Ahova a Google sem jut el

Kömlődi Ferenc (agent.ai)

2009. március 8., vasárnap 10:10 | Frissítve: 16 órája

Hiába ismer háromtrilliónál több honlapot a Google, az irdatlan mennyiség a világháló viszonylag csekély része. Mély, rejtett, láthatatlan, sötét jelzőkkel illetett felszín alatti rétege kívül esik a szabványos keresők indexelésén, a botok kutakodásán.

Az először 2001-ben használt Mély Web (Deep Web) kifejezés a neves számítástudományi szakembertől és konzultánstól, Mike Bergmantól származik. A rejtett adatbirodalom különböző forrásokból tevődik össze. Dinamikus tartalmakból, más oldalakhoz nem linkelt honlapokból, regisztrációt és bejelentkezést igénylő site-okból (private web), korlátozott hozzáférésű, szkriptelt, keresők által nem kezelt, specifikus formátumú - nem HTML/szöveg - tartalmakból. Az adatbázisokban tárolt pénzügyi információk, bevásárló katalógusok, repülő-menetrendek, orvosi kutatások és mindenfajta más anyagok sokasága mostanáig láthatatlan maradt a Google és a többi keresőmotor számára.

A Mély Web indexelhetetlensége a magyarázat arra, hogy a Google képtelen megválaszolni az olyan kérdéseket, mint például: „mennyibe kerül jövő péntekre a legolcsóbb Budapest-Frankfurt repülőjegy?" Mi könnyedén rátalálunk a feleletre, a keresők azonban nem. Fogalmuk sincs, hogyan bukkanjanak rá a különösebb nehézség nélkül beszerezhető információra.

A láthatatlan láthatóvá tétele

Az internetes keresőmotorok két részből állnak: az egyik összegyűjti, a másik rendszerezi az információt. Az előbbit hiperlinkeket követő automatizált böngészőprogramok (*robot*, *spider*, *web crawler*) végzik. Ez a módszer a web felszínén ugyan optimális, a mélyben viszont általában nem hatékony. Oda általában integrált keresőkkel (*federated search*) vagy algoritmusok helyett a kapcsolódásokat könnyebben észrevevő „humán böngészőkkel" igyekeznek eljutni.

2005-ben indult Search Subscriptions szolgáltatásával a Yahoo! lehetővé tette az előfizetett tartalom töredékében való kutakodást, amivel a láthatatlan web ugyan kis része, de valamennyi mégiscsak fokozatosan láthatóvá válik. Az integrált keresők, a humán böngészők és a Search Subscriptions csak a kezdetet jelentették, hiszen manapság egyre többeket foglalkoztat a Mély Web „felszínre hozatala."

Új technológiák alakulnak ki, s ha tevékenységüket siker koronázza, nemcsak a keresés minősége javul fel, hanem - Bergman szerint - hatásukra hosszabb távon az üzleti tevékenységek is megváltoznak. Például egy egészségügyi oldal kereszthivatkozásokkal kapcsolhatja össze a gyógyszeripari cégeket és a medicina legújabb kutatásait. Egy másik terület: helyi híroldalak kiegészülhetnek a kormányzati adatbázisokban tárolt közzétehető rekordokkal.

Warholt keresve

„Az automatizált böngészőprogramok csak a jéghegy csúcsáig jutnak el" - jelentette ki Anand Rajaraman, a Mély Web kutatására szakosodott Kosmix cég egyik alapítója (talán nem véletlen, hogy az egyik befektető Jeffrey P. Bezos, az Amazon.com vezetőségéből). - „A keresők tű után nyomoznak a szénakazalban, mi viszont a szénakazalt igyekszünk feltérképezni."

Ahhoz, hogy használható információt nyerjen ki a mélyrétegből, a motornak elemeznie kell a felhasználó keresési kulcsszavait, majd ki kell találnia, miként alkalmazza ezeket speciális adatbázisokra. Például, ha valaki begépeli, hogy „Warhol", a keresőnek tudnia kell, mely adatbázisokban lehet a legnagyobb valószínűséggel művészetre vonatkozó információt találni (múzeumok, aukciós házak), illetve milyen típusú kulcsszavakat fogadnak el ezek az adatbázisok. Az eljárás hiába tűnik könnyen kivitelezhetőnek, az adatbázis-szerkezetek változatossága és a lehetséges keresési terminusok komoly számítási kihívást jelentenek.

„Ez az elképzelhető legérdekesebb adatintegrálási probléma" - véli a Google Mély Web kutatásait vezető Alon Halevy.

A Google stratégiája a következő: egy program elemzi az összes „útjába került" adatbázist. Példánknál maradva, ha talál egy képzőművészeti oldalt, valószínűleg működő kulcsszavakkal - „Warhol", „pop art" stb. - kezdi, s mindaddig így jár el, míg nem kap egy linket. Ezt követően, a keresőmotor elemzi a találatokat, s kidolgoz az adatbázis tartalmára vonatkozó előrejelző modellt.

Eredmények és kételyek

A Utah Egyetem DeepPeep projektjét vezető Juliana Freire célja a nyilvános web összes adatbázisának átböngészése és indexelése. Természetesen sem ő, sem munkatársai nem tervezik a szótár valamennyi szavának felhasználását, hanem csekélyszámú egyszerű kulcsszóval kezdenek, s statisztikai módszerrel jutnak el a kívánt eredményhez. Freire szerint bármely adatbázisban tárolt tartalmak több mint kilencven százalékát képesek kezelni.

A Mély Web fokozatos megismerése felveti azt a kérdést is, hogy miként jelenítődjenek meg a különböző adattípusok a találati oldal agyonbonyolítása nélkül. Az olyan óriásoknak, mint a Google különösen óvatosan kell eljárniuk, különben a megszokottól eltérő és esetleg áttekinthetetlen megjelenítés elriasztja a felhasználókat.

Nem meglepő módon, többen vonnak párhuzamot a szemantikus webbel: az adatbázisok szerkezetének elemzése és a találatok keresztreferenciáinak automatizálása hasonló előnyökkel járna, ráadásul sokkal kevesebb pénzből kivitelezhető. „Óriási dolog ennyire szanaszét található adatforrások összekapcsolása" - lelkendezik Mike Bergman.