nyilvános kereső működtetésével, helyette pedig vállatoknak szánt kiadást vitt a piacra (l. pl. *Northern Light*). Sok kereső belekerült ebbe a gazdasági buborékba, ami egy spekuláció-vezérelt piaci robbanásnak volt tekinthető. Ez a folyamat a tetőpontját 1999-ben érte el, és 2001-ben ért véget.

A *Google* 2000 környékén jelent meg a keresőpiacon és fokozatosan prominens keresővé vált. A cég felemelkedését annak köszönheti, hogy *PageRank* algoritmusával pontos találatokat szolgáltatott a használók kulcsszavas kereséseire. Ezenkívül a Google egyszerű keresőfelületet kínált. 2000-re a Yahoo! olyan szolgáltatásokat nyújtott, amelyek az *Inktomi* keresőre épültek. 2002-ben felvásárolta az Inktomit, valamint 2003-ban az *Overture*-t. 2004-ig áttért a Google kereső használatára, ekkor azonban útjára indította a saját keresőjét, amely a felvásárolt cégeinek a technológiáin alapult.

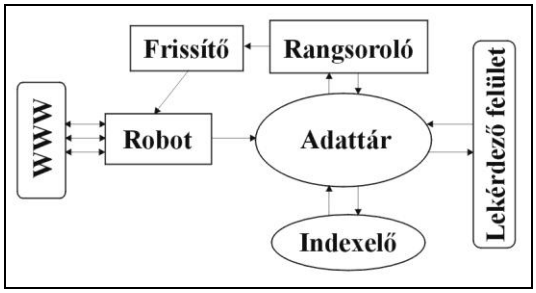
1998 őszén a *Microsoft* elindította az *MSN* keresőjét, amely az Inktomi-ból származó keresési találatokra támaszkodott. 1999 elején az MSN a *Looksmart* és az Inktomi keresési eredményeit jelenítette meg a találatlistájában, kivéve ugyanebben az évben azt a rövid időszakot, amikor az

Altavista találataira támaszkodott. 2004-ben a Microsoft elkezdett áttérni a saját keresőtechnológiájára, amelyet a keresőrobotjával, az ún. *msnbot*-tal támogatott. 2009 júniusában a Microsoft egy új, *Bing* nevű keresővel jelent meg a piacon. 2009 júliusában a Yahoo! és a Microsoft olyan egyezséget kötött egymással, amely szerint a Yahoo! is a Microsoft Bing keresőtechnológiájára épül.

2009 júliusában a keresők világméretű népszerűsége a következőket mutatta: Google (78,4%), *Baidu* (8,7%) és a Bing (3,17%). A Yahoo! 7,16%-os és az *AOL* 0,6%-os piaci részesedése szintén csökkent az előző évhez képest. 2009 májusában a Google használatának aránya 63,2% volt az Egyesült Államokban. 2009 júliusában a Baidunak 61,6%-os népszerűsége volt a Kínai Köztársaságban [12].

Weboldalak begyűjtése és indexelése

A továbbiakban részletesen ismertetjük egy keresőrendszer alkotóelemeit és a rájuk bízott elvégzendő feladatokat. Az *1. ábra* vázlatos áttekintést nyújt egy kereső szerkezeti felépítéséről.



1. ábra **Egy keresőrendszer felépítése**
Forrás: FRIEDMAN E. – UHER M. – WINDHAGER E.: Keresés a világhálón

A keresők első, lényeges feladata az oldalak meglátogatása és begyűjtése, amit speciális szoftverek, ún. keresőrobotok (*crawlers, web robots, web spiders, bots*) segítségével valósítanak meg. Ezek a programok folyamatosan és bizonyos időközönként újra és újra átfésülik a webet. Egy keresőrobot választhat egy népszerű, de megbízható oldalt kiindulópontjául, illetve dolgozhat egy korábbi, meglévő adatbázis alapján is. A robotnak le kell töltenie az általa meglátogatott oldalt, és át kell adnia azt az indexelőnek*. Ezután az oldalon lévő hiperhivatkozásokat nyomon követve ugyanígy kell eljárnia a hivatkozott oldalakkal is.

Számos esetben bizonyos időkülönbség jelentkezik a begyűjtés és az indexelés, valamint az eredmény keresőbe történő beépülése között. Ezért az oldalak begyűjtését és indexelését két, párhuzamosan zajló feladatnak kell tekintenünk. A keresőrobotok tehát nem végeznek semmilyen elemzést a meglátogatott dokumentumon, hanem csak nyomon követik a hivatkozásokat és letöltik a felfedezett oldalakat. Látszólag a robotok nagyon hasonló módon működnek, azonban jelentős különbségek figyelhetők meg a viselkedésükben. A robotok további feladata a meglévő, begyűjtött dokumentumok frissítése az adattárba. Ha például módosul egy oldal, akkor annak az újabb verzióját a megváltozott metaadataival együtt a robotnak el kell helyeznie az adatbázisban, a régit pedig törölnie kell.

Egy robot számára fontos szempont az, hogy mely hivatkozásokat kövessen nyomon, és mely oldalakat keressen fel, valamint lényeges kérdés, hogy milyen gyakran végezze el az oldalak begyűjtését. Egy keresőrendszer általában több robotot is alkalmaz a weblapok begyűjtésére. Emiatt a hálózati forgalom megnövekszik. A robotok igyekeznek nem folyamatosan terhelni egy szervert különböző kérésekkel, hanem időben elosztva küldik neki a kéréseket [10].

A „crawler control” modul látja el a robotok irányítását és munkájuk összehangolását. Az oldalak begyűjtése közben egy prioritási sort használ, amelyben a még meg nem látogatott oldalak címei szerepelnek fontossági sorrendben. A sor elejéről kiveszi a címeket és a hozzájuk tartozó oldalakat, letölti és kigyűjti belőlük a hivatkozásokat. A felderített linkekről eldönti, hogy melyiket kell követniük a robotoknak, ezeket beteszi a prioritási sorba, a többit pedig elhagyja. A begyűjtés addig tart, amíg a helyi erőforrások, mint például a tárolókapacitás, vagy egyszerűen a meglátogatott oldalak el nem fogynak [3].

A webmestereknek vagy a tartalomgazdáknak módjukban áll a robotok számára megtiltani egyes oldalak begyűjtését, az oldalon lévő hivatkozások követését és az oldal archiválását. Ezt a *Robot Kizárási Szabvány*ban (*Robot Exclusion Standard*) megszabott módon tehetik meg [6]. Ha egy weblapra nem hivatkozik egy másik oldal, akkor a keresőrobot nem fogja azt megtalálni. Ezért az új

* A fenti ábra tükrözi, hogy a robot és az indexelő között csak közvetett kapcsolat valósulhat meg az adattáron keresztül.

honlapokat tanácsos manuálisan regisztrálnunk az egyes keresőknél, amelyek így indexelni tudják őket. A keresőket lekérdezhetjük arról, hogy egy adott oldal indexelve van-e náluk. Azonban a különböző keresőknél eltérő lekérdezéseket kell alkalmaznunk erre a célra [9].

A keresőrobotok által begyűjtött oldalak az adattárba (*repository*) kerülnek, amelynek elsődleges feladata az oldalak egyenkénti tömörítése és szekvenciális tárolása. Ezenkívül a rendszer nyilvántartja egy állományban a dokumentumok pontos elhelyezkedését [3].

A keresők másik lényeges összetevője az indexelő (*indexer*), amelynek fő feladata az adatbázisban lévő begyűjtött oldalak elemzése és az indexelendő kifejezések kigyűjtése [3, 10]. Az indexelő tulajdonképpen az adattárra támaszkodik. A feldolgozás elején két problémával találkozik az indexelő. A weben előforduló oldalak elemzése összetett feladat. Ezt nem csupán a dokumentumok heterogén kialakítása okozza, hanem az egy-egy adott formátum esetén előforduló hibák is, például szintaktikai hibák a HTML dokumentumokban. A másik probléma az, hogy az indexelőnek szét kell tudnia választani a fontos és a kevésbé fontos kifejezéseket egy dokumentumban.

Erre egy lehetséges megoldás az, hogy figyelembe vesszük a szavak gyakoriságát és eldobjuk a legkisebb, valamint a legnagyobb gyakoriságú szavakat. Az előbbieket azért, mert azok nem lehetnek fontosak, hogyha csak néhány alkalommal fordulnak elő, az utóbbiakról pedig nagy valószínűséggel állítható, hogy felesleges szavak a dokumentumban. Azt is feltételezzük, hogy a töltelék- és egyéb szavak eloszlása eltérő egy dokumentumban. Tehát a szavak eloszlásának elemzésével a szavak gyakorisági kategóriákba sorolhatók.

A gyakorlatban azonban elterjedt egy másik megközelítés is, amelyben nyelvenként hoznak létre egy ún. tiltott szólistát (*stopwords*): ez foglalja magába a tartalmi szempontból feleslegesnek tekintett szavakat [10]. Az ilyen lista meggátolja a névelők, a kötőszavak és más, szinte minden dokumentumban előforduló szavak indexelését [3]. Tehát ez a módszer rendkívül gyors, egyszerű és könnyen használható.

A megmaradt releváns kifejezéseket bizonyos jellemzőikkel együtt gyűjti ki a dokumentumból az indexelő. Fontos jellemzőnek minősül a szó előfordulásának helye, mint például az oldal címe, a metaelemek, az oldalon belüli pozíció [10]. Továbbá az indexelő létrehoz egy indexet, amely minden releváns kifejezéshez hozzákapcsolja az őt tartalmazó URL-ek listáját [3]. A kigyűjtött indexelendő kifejezéseket és jellemzőiket a tényleges keresés és sorrendezés során veszik alapul a keresők.

A találatok sorrendezése, rangsorolása

A keresők működésének a leglényegesebb vonása a találatok megfelelő fontossági sorrendben történő megjelenítése a felhasználók számára. Ezért a keresőknek jelentős alkotóeleme a *Rangsoroló modul*, amely egy adott keresésre automatikusan sorrendezi a találatokat fontosság szerint [3]. Az indexelt adatmennyiség megnövekedésével vált egyre fontosabb feladattá a találatok pontos sorrendezése. Mivel a felhasználók csak az első 10-20 találatot szokták áttekinteni egy adott keresésnél, ezért rendkívüli fontos, hogy a kereső mely találatokat jeleníti meg a találati lista elején [10]. Kereséskor célunk a témában íródott legszínvonalasabb weblapok felkutatása, melyhez az oldalakat rangsorolni kell [3]. Az egyes keresők által használt rangsorolási szempontokról általában keveset tudunk, de a fő elvek ismertek.

Az egyik legalapvetőbb sorrendezési szempontnak minősül *a keresőkifejezés helyének vizsgálata* a dokumentumban. A keresők nagyon gyakran előnyben részesítik azokat az oldalakat, amelyeknek a címében is megtalálható a keresendő kifejezés. A találatok sorrendezésénél azt is figyelembe vehetik, hogy a dokumentum mely részében jelenik meg először a keresőkifejezés. Itt az alapelv az, hogy a weblap szempontjából releváns kifejezések nagy valószínűséggel fordulnak elő már a bevezetésben is, vagy legalábbis a dokumentum elején. Egyes keresők az oldal fontosságának meghatározásához számításba veszik a fontméretet is, és következtetésekre jutnak a szavak közti távolságokból is, valamint elemzik a HTML metaelemeket. A metaadatok segítségével közölhetjük honlapunk tartalmának összefoglalóját, valamint az oldalunkra vonatkozó kulcsszavakat. Ezeket a háttér-információkat is hasznosíthatják a keresők a rangsorolás, valamint a keresés közben is.

Másik jelentős vizsgálati szempont *a keresőkifejezések előfordulási gyakorisága*. Itt azzal a feltételezéssel élhetünk, hogy ha egy dokumentumban egy bizonyos kifejezés gyakran fordul elő, akkor az fontos a téma szempontjából.

Ebben az esetben természetesen kivételt képeznek a tiltott szavak listáján lévő kifejezések. Továbbá lényeges, hogy ne csak az egyes szavak előfordulási gyakoriságát kövessük nyomon, hanem az adott szóösszetételekét is. A keresők sokszor tanulmányozzák felhasználóik reakcióit is. Ha például a felhasználók többsége nem az első találatra kattint a szolgáltatott találatlistában, akkor nagy a valószínűsége annak, hogy rossz a találatok rangsorolása, és nem az első helyen szereplő oldal a legrelevánsabb.

Ezek a felsorolt sorrendezési szempontok sajnos lehetővé teszik, hogy könnyedén befolyásoljuk a találatok rangsorolását. Megfigyelhető az a tendencia, hogy az egyszerűen manipulálható rangsorolási szempontok egyre inkább háttérbe kerülnek és csökken a súlyuk a végső sorrend kialakításában. Helyettük pedig olyan kritériumokra helyeződik a hangsúly, amelyeket nehezebb befolyásolni. Itt megemlíthetők például olyan módszerek, amelyek az oldalak közti linkstruktúrát veszik figyelembe [10].

A találatok rangsorolásánál kényes etikai kérdésként merülhet fel az, hogy a kereső jó pénzért nem árul-e kulcsszavakat a cégek számára. A megvásárolt kulcsszóért cserébe az adott cég webhelye az első 10 találat között szerepelhet. Ez nem jellemző a nagyobb keresőkre, azonban a felhasználói kulcsszavakhoz kapcsolódó reklámok eladása széles körben elterjedt gyakorlat. Ezekben az esetekben a szoftverfejlesztők úgy változtatják meg a keresők relevanciarangsorolási algoritmusát, hogy az eladott kulcsszó a felhasználót rögtön arra a webhelyre vezesse, amely korábban megvásárolta azt [4].

Egyes keresők a linkhez tartozó szöveget nem a linket tartalmazó, hanem a link által hivatkozott oldalhoz tartozónak veszik. Az ilyen típusú linket horgonynak hívjuk, amit bizonyos keresők a találatok rangsorolásakor használnak fel [10]. A Google együtt kezeli a linkek szövegét azokkal a weboldalakkal, amelyekre azok ténylegesen hivatkoznak. Ennek a módszernek számos előnye van: a linkek sokszor pontosabb leírást nyújtanak a hivatkozott oldalakról, mint maguk az oldalak; továbbá olyan oldalakat is megkaphatunk, amelyeket a keresőrobot nem gyűjtött be a webről [1, 10].

Sokan vélekednek úgy, hogy a Google kereső népszerűségét annak köszönheti, hogy a találatokat minőségileg jobban rangsorolja, mint a többi kereső. A Google alkalmazza a fentebb ismertetett általános módszereket, továbbá kialakít egy olyan speciális algoritmust is, amely kizárólag a linkstruktúrát alapul véve határozza meg az egyes dokumentumok fontosságát. Ezt a fontosságot a kereső megfelelően súlyozza és a többi faktort egyaránt figyelembe véve, dönt a végső sorrend kialakításáról [10].

A Google kereső PageRank algoritmusa

A PageRank (PR) valós szám, amely egy adott oldal fontosságát tükrözi. A Google kereső a PageRank algoritmust alkalmazza az általa indexelt oldalak fontosságának meghatározásához, amit figyelembe vesz a rangsorolás során. A Google más egyéb szempontokat is felhasznál a sorrend kialakításakor, amelyek közül csak egy a PageRank érték, azonban ez az egyik legfontosabb. A PageRank-kel kapcsolatos eredmények megtalálhatók [10]-ben. Fontosságuk miatt néhány alapvető megállapítást részletezünk a továbbiakban.

A PageRank algoritmus alapgondolata, hogy amikor egy oldal hivatkozik egy másik weblapra, akkor a forráweboldal tulajdonképpen ajánlja a hivatkozott weblapot. Tehát az oldal létrehozója azért tüntette fel a linket az oldalán, mert a másik lapot valamilyen szempontból fontosnak tekintette. Emellett azt is figyelembe kell vennünk, hogy a hivatkozó oldal mennyire fontos, mert egy fontos oldalnak többet ér a hivatkozása. Eredményül egy rekurzív algoritmust kapunk, ami azt fejezi ki, hogy egy oldal fontos, ha mérvadó oldalak hivatkoznak rá. Ez a modell természetesen vitatható, hiszen lehetséges, hogy csak rossz példaként hozunk fel egyes weboldalakat, és nem arra szeretnénk velük célozni, hogy ők értékes oldalak. A gyakorlat azonban az eredeti alapötlet sikerességét igazolja, hiszen kevésbé meghatározók ez utóbbi linkek az interneten [10].

Az alapalgoritmust [7]-ben közölték először. (Nagy valószínűséggel feltételezhetjük azt, hogy a Google most már egy másik változatát használja az itt tárgyaltaknak, amiről azonban nem tájékoztatják a nyilvánosságot [10]). Ez a rekurzív egyenlet a weboldal fontosságára egy megközelítőleges becslést nyújt [1]. Érdekesség, hogy a szerzők egyik cikkükben pontatlanul adták meg az egyenlet első tagját és az így terjedt el a szakmában széles körben. Ez a változat megtekinthető az alábbiakban:

$$PR(A) = (1-d) + d \cdot \left(\frac{PR(t_1)}{C(t_1)} + ... + \frac{PR(t_n)}{C(t_n)} \right)$$

Az egyenlet az *A* oldal PageRank értékét határozza meg. Az egyenletben $t_1...t_n$ jelöli azokat az oldalakat, amelyek *A* oldalra mutatnak. $PR(t_i)$ fejezi ki az *i.* ilyen oldal PageRank értékét, azaz annak a fontosságát. A *d* paramétert egy skálázó faktornak tekintjük, aminek értéke *0* és *1* közé eshet. A *d* értékét a szerzők *0,85*-nek határozták meg [1, 10]. *C*-vel jelöljük az egy oldalon lévő összes kimenő hivatkozás darabszámát. Például, ha $C(t_i)$ értékét 24-nek vesszük, az azt jelenti, hogy az *i.* oldal összesen 24 darab kimenő hivatkozást tartalmaz, amelyek közül egy biztosan az *A* oldalra hivatkozik. Az eredeti algoritmus nem számol azzal az esettel, hogy mi történik akkor, ha egy oldalról több link is hivatkozik egy másik oldalra.

Az egyenlet tehát a következőt jelenti: az *A* oldal az első olyan oldaltól, amely hivatkozik rá, $PR(t_1)/C(t_1)$-nyi szavazatot kap, azaz a t_1-es oldal egyenletesen elosztja a saját fontosságát a kimenő hivatkozásai között. Ha t_1 oldalon egyetlen kimenő link található, akkor *A* megkapja a teljes $PR(t_1)$ értéket, ha három, akkor csak t_1 fontosságának a harmadát stb. Ugyanezt az elvet követjük az összes többi olyan oldal esetén, ahonnan találunk hivatkozást *A*-ra. Ezután ezeket a fontosságokat összeadjuk és megkapjuk *A* oldal fontosságát. Ebből tehát az következik, hogy kedvezőbb PR értéket kapunk, ha egy alacsonyabb PR értékű lap mutat ránk, mintha egy magasabb, ha az alacsonyabb fontosságú lapon nem sok kimenő link található. Egy dolgot azonban biztosan kijelenthetünk: ha oldalunkra több oldal hivatkozik, nem számít, hogy milyen rangos oldalak, valamilyen mértékben nőni fog a fontosságunk.

A *d* faktornak köszönhetően egy bizonyos oldal nem a teljes fontosságát osztja szét a kimenő linkjei között, hanem annak csak a 85%-át. Ahhoz, hogy megértsük ezt az összefüggést, szükségünk van egyrészt a javított PageRank egyenletre és a PageRank algoritmus egy újabb jelentésének bemutatására. A javított PageRank egyenletet tehát a következőképpen adhatjuk meg, ahol *N* az összes indexelt weblap számát jelenti.

$$PR(A) = \frac{(1-d)}{N} + d \cdot \left(\frac{PR(t_1)}{C(t_1)} + ... + \frac{PR(t_n)}{C(t_n)} \right)$$

A PageRank algoritmus egy olyan modellnek is tekinthető, amely a „véletlen szörfölő” viselkedését tükrözi. Egy ilyen felhasználó véletlenszerűen elindul egy weboldalról és a hivatkozásokra véletlenszerűen kattintva folyamatosan előrehalad. Nem is figyeli meg, hogy hova kattint, hanem egyenletes eloszlás szerint választ a meglévő hivatkozások közül. Ezzel magyarázható az, hogy a PageRank algoritmus a kimenő linkek számával elosztja egy bizonyos oldal fontosságát. Mindez addig tart, amíg szörfölőnk meg nem unja a kattintgatást és egy másik véletlenszerűen kiválasztott weboldalon nem indul el. Ez az egyenlet egy valószínűségi eloszlást határoz meg, ahol egy-egy weboldal PageRank értéke egy valószínűségnek (0 és 1 közötti valós szám) felel meg. Ebben a modellben az összes weboldal PageRank értékeinek összege maximum 1 lehet. Ez a megállapítás csak abban az esetben igaz, ha a felhasználónk egy adott oldalon mindig talál legalább egy hivatkozást, amelyen továbbhaladhat [1,10].

Ha webszájtunk olyan oldalt tartalmaz, amelyre ugyan mutat link, de belőle nem indul kimenő hivatkozás, akkor a szájt nem veszi fel a maximális PageRank értéket. Lógó (*dangling*) oldalnak hívjuk az ilyen oldalakat. A Google figyelmen kívül hagyja a lógó oldalakat, mert azok ellentmondanak a PageRank algoritmus által használt „véletlen szörfölő” modellnek. A Google tehát szűri a lógó oldalakat (az elhagyások miatt esetlegesen újonnan keletkezett lógó oldalakat rekurzív módon szintén figyelmen kívül hagyja). A megmaradt linkstruktúrában kiszámolja a pontos PR értékeket. Ezután fokozatosan visszahelyezi a lógó oldalakat és meghatározza azok fontosságát is a már kiszámított PR értékek alapján.

A Google nem csupán a linkstruktúrát elemzi, hanem egyéb tényezőket is figyelembe vesz az oldalak rangsorolásakor. Például sokszor negatívan értékeli azt, ha bizonyos, megbélyegzett oldalakra mutató hivatkozásokat tüntetünk fel az oldalunkon. Nyomon követi azt is, hogy az oldalra történő hivatkozások ugyanabból a doménből, földrajzi területről származnak-e. Tehát a rangsorolás szempontjából többet ér az, ha egy „független” valaki hivatkozik ránk, mint ha egy „ismerős” szavaz nekünk bizalmat [10].

A PageRank módszer manipulálása sokkal nehezebb feladat, mint a szöveges dokumentumok tartalomalapú sorrendjének befolyásolása. Ennek az az oka, hogy a web nagyobb részét kell módosítanunk, valamint hivatkozások sűrű szövevényével kell azt ellátnunk. A Google által alkalmazott rangsorolási módszer nagyjából ismert a nagy

nyilvánosság számára, ezért a világban számos cég specializálódott különféle manipulatív megoldások használatára, amelyekkel a saját forgalmukat tudják indokolatlanul befolyásolni. A cégeknek ezt a törekvését finomabb változatban „kereső optimalizálásnak" hívjuk, erősebb változatban pedig „hivatkozás spam-nek". A PageRank támadásának egyik közkedvelt módszere a linkfarmok létrehozása. Ilyenkor nagyszámú és sok szerverre kiterjedő, részben értékes oldalak másolatát, részben számítógéppel generált weblapokat tartalmazó oldalcsoportot alakítanak ki. Itt az oldalak mindegyike a céloldalra hivatkozik, ezáltal magas fontosságot tulajdonítanak nekik [2].

Problémák az internetes kereséssel és a megoldási kísérletek

A kereséssel kapcsolatos problémákat öt fő csoportba soroljuk, amelyek a következők:

1. Általános problémának tekinthető az internet hatalmas mérete, ami nemcsak a keresést, hanem az oldalak begyűjtését is nagymértékben befolyásolja. A weblapok meglátogatása és feltérképezése időigényes feladatot jelent még a legjobb keresők számára is.
2. Az utolsó begyűjtés óta eltelt idő alatt az internet tartalma és szerkezete megváltozik, ami további nehézségeket eredményez [5].
3. A keresőrendszerek számára általában elérhetetlenek azok az interneten meglévő tartalmak, amelyek a mély web (*deep web*) körébe sorolhatók.
4. A keresőrobotok nem gyűjtik be a dinamikus weblapokra mutató hivatkozásokat.
5. Az internetes keresők nem a felkutatható dokumentumok és a keresőkérdés jelentésével foglalkoznak, hanem csupán a szöveges alakkal.

A továbbiakban részletezem, hogy milyen módszerekkel próbálják megoldani ezeket a felmerülő problémákat.

Az óriási adattömeg kérdését oldják meg a metakeresők, amelyek párhuzamosan több rendszerrel kerestetnek. Így azok az internet nagyobb részét képesek átfésülni [10]. Növelik a találati esélyünket az ismeretlen témák esetében, valamint átfogóbb képet nyújtanak számunkra a weben fellelhető információkról egy adott témában.

A gyorsan változó tartalom kezelésére használt leglényegesebb módszer, hogy a változás mértékét és gyakoriságát is eltárolják a weblapok tartalmával együtt, majd a gyakran és jelentősen változó oldalakat sűrűbben látogatja újra a robot. Fontos továbbá az RSS csatornák indexelése is, mert így gyorsan értesül az új tartalmakról a keresőgép. További lehetséges módszer az oldalak begyűjtésének fókuszált módja (*focused crawling*). Ennek a módszernek lényege az, hogy nem követünk minden hivatkozást, hanem valamilyen szempontrendszer szerint egy bizonyos területhez kapcsolódó oldalakra szűkítjük a keresési teret, például nevezetes hírportálok meglátogatására. A fókuszált begyűjtést végző robotokkal kialakíthatunk egy-egy adott területre specializálódott keresőt is. Létrehozhatunk például egy olyan keresőt, amely orvosi tartalmak indexelésére és orvosi szakterületen feltett kérdések megválaszolására alkalmas.

A mély web csoportjába tartoznak a weben keresztül lekérdezhető adatbázisok, a nem szöveges formában található tartalmak, valamint a jelszavas vagy IP címalapú védelem mögé rejtett statikus oldalak. Ez az adatmennyiség azért nem elhanyagolható, mert becslések szerint a mély weben nagyságrendekkel több információt tárolnak, mint a hagyományos weboldalakon. A mély web és a sekély web (*surface web*) közötti lényeges különbség, hogy ez utóbbit az általános keresőgépek visszakeresik, a másikat azonban nem. A szerver mindkét típusú web esetében eljuttatja a kért weboldalt a klienshez – ha az jogosult rá. A kettő között húzódik az ún. szürke zóna (*gray zone*), amit egyes keresők látnak (pl. a torrent keresők, a Flickr képkeresője), mások viszont nem. A mély web kezelését úgy is támogathatjuk, ha a keresők számára is elérhető metainformációkat közlünk az adatbázisok tartalmáról, valamint különböző csatolóprogramokat hozunk létre a nem szöveges állományokhoz (PDF, Excel, JPG stb.).

Újabb nehézség, hogy a keresőrobotok nem követik a dinamikus weblapokra mutató hivatkozásokat, ezáltal sok információhoz nem férnek hozzá. Ennek az az oka, hogy a dinamikus linkek gyakran hoznak létre hatalmas vagy esetleg végtelen keresési tereket. Ezeket keresőcsapdának (*spider trap*) nevezzük, amelyeket a keresőrobotok megpróbálnak elkerülni. Előfordul az is, hogy bizonyos szerverek megkísérlik álcázni magukat, és egy keresőrobotnak más tartalmat nyújtanak, mint amit egy böngészőnek. Napjainkban számos technika terjedt el a dinamikus oldalak indexelésének támogatására; lényegük, hogy elhitetjük a keresőrobotokkal, hogy statikus hivatkozást követnek.

Az internetes keresők számára a legnagyobb probléma, hogy nem a fellelhető dokumentumok és a keresőkérdés jelentésével foglalkoznak, hanem csupán a szöveges alakkal. A nyelvi problémákat tulajdonképpen az okozza, hogy a mai eszközökkel történő információ-visszakeresés túlságosan a letárolt szöveges információ tényleges alakjára épül. Ennek egyik következménye, hogy a nem szöveges dokumentumok által hordozott információk nem kereshetők vissza automatikusan. További hiányosságként kiemeljük, hogy a keresőrendszerek nem ismerik a fogalmak jelentését és a fogalmak közötti kapcsolatokat, ezért nem képesek különféle következtetések levonására [10]. Ezt a problémát a szemantikus keresők orvosolják hatékonyan.

Az internetes keresőknek létezik egy másik fajtája, a webes katalógusok, amelyek emberek által összegyűjtött oldalakat tesznek visszakereshetővé. Ezek a katalógusok eredményesen oldják meg a jelentés, azaz a szemantika megragadását, ami az oldalak begyűjtését és indexelését végző emberek feladata. Előnyük, hogy oldalaik megbízhatók és minőségük garantált, mert emberek válogatják ki őket. Itt nemcsak szöveges keresésre van lehetőségünk, hanem témakategóriák között is böngészhetünk. Ezekben a gyűjteményekben nagy segítséget jelent, hogy az oldalakat emberek olvasták végig kategorizálásukkor. Legnagyobb hátrányuk viszont, hogy a létező weboldalaknak csak kis hányadát tartalmazzák. Ezenkívül meg kell említenünk a kérdésátalakító keresőket is, amelyek szintén a jelentés megragadására törekednek. Feladatuk, hogy megpróbálják jobban értelmezni a feltett keresőkérdést és azt úgy átalakítani, hogy az új keresőkérdés már jobb találatokat eredményezzen. Egy ilyen átalakításhoz a keresőknek kell, hogy legyen valamilyen matematikai formalizmussal leírható háttértudásuk.

A szemantikus webirányzat hatékonyan oldja meg a jelentéssel kapcsolatos problémakört, amelynek fő célja, hogy jelentést vigyen a webre. Ez úgy válik lehetségessé, hogy a weboldalak előállítói a webes tartalmakhoz szabványos formában metaadatokat rendelnek, a szemantikus web pedig biztosítja számunkra, hogy ezen metainformációk alapján következtetéseket vonjunk le. Jelenleg a metainformációk ugyanolyan heterogén formában fordulnak elő, mint maguk a webes dokumentumok. Ezért a szemantikus web fejlesztőinek elsődlegesen a metainformációk és a következtetéshez szükséges háttértudás egységes és feldolgozható alakban történő leírására kell törekedniük [10].

Irodalom

1. BRIN, S. – PAGE, L.: The anatomy of a large-scale hypertextual web search engine. = Computer Networks and ISDN Systems, 30. köt. 1-7. 1998. p. 107–117. *http://infolab.stanford.edu/pub/papers/google.pdf* (2007.03.02.)
2. BENCZÚR A. – BÍRÓ I. – CSALOGÁNY K. – RÁCZ B. – SARLÓS T. – UHER M.: PageRank és azon túl: Hiperhivatkozások szerepe a keresésben. = Magyar Tudomány, 167. köt. 11. sz. 2006. p. 1325–1331. *http://www.matud.iif.hu/06nov/07.html* (2007.07.17.)
3. FRIEDMAN E. – UHER M. – WINDHAGER E.: Keresés a világhálón. = Híradástechnika, 58. köt. 3. sz. 2003. p. 20–24. *http://www.cs.elte.hu/~hexapoda/kereses.pdf* (2010.04.10.)
4. FROEHLICH, T. J.: Case study 5.1: Developing search engine evaluation criteria. = Library evaluation. Libraries Unlimited, 2001. p. 185–200.
5. HAWKING, D. – CRASWELL, N.: Very large scale retrieval and web search. = TREC: Experiment and evaluation in information retrieval / Ellen Voorhees, Donna Harman editors. MIT Press, 2005. *http://es.csiro.au/pubs/trecbook_for_website.pdf* (2007.07.10.)
6. KOSTER, M.: A method for web robots control. Technical report, Internet Engineering Task Force (IETF), 1996. *http://www.robotstxt.org/norobots-rfc.txt* (2010.04. 22.)
7. PAGE, L. – BRIN, S. – MOTWANI, R. – WINOGRAD, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. *http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1999-66&format=pdf&compression=&name=1999-66.pdf* (2007.07.17.)
8. SCHWARTZ, C.: Web search engines. = Journal of the American Society for Information Science, 49. köt. 11. sz. 1998. p. 973–982.
9. SULLIVAN, D.: Checking your listing in search engines, October 2001. *http://searchenginewatch.com/webmasters/article.php/2167861* (2007.07.10.)
10. SZEREDI P. [et al.]: A szemantikus világháló. = A szemantikus világháló elmélete és gyakorlata. Budapest, Typotex, 2005. p. 17–59.
11. A világ előrehaladása az információs társadalom terén 1998–2008. World Progress Report 2008. Készít. a BME-UNESCO Információs Társadalom-

és Trendkutató Központjának (ITTK) kutatócsoportja. Budapest, 2007. március *http://www.ittk.hu/web/docs/ITTK_WPR1998-2008.pdf* (2008.01.28).

12. Web search engine (Webes keresőgép) szócikk. = *http://en.wikipedia.org/wiki/Web_search_engine#History* (2010.04.22.)

Beérkezett: 2010. V. 25-én.

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

Hangtalan társalgás a számítógéppel

A *Silent Speech Interface* lényege, hogy az emberek beszédmozgásával kapcsolatos izommozdulatait figyeli, és ezek alapján értelmezi a hangtalanul közölt szavakat.

Tanja Schultz a *Karlsruhei Technológiai Intézet* (*KIT*) *Antropomatikai Intézet*éhez tartozó *Kognitív Rendszerek Tanszék*én dolgozik. Az intézményt februárban alapították, összesen 120 kutatója van. Schultz szakterületének az emberi biojeleken alapuló technológiák és alkalmazások számítanak, ide értve az izom- és agyi tevékenységen alapuló beszédfelismerést és -interpretációt. Az antropomatika egy mesterséges szó, amelyet a görög anthropos (ember) és az automatikára utaló matik szavakból hoztak létre. Az antropomatika alatt az ember és a gép szimbiózisát érti a tudomány. A célja az embereken alapuló rendszerek kutatása és fejlesztése informatikai eszközök segítségével.

„A Silent Speech Interface lehetővé teszi, hogy hangtalan beszédet továbbítsunk. Az alapötlet a következő: a beszéddel kiküldött akusztikus jeleket gyakran eltorzítják a háttérzajok, például egy vonaton ülve vagy egy zsúfolt csarnokban. Éppen ezért működik annyira rosszul a beszédfelismerés hangos környezetekben. A technológia az EMG-n, vagyis az izomtevékenység elektromos jeleinek felderítésén és rögzítésén alapul. Vagyis az akusztikai jelek helyett mi ezekre építünk. Ha valaki hangtalanul mozgatja az ajkait, akkor a szájmozgásával kifejtett EMG-jeleket továbbítjuk a számítógépnek. Egy szoftver képes e jelek alakja alapján megmondani, hogy az illető melyik izma mozgott, majd pedig kiszámolja a hangot. A jeleket később beszéddé alakítja át, majd ezt egy számítógépes hang segítségével közvetíti. Ez a technika nemcsak a beszédfelismerésben használható, hanem lehetőséget ad arra is, hogy akik egy betegség miatt nem tudnak beszélni, újra megtehessék ezt. Ilyenek lehetnek, mondjuk a gégerákban szenvedők, akiknek megsérültek a hangszálaik. Ezenkívül lehetőség nyílik a hangtalan telefonálásra. Ez elsősorban olyan embereknek lehet érdekes, akiknek telefonálniuk kell, de nem szeretnének hangosak lenni. Érdekes lehet a szoftver az elektronikus banki és más beszédinterfészt használó programok esetében is. A Silent Speech Interface használatával még a bizalmas információkat (kódokat, jelszavakat) is biztonsággal lehetne kimondani.” – jelentette ki Tanja Schultz.

Az új interfész hibaaránya erősen függ a viselőtől és attól, hogy az illető mennyire jól artikulál. A jól artikulálóknál a hibaarány 5-10 százalék, az átlagos hibaarány pedig 100 szó esetében 10-20 százalékos. A rendszert tesztelték már kínai és német nyelvű embereknél is, valamint hamarosan kezdődnek a japán tesztek, de a technológia gyakorlatilag nyelvfüggetlen.

"A technológia még kiforratlan, ráadásul az embereket zavarja, ha kábelekkel az arcukon kell beszélniük. Ez a megoldás nem szép, de számos előnye van. A szenzortechnológia tovább fog fejlődni. Már most vannak kísérletek emberi szervezetbe ültethető EMG-elektródákkal. Az elektródák a jövőben még kisebbek lesznek; eljön az idő, amikor majd egyszerűen beinjekciózhatjuk őket.” – mondta Tanja Schultz.

A KIT-nél külön intézet foglalkozik azzal, hogy milyen következményekkel járhat a technika alkalmazása. Ha ugyanis valaki valóban adaptálható rendszereket fejleszt, akkor foglalkoznia kell azzal is, hogy ezeknek milyen morális, etikai és szociális következményei lehetnek a társadalomra nézve. Egy gép csak akkor hasznos, ha képes alkalmazkodni az emberi igényekhez. Sok technika vesz körül minket, de rengeteg időt és energiát fordítunk arra, hogy ezeket a saját igényeinkhez igazítsuk. Valójában ennek fordítva kellene lennie: a technikának kellene tudnia, hogy mit akarunk, és ahhoz kellene megfelelő szolgáltatásokat kínálnia.

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

(SzP)