A szemantikus web visszakereső rendszereinek értékelése

A tanulmány a különböző szemantikus visszakereső rendszerek elemzését és értékelését tűzte célul, különös tekintettel a szemantikus dokumentumok kezelésére és visszakereshetőségére.

A szerzők egy rövid áttekintést adnak a jelenlegi tudásmodellekről és szemantikus visszakereső rendszerekről, hogy a főbb problémákat azonosítani tudják; valamint az értékelhetőség érdekében osztályozzák a szemantikus dokumentumok kezelésének főbb jellegzetességeit. Az elemzéshez 12 visszakereső rendszert választottak a fentebbi szempontok alapján. Az értékelés azt a módszert követi, amelyet a Desmet-projektben[1] alkalmaztak minőségi tulajdonságok értékeléséhez.

A szakirodalom áttekintése rávilágít néhány problémára, mellyel a szemantikus web a jelenlegi állapotában nem tud megbirkózni. A szemantikus kereső rendszerek elégtelenséget mutatnak a szakirodalom által legfontosabbnak ítélt tulajdonságaikban. Összességében a szemantikus webet e pillanatban a használhatóság alacsony szintjével lehet jellemezni, mely problémák többnyire a szemantikus dokumentumok kezeléséből erednek.

A szemantikus web koncepcióját a könnyebb hozzáférés és tudáselérés igénye hozta létre. Jóllehet, a különböző szerzők eltérő módokon közelítik meg a szemantikus web fogalmát, de az elképzelés lényege, hogy a web gépek által olvasható, egymással szemantikus viszonyban álló (ilyen módon kapcsolódó) adatok összessége, amely azt jelenti, hogy a köztük történő navigáció a dokumentumokat összefűző, jelentéssel bíró, hyperlinkekben megtestesülő kapcsolatok által történik. Ebből adódóan a szemantikus web szemantikus dokumentumokat kíván meg, amelyek lehetővé teszik a tudásreprezentációt és kölcsönhatásokat.

A szemantikus dokumentumokat olyan összetett információforrásokként tartjuk számon, mint amelyek egyedileg azonosítottak, szemantikusan annotáltak, és szemantikusan összekapcsolt adatelemekből állnak.

A metaadatszótárakat (*vocabularies of metadata*), ontológiákat és sémákat (*schemas*) szintén szemantikus leíró forrásoknak tartjuk, mivel ezek mindegyike kulcsszerepet tölt be a felhőben tárolt adatok (*linked data cloud*) közti együttműködésben, illetve a szemantikus webes alkalmazásokban. Mivel azonban az egyes reprezentációs modellek más nyelvtant és meghatározásokat használnak, ez problémákat okozhat, amikor össze akarjuk fűzni őket. A szemantikus weben túlnyomórészt uralkodó az RDF mint szabványos adatmodell használata; valamint egy szerializációs formátumé, ami XML-szintaxison alapul, ahol az adatforrásokat és a szótárelemeket, fogalmakat URI-k segítségével lehet elérni. Ezen tendencia ellenére szükség van egyéb nyelvekről és modellekről (XML Schema, KIF stb.) való transzformálásra.

A szemantikus weben létező dokumentumformátumok sokszínűsége magyarázza, hogy például amikor elmentünk egy keresést egy adattárolóba, akkor a CSV-, XML-, HTML-, és köztük RDF-fájlok sokaságát találjuk.

A tanulmány kitér arra, hogy az utóbbi idők egyik legsikeresebb vállalkozása a *Linked Data* projekt. Főbb előterjesztései négy elv köré csoportosíthatók: 1. minden entitás legyen pontosan identifikálva (pl. URI-k által); 2. minden objektum legyen hozzáférhető (pl. HTTP URI-k által); 3. az információk strukturálása szabványok által meghatározott módon (pl. RDF és SPARQL); valamint 4. integrálni az egyes entitásokat a köztük lévő kapcsolatok (pl. alkalmazási profilok, metaadatszótárak és hasonlók) által. A Linked Data ezáltal képes megosztott adatokat, illetve szemantikus források és adatelemek különböző generációinak repozitóriumait kezelni. Sikerét meghatározza, hogy alkalmazói képesek azonosítani, összekapcsolni, újrafelhasz-

nálni a Linked Data által már elérhetővé tett forrásokat.

Következésképpen, a szemantikus webre írandó kereső rendszer specifikációja nem csupán az RDF-dokumentumokra kell, hogy koncentráljon/korlátozódjon, de magába kell, hogy foglalja az eltérő típusú formátumokat és nyelveket is.

Tudásmodell a weben

A szemantikus web tudásmodelljének megalkotásához azonban szét kell választani a metaadatszótárakat az összetett, kidolgozott szemantikus dokumentumoktól. Mindezzel együtt a tulajdonképpeni mataadatszótárak elkülönítése magában rejt néhány nehézséget és kihívást: *A metaadatszótárak a felhőben jelen lévő, összetett szemantikus adatok (linked open data cloud, LOD Cloud) szétválasztása.*

Nagyon sok szótár ír le forrásokat, személyeket, intézményeket, de a legmegfelelőbb kiválasztására nincsenek meg az elégséges kritériumok. Egy korábbi elemzésben megfogalmaztak ugyan néhány szempontot a minősítéshez, mint: a szabályozottság szintje; állandóság; elemek száma; használati statisztika; népszerűség stb. Másik probléma a hasonló elemkészletek közti egyetértés hiánya, a sok átfedés, illetve a különböző szótárakban található, azonos elemeknek egymástól eltérő definícióval való szerepeltetése.

A megfelelő hozzáférés kiválasztása

Az internet örök változékonysága miatt az URI-k jelenléte és az általuk való navigálás több problémát is felvet. Az egyik legfontosabb az ún. „üres csomópontok” (*blank nodes*) jelenléte az adatszerkezetekben, amelyek olyan fölérendelt, összefoglaló kategóriát képviselnek, amelyek önmaguk nem rendelkeznek adatelemmel, az általuk „bennfoglalt” kategóriák viszont igen; URI-kkal hivatkozni azonban csak a valós tartalommal bíró elemekre lehet. Az „üres csomópontok” értéke legtöbbször szöveges meghatározás, mely önmagában kétségessé teszi a rájuk való, automatizált hivatkozást.

A keresési művelet és a fogalmi navigáció

A keresési művelet modellezését megnehezíti, hogy ez a műveletsor SPARQL nyelven van szabályozva, és a reprezentációhoz nemcsak ennek a nyelvtanát kell ismerni, de azoknak a fogalmaknak részletes környezetét is, amelyek a találatokban szerepelnek. Másként fogalmazva, szinte bizonyíthatatlan, hogy egy általánosan megfogalmazott kérdés a különböző elemkészletek elemeinek eltérő kombinációira vajon teljes mértékben a kívánt eredményt hozta-e.

A jelentésen alapuló navigáció az egyes szótárak elemei között hierarchikusan felépülő kapcsolatok mentén zajlik. Azonban az egyes szótárak között jelentős eltérés van mind a hierarchia részletezettségét, az egyes fogalmak alá tartozó elemek mennyiségét, illetve a meghatározásokat illetően. A szerzők a *Dbpedia*ból, *Wikipedia*ból vett, részletes példákkal igazolják, hogy nem egy esetben ugyanaz a fogalom mind alá- mind fölérendelt kategóriaként előfordul a létező szótárakban, így csaknem lehetetlenné téve a fogalmi navigálást.

Az értékelés módszere

A szakirodalom alapján a szemantikus webbel kapcsolatban az eddigiekben azonosított problémákat táblázat mutatja be összefoglalóan. Mivel ezek mindegyike „minőségalapú” elv, éppen ezért nehéz összemérni a klasszikus információkereső rendszerek teljesítményével. A szerzők a *Desmet-módszer*t ajánlják a különböző típusú szemantikus visszakereső rendszerek elemzésére és értékelésére, különös tekintettel az ilyen jellegű dokumentumok visszakeresésére és kezelésére való képességüket. A vizsgálat célja, hogy tisztázza, vajon ezek a szemantikus rendszertípusok megvalósítják-e a korábbi szakirodalom által velük szemben támasztott követelményeket, illetve, hogy jelen problémáik összefüggésben vannak-e a szemantikus webbel.

A Desmet-módszer egy egyszerű, megbízható, független vizsgálati módszer az informatikai fejlesztésekkel kapcsolatban, hasonlóan a funkcióanalízishez. A módszer segítséget kíván adni az értékelő vizsgálatokhoz, például maximalizálva a legjobb eszköz, metódus stb. azonosítását. A cikk nem tűzi célul a legjobb visszakereső rendszer kiválasztását, de segítséget nyújt az egyiknek vagy a másiknak kontextustól függő kiválasztásához. A szerzők azért is tartják alkalmasnak a módszert, mert képes alkalmazkodni a webes visszakereső rendszerek állandó fejlődéséhez, változásához.

A Desmet-módszer lépéseit követve, először meg kell határozni néhány speciális körülményt és követelményt a metaadatszótárakkal és visszakereső

rendszerekkel (*retrieve ontologies*) kapcsolatban. Másodszor, el kell végezni a tulajdonságok analízisét – amely lényegében a követelmények megfogalmazásán alapul –, és ezek összefüggéseit azokkal a jellegzetességekkel, amelyeket az adott specifikációk támogatnak. Végül el kell végezni az értékelést, és a Desmet-módszer szerinti értéket, valamint helyezési szintet hozzárendelni a vizsgált rendszerhez.

A szemantikus dokumentumok visszakereső rendszerének kiválasztása

A szerzők 12 szemantikus visszakereső rendszert gyűjtöttek össze, és úgy találták, hogy ezek nagyon is különbözők a funkciók tekintetében. Ennek eredményeképpen a visszakereső rendszereket a szemantikus keresők négy típusa szerint csoportokba sorolták, hogy az eredményeket értékelni lehessen.

A kereső rendszerek – az általuk a keresés során forrásul használt dokumentumok típusai alapján – a következők:

- Ontológiakereső rendszerek *(ontology search engines)*: ezek az alkalmazások szemantikus dokumentumok után kutatva pásztázzák a webet. A kereső motor indexeli a különböző ontológiákat. Ilyen pl. a *Swoogle*; *Sindice*; *Watson*.
- Metaadatra kereső rendszerek *(search engines for matadata)*: az effajta rendszer a metaadatok visszakeresésére törekszik, mint például *Linked Open Vocabulary* (LOV); *DataHub*.
- Ontológiatárak *(ontology directories)*: fogalomkészletek katalógusa, „kézi erővel” összegyűjtve; például *DAML Ontology Library*; *Protégé Ontologies*.
- Metaadattárak *(metadata directories)*: metaadatkatalógusok, források; mint például az *UKOLN*; a *Topic Maps' PSIs*; *RDA Vocabulary*; *Open Metadata Registry*.

Néhány olyan rendszert eleve kihagytak a vizsgálatból, amelyek technológiája nem metaadatleírásokon és szemantikus dokumentumok visszakeresésén, hanem információbányászaton alapul.

A szemantikus dokumentumok visszakereső rendszereinek értékelési szempontjai

Három táblázat mutatja azokat a kritériumokat, amelyeket a szerzők meghatároztak a források értékeléséhez. A szempontok az előző évek szakirodalmának megállapításain alapultak, és három fő csoportba sorolták őket:

- Sématámogatás *(schema management)*. Az ehhez kapcsolódó fogalmak: együttműködés; szabályozottság; interaktivitás és szemantikus keretrendszer.
- Jelentéstámogatás *(semantic management)*. A fogalmak jelentésével és kezelésével összefüggésben, a kapcsolódó kritériumok a következők: egyértelműség; többnyelvűség; szinonimák kezelése; kiterjeszthetőség stb.
- Kérdezhetőség *(queries)*. A keresési folyamatban a kapott találatok kezelése. Ez a kategória kiterjed a jelentés-meghatározásra; a fogalmi keresésre; a szövegösszefüggésen alapuló keresésre; és a dokumentum-visszakeresésre.

Az egyes kategóriákhoz tartozó fogalmakat tovább *egyszerű*nek vagy *összetett*nek minősítették. Az egyszerűek azok, amelyek mint feltételek fennállhatnak, vagy hiányozhatnak, továbbá a Boole-algebra műveleteivel kifejezhetők; az összetettek pedig egy számsoron kaphatnak különböző értékeket. Mind a két kategóriába tartozó fogalmaknál külön értékelték – hozzárendelt pontszámokkal – a fontosságot. Az egyes kategóriák fogalmait a hozzárendelt értékekkel táblázatokba foglalva rendszerezték, illetve kifejtve az eredményeket, részletesen elemezték.

Eredmények

Az összesítés szerint például egyetlen visszakereső rendszer sem támogatta a szabályozottságot, amely meghatározta volna azok létrehozását és működését. Ebből adódóan továbbá egyetlen rendszer sem követelte meg az egyértelműséget minden egyes alkotó fogalomra; valamint nem jelent meg az értékelhető tulajdonságok között a többnyelvűség.

A sématámogatás eredménye

A metaadat-regiszterek nem támogatják az átjárhatóságot, mivel „egy-az-egyben” megfeleltetésekkel dolgoznak az egyes sémák között. Csak néhány ontológiakereső, mint például a Watson elemzi a fogalmak között kapcsolatokat.

A metaadat-regiszterek és ontológiatárak gyakran biztosítanak extra funkciókat a felhasználóknak, így azok újabb és újabb elemeket tudnak beemelni a rendszerbe, szemben azokkal az alkalmazásokkal, ahol a használói beavatkozás – kivéve magát a keresési folyamatot – eleve korlátozva van.

A jelentéstámogatás eredménye

Tekintetbe véve a szemantikus keretrendszert, a metaadattárak nem használják ki a fogalmakban rejlő szemantikus lehetőséget, inkább jelképes leírásokat alkalmaznak. A metaadat-keresőkkel szemben az ontológiakeresők és ontológiatárolók kihasználják a *jelentést* mint a sémák sajátosságát, beleértve az egyéb sémákkal való kapcsolatokat is.

Sajátos módon ennél az egy kategóriánál jelent meg a *nyelviség* és a *változtathatóság*. Az általában használt sémadefiníciós nyelv RDF vagy XML. Másrészt, a sémák és a szemantikus reprezentációs modell közötti megfelelés „egy-az-egyben" típusú kapcsolat, amely az összes többi megfelelés felülvizsgálatát és frissítését maga után vonja.

Kérdezhetőség

Az elnyert találatok vonatkozásában elemezték a fogalmi és a jelentésen alapuló keresést. A metaadat-tárolók visszakeresési mechanizmusa a megadott címkéken és tulajdonságokon alapul. Ezzel szemben, a metaadat-keresők, ontológiakeresők, illetve ontológiatárak kiterjesztik a keresést a jelentés figyelembe vételével az általános kategória szintjére is, ugyanakkor támogatják az eredeti keresőfogalom jelentésén alapuló találatokat. A metaadat-tárolók a találati halmazt nem terjesztik ki a fogalmak kapcsolódási tartományára.

Végezetül, a dokumentum-visszakereshetőség szempontjából a metaadat- és ontológiatárak csak a sémán alapuló keresést tették lehetővé, az összehasonlításban szereplő, egyéb keresőmotorok a sémákba ágyazott dokumentumokat is megtalálták.

Az egyes kategóriákhoz korábban hozzárendelt fontossági értékek alapján számszerűen és százalékosan kiértékelték az egyes rendszereket, majd a kapott értékeket a kereső rendszerek szerint táblázatba foglalták, így ábrázolva, hogy a mataadat-keresők, metaadat-tárolók, ontológiakereső rendszerek, ontológiatárak összesítve milyen eredményeket értek el az egyes – sématámogatás, jelentéstámogatás, visszakeresés támogatása – kategóriákban.

A sématámogatás szempontjából a metaadat-keresők, ontológiakeresők és ontológiatárak egyaránt magas, nagyjából azonos eredményt értek el. A legalacsonyabb érték (metaadattárak) részben annak is köszönhető, hogy ezek a rendszerek kevésbé támogatják az elemek közötti megfeleltetést, valamint alacsonyabb a használói interaktivitás lehetősége.

A jelentéstámogatás kategóriában az ontológiatárak és ontológiakeresők érték el a legjobb eredményt; köszönhetően – többek között – az elemek közötti kapcsolatok széleskörű támogatásának. A legalacsonyabb értéket a metaadattárak kapták, ami leginkább az alkalmazási környezet korlátainak tudható be.

A kereshetőség kategóriában a metaadat- és ontológiakeresők végeztek az első helyen, mivel a keresés során figyelembe tudják venni a fogalmi összefüggéseket, illetve el tudják érni a találatok között a szemantikus dokumentumokat; a legalacsonyabb helyezést pedig – a metaadattárak esetében – éppen eképességek hiánya eredményezte. A végleges összesítésben az 1. ontológiakeresők, 2. metaadat-keresők, 3. ontológiatárak, 4. metaadattárak sorrend alakult ki.

Értékelés

Ebben a vizsgálatban a szerzők „szemantikus dokumentumnak" tekintettek minden olyan sémát és szabályozási dokumentumot is, amely szemantikus leírást alkalmaz a dokumentumok tartalmára vonatkozóan. Bizonyos szabványok – mint például a tématérképek, vagy az OWL – XML-sémával írhatók le, az RDF alkalmazása nélkül. Éppen ezért a téma alapos vizsgálata sem korlátozódhat csupán az RDF-dokumentumok visszakereshetőségére, karbantartására, tárolására. A kiterjesztett vizsgálat mindenképpen egyfajta kulcsot ad az egyéb források szemantikus leírásához is. És miután ezeket a dokumentumokat „szemantikusnak" minősítettük, ebből adódóan a szemantikus keresők által visszakereshetőknek, megtalálhatóknak kell lenniük. Kétségtelen továbbá, hogy a szemantikus web közössége előnyben részesíti az olyan nyílt szabványokat, mint az OWL, RDFS, szemben a tulajdonosok által védett, kódolt eredményekkel.

Már 2008-ban körvonalazódott egy szabályozott szemantikus kereső környezet kialakításának igénye. A kutatók remélik, hogy a szemantikus keresővel szemben elvárt és megfogalmazott igények valóban beillesztésre kerülnek majd a fejlődés során ebbe a környezetbe. A szemantikus web fejlődésében rejlő kihívások példájaként vizsgálták meg közelebbről a „linked data" fejlesztések kapcsán létrejött elemkészleteket, illetve értékelték a

szótárakat. Az első probléma a hozzáférés linkekre épülő természete. Számos probléma vetődik fel, amikor a kapcsolódó források „egy-az-egyben" megfeleltetést alkalmaznak a szótárak között. Ha ki akarnánk számolni a létező szótárak elemeiből összesen létrehozható sorozatok számát, elképesztően nagy – *(n-2)! – mennyiséget kapnánk, ahol „n" a létező szótárak száma. A *W3C Library Linked Data Incubator Group* nagy erőfeszítéseket tesz a szótárak összegyűjtése és értékelése érdekében, hogy ezeket a számokat csökkentse, de ez egy hosszabb folyamat. Egy másik lehetőség lenne egyetlen, központi forrás használata, mely az összes többi szótárhoz kapcsolódna; ekkor n-1 megfeleltetés jöhetne létre az összes lehetséges fogalom között. De a megoldásra váró probléma még akkor is az lenne, hogy a megbízhatóság érdekében milyen hatásoknak kell érvényesülniük a szótárak frissítése, javítása során.

Végigkövetve a vizsgálatot, amellyel a szemantikus visszakereső rendszerek fejlesztéseinek legfőbb követelményeit – mint például használhatóság (használó-központúság?) – azonosították, a kritikus pont ezen eszközök, források népszerűségének biztosítása lesz. A minél nagyobb számú felhasználó bevonása elemi feltétel ahhoz, hogy a szemantikus web és a linked data-technológiák elérhessék saját lehetőségeik legnagyobb kihasználtságát.

Következtetések

A tanulmány a szemantikus dokumentumok visszakeresésére szolgáló módszerek vizsgálatát és értékelését mutatta be. Az eredmények jelen pillanatban azt mutatják, hogy több, a szemantikus források visszakeresésére szolgáló rendszer hiányosságokat mutat azokban a minimális követelményekben is, amelyek pedig növelhetnék népszerűségüket. Ezek közül csak néhányat említve, a problémák a megbízhatóság hiányában; a források leírásának kevésbé szabályozott voltában; a használó-központúság alacsony szintjében; valamint az elemek elszigeteltségében keresendők, mely utóbbi nehézkessé teszi a keresést, és bonyolulttá a fogalmi navigációt.

Ahogy az elemzés kimutatta, egyedül az ontológiakeresők (NEM AZ ONTOLÓGIATÁRAK!) érték el az alig 50 százalékot a minősítés során. A szerzők értékelése szerint fejleszteni kell azokat a tulajdonságokat, amelyek elsődlegesek a szemantikus dokumentumok kezelésében; ilyenek például a szabályozottság, megbízhatóság, többnyelvűség és a keresés szemantikus jellegének kiterjesztése.

A folyamatban lévő kutatások jelenleg arra irányulnak, hogy hogyan lehet meghatározni a megbízhatóság és a minőségi linkek követelményeit. Noha néhány kereső eszköz statisztikák közzétételével segíti a felhasználókat a megfelelő szótárak kiválasztásában, az eddigi vizsgálatok hiányosságaként lehet említeni, hogy ezeknek az adatoknak a felhasználói viselkedésben és a kiválasztásban játszott szerepével még nem foglalkoztak kellőképpen.

Hivatkozás

[1] KITCHENHAM, Barbara: DESMET: A method for evaluating Software Engineering methods and tools: Technical Report TR96-09. University of Keele, Department of Computer Science, 1996.

/MORATO, Jorge Luis – SANCHES-CUADRADO, Sonja – DIMOU, Christos: Evaluation of semantic retrieval systems on the semantic web. = Library Hi Tech, vol. 31, no. 4 (2013) p. 638−656./

(Vass Johanna)

A *linked data* és a *big data* találkozása – a tudásszervezési rendszerek szempontjából

Big data

A változatos, komplex és hatalmas méretű digitális adathalmazok megjelenése a weben életre hívta a *big data* nevű jelenséget. Ilyen „nagy adatokat" termel például a közösségi média, az elektronikus kereskedelem, a kormányzat, a tudományos kutatás... A fogalomnak sokféle meghatározása született; másként közelítik meg ezt a témát a tudósok, a számítástechnikával és az információtudománnyal foglalkozó szakemberek, a tudománypolitikát irányítók és a finanszírozó szervek vezetői. Vannak, akik azt a technológiai kihívást hangsúlyozzák, amit az igen nagy méretű adatállományok –