Nagytömegű, struktúrált szövegek online szolgáltatása

Lépések a szemantikus web felé

Király Péter, pkiraly@tesujionline.com
Tesuji Magyarország Kft. (http://www.tesujionline.com)

W3C Szemantikus Web Műhelykonferencia
2006. április 13. csütörtök, Budapest, MTA SZTAKI

Az Arcanum Adatbázis Kft. 1991 óta folytat szövegdigitalizációs tevékenységet, aminek eddigi eredménye (formázási és szemantikai jelölőelemek nélkül számolva) körülbelül 10 gigabájtnyi 'tiszta' szöveg, ami kb. 4 millió nyomtatott oldalnak felel meg. Az offline kiadványokon alkalmazott Folio adatbáziskezelő egy olyan rendszer, ami valamiképpen megvalósítja a szemantikus rendszerek bizonyos célkütűzéseit, amennyiben lehetőséget teremt arra, hogy tetszőleges szövegben hierarchikus (szintek) és 'inline' (mezők) szemantikus egységeket hozzunk létre.

A hierarchikus jelölés célja, hogy a szövegben alá- és fölérendelt relációkat határozzunk meg (pl. egy könyv fejezetekből, ezek bekezdésekből állnak, egy újság évfolyamokból, lapszámokból rovatokból stb.). A mezők segítségével pedig tetszőleges predikátumokkal láthatunk el kisebb szövegegységeket (pl. a „Petőfi Sándor: Anyám tyúkja" karakterláncban a „Petőfi Sándor" kifejezést „szerzőnek" jelöljük, az „Anyám tyúkjá"-t pedig címnek).

A Folio ebben a vonatkozásban az XML-re emlékeztet, vagyis magában a szövegben kódolja a struktúrát, nem pedig valamilyen katalóguscédulán.

A feladat az, hogy hogyan tudjuk ezt a rendszert a weben reprodukálni, vagyis hogyan lehetne megvalósítani egy rendszert, ami olyan fájlformátumokkal tud dolgozni, amelyek függetlenek valamely konkrét programrendszertől (nem úgy mint például a Word .doc), ugyanakkor képes a meglévő szemantikai struktúrákat megőrizni és ezeken műveleteket végrehajtani. Ebből az igényből – hosszú keresgélés után – nőtt ki az Anacleto digitális könyvtári szoftver (illetve az ezt létrehozó cég). Jelenleg az 1.0 változatnál tartunk, az Arcanum mellett a világ legrégibb digitális könyvtára, a Project Gutenberg Consortia Center[1] és az „olasz MEK", a LiberLiber vette alkalmazásba.

A rendszer fő célja, hogy a különféle forrásból származó (intranet, internet, FTP, WebDav, helyi és hálózati fájlrendszerek), különféle formátumú (HTML, Word, Rtf, PDF, XMP/RDF, TEI-XML stb.) dokumentumokat kontrolláltan indexelje és – különféle kényelmi szolgáltatásokkal – egységes kezelői felületbe integrálja.

Mivel a rendszer kialakításának még csak egy kezdeti fázisán vagyunk túl, ambivalens érzések vannak bennem azzal kapcsolatban, hogy az Anacleto vajon belefér-e a „szemantikus web" kategóriába. Remélem a nap végére okosabb leszek e tekintetben, most azonban sorbaveszem azokat a szempontokat amelyekben megvalósítja a közös célkitűzéseket és amelyekben nem. Egy biztos: formailag – jelen pillanatban legalábbis – nem használjuk a vonatkozó szabványokban leírt fájlszerkezeteket, noha…

Szemantikus elemek és hasznuk

Az RDF és az Anacleto közötti fő különbség abban áll, hogy az előbbi metaadatokkal dolgozik,

1 http://www.gutenberg.cc/, http://www.gutenberg.com, http://www.gutenberg.us

azokat egy elkülönített helyen, formailag rögzített módon tárolja, az utóbbi pedig 'in situ' azon a helyen, ahol az a szövegben található. Ennek ellenére nem jelentene komolyabb gondot olyan állományok előállítása, amelyek a szövegből kiemelt mezőket RDF módon formázva egy elkülönített helyen tárolnák vagy mutatnák meg (hiszen az adatok adottak, és az indexelés ugyanezt a kinyerési technológiát (XSLT sablon) alkalmazza – sőt az adminisztrációs felületen létezik egy index-nézegető, ami „kiterítve", táblázatos formában mutatja meg a dokumentum szemantikus elemeit.)

Egy másik különbség, hogy a szemantikus web jelenlegi „mintaalkamazásai" olyan dokumentumok metaadatai, melyekben a szöveg és a metaadat jelentősen nem különbözik: képek, hírek, tárgyleírások stb. Ahol hosszabb szövegre használnak metaadatokat, ott az információ jelentős tömörítésére, összefoglalására van szükség (pl. a Dublin Core 12 mezőben leír egy teljes könyvet). Ilyen módon a szövegben való keresést jelentősen nem könnyíti meg az RDF használata.

A szemantikus adatok legfőbb jelentőségét az adja, hogy sokkal finomabb kereséseket lehet végrehajtani és pontosabb találati listánk lesz, mintha pusztán a „lapos" szövegben tennénk ugyanezt. Az Anacleto lehetőséget ad a különféle trükkös keresőkérdések megfogalmazására, de ennél tovább is megy: a felhasználónak megmutatja a keresett mezőben található értékek halmazát, sőt ezt a listát is lehet kereséssel szűkíteni. Így a felhasználó nem vakon tapogatózik, hanem a keresés tényleges elindítása előtt tisztában van azzal, hogy az adatbázisban létezik-e egyáltalán a keresendő terminus és ha igen, hány előfordulása van. Mivel az indexlistán is lehet balról és jobbról csonkolni, könnyen végigfuthatunk pl. 'szony' végződésű szavakon. Az alapvetőnek számító logikai operátorokon kívül lehetőségünk van közelségi és hasonlósági keresést futtani (példa az elsőre: *Kossuth* és *Deák* 5 szó távolságra; a másodikra: a vakondhoz – a karaktereket tekintve – 80%-nyira hasonló kifejezések).

Mivel – mint korábban szó volt róla – a szövegek alapesetben egy hierarchiát alkotnak, a keresésbe be lehet kombinálni ezt a szemantikus elemet is: vagyis leszűkíteni a keresés tartományát az általunk kijelölt tartalmi ágakra. A találati listán mindig látjuk a találat szókörnyezetét, amit ki is lehet vetíteni egy speciális tartalomjegyzékre, amiben csak a találati listán szereplő dokumentumok (illetve ezek szülői) vannak. A találati listáról kattintva mindig az adott találatra ugrunk, ami ki van világítva (ellentétben a Google-lel). A dokumentumban mindig látjuk az aktuális hierarchiát, ami szintén azt segíti, hogy felhasználó a dokumentumot a maga tágabb értelmezési kontextusába helyezhesse. (Kedvenc levéltári hasonlatomat idézve: egy szerelmes levélnek más a jelentése egy irodalmi hagyatékban, mint az ÁVH archívumában.)

Mindezek az információk ugyanabból a forrásból keletkeznek amit a szemantikus web az RDF fájlokban tárol.

Együttműködés

Az 1.0 változat jelenleg a saját kommunikációs szabályait használja, a következő változat elsődleges célja az, hogy különféle webszolgáltatási felületeket biztosítsunk és – amennyiben erre valamely felhasználótól igény jelentkezik – elkészítsünk egy automatikus RDF-előállító segédletet (említettem, hogy az ehhez szükséges részek rendelkezésre állnak).

Nyelvi elemek

Az Anacleto képes kezelni és a keresésbe bevonni különféle ontológiákat – köztük a WordNet 'Prolog' formátumát natívan is. Pl. A „The quick brown fox jumps over the lazy dog" keresőkérdést átengedve egy szinonímakereső szűrőn és a kapott értékeket boole operátorokkal ellátva a „(quick OR agile OR fast OR flying) AND (brown OR brownish) AND (fox OR bedevil OR befuddle) AND jumps AND (over OR across) AND (lazy OR slothful) AND dogs" keresőkifejezést kapjuk.

Látható, hogy a ragozott alakokhoz (jumps, dogs) nem találtunk szinonímát, vagyis először a keresőkifejezést meg kellene megtisztítanunk a ragoktól (stemmatizálnunk). Ez az angol nyelvre nem is

jelent különösebb akadályt, viszont magyar szövegekre ez csak kompromisszumokkal járható út. Az Arcanum szövegei ugyanis nem vegytiszta magyar szövegek, rengeteg régies kifejezés (pl. „fogas vakony" – a földikutya az 1850-es években)[2], régies helyesírású szó („utcza"), változatos helyesírású idegen kifejezés (ugyanannak latinos, görögös, németes, franciás alakjai), ritkább személy- és helynevek, illetve mindezek ragozott alakjai találhatók bennük. (És akkor még nem is szóltunk az eredetiben vagy a digitalizálás során elkövetett melléütésekről).

Summa summarum: egy elemzés során kiderült, hogy az Arcanum által használt szóalakok durván egyharmadát ismeri fel egy magyar szóelemző program. Felvetődik a kérdés, hogy érdemes-e, illetve lehet-e ebben az esetben erre az egyharmadra megígérni a szótöves keresés lehetőségét. A probléma akkor jelentkezik, ha a felhasználó azt hiszi, hogy a *vakony* minden ragozott előfordulását megtalálta, pedig erre a szóra nem áll rendelkezésünkre a szótári alak, következésképpen a ragozott alakok nem fognak erre a keresőkérdésre illeszkedni.

Az ontológiák kérdése azonban még ennél is fogasabb kérdés: a szövegek túlnyomó többségét kitevő bölcsészettudományi jellegű (történeti, néprajzi, irodalmi) alkotásokhoz nem áll rendelkezésre megfelelő mélységű tezaurusz vagy ontológia. A Széchényi-könyvtár Köztaurusza 35 500 kifejezést tartalmaz és számos olyan kifejezés nincs benne, ami még az általános iskolás tananyagnak is szerves része (pl. ispán, jobbágy – igaz van főispán illetve jobbágyság), nem említve a felsőbb iskolák anyagát. Ennek a fogalomkészletnek az alkalmazása szintén nem technikai problémát jelent, a gond ugyanaz, mint a stemmatizált változatok esetében: nem-e a felhasználó becsapása olyat ígérni, amit nem tudunk megbízhatóan teljesíteni.

Néhány éve, egy, a Kossuth-év kapcsán kiadott munka körül vita alakult ki arról, hogy vajon nem Hermann Róbert (a szövegközlő) alkotta-e kossuthiánus stílben a „kompromisszió" szót? Mint kiderült, a szó a kompromittálás korabeli alakja, nem Hermann találmánya, viszont a Kossuth-összes elektronikus változa ezzel kapcsolatban további kérdésekre tud válasszal szolgálni: ezt a szót Kossuth számos helyen használta, többféle alakban (sajnos a mai, kifogásolt formájában soha)[3] – de rajta kívül Jókai és Deák is. Ez tehát tipikus példája azon szavaknak, amelyek nem egyediek de nem is szótározottak, tehát kibújnak a stemmatizált keresés hatálya alól.

Talán nem tartozik szorosan ide, de van még egy nyelvi elem, amit jelenleg is használunk: a karakter-behelyettesítés: az Anacletot az adminisztrációs felületen el lehet látni egy karakterhelyettesítő táblázattal, amibe azokat a karaktereket (és ékezet nélküli párjukat) vesszük fel, amit a felhasználó nem tud a billentyűzetről begépelni („Košice"). Azokat a szavakat, amiben ilyen karakterek vannak karakter nélkül keresve is meg lehet találni.

Relevancia

Az Anacleto a találati sorrendet többfajta módszerrel tudja megrendezni, ezek közül egyik a relevancia-számítás. Mivel a *szemantikus web*-könyv szerzői ennek a témának viszonylag nagy teret szenteltek itt csak néhány tapasztalatról számolnék be. Bár az Arcanum adattára elég nagy kiterjedésű, tematikáját tekintve – és néhány év után – viszonylag átlátható, vagyis elég pontos fogalmaink vannak arról, hogy „emberi ésszel" mi tekinthető néhány esetben relevánsnak; és mi nem. Első próbálkozásaink (az Anacleto alapjául szolgáló Lucene-motor a bevett dokumentum-vektor algoritmust használja), sajnos nem vezettek kielégítő eredményre. Valószínűleg alapos és megbízható nyelvi elemzési eljárások alkalmazása és a különféle predikátumok (mezők) átgondolt súlyozása nélkül nem is lehet jó eredményt elérni, továbbá bizonyosan gátló tényező a nyelvileg is heterogén szövegkorpusz.

De gátolhatja ezt a más körülmény is. Egyetlen kirívó példát idéznék: Németh László Iszony

2 Megjegyzendő, hogy e név már a kortársakban sem váltott ki osztatlan elismerést. A nagy Toldi Ferenc véleménye: «Hiják még ezen állatot „vaktur"-nak is, s ezen nevet még a legjobbnak itélem.» (Vasárnapi Ujság, 1859. január 29.) Azonban fájdalom! a vaktur sem kapott sok bizalmat.

3 Hermann eljárása teljes mértékben megfelelt a vonatkozó szövegkiadási szabályzatoknak.

című regényében a címen kívül pusztán három helyen fordul elő az „iszony" kifejezés, miközben a szót Németh egyéb írásaiban gyakorta használta, s a regényt – erről egy visszaemlékezésben számol be – „az iszony lélektani rajzának" szánta. De a relevanciát számító képlet alapján sem a szó egyedisége (nem egyedi), sem a szövegbeni gyakorisága (más szövegekhez képest ritka) nem teszi lehetővé a jó helyezést.

Összefoglalás

Az Arcanum szempontjából az Anacleto haszna elsősorban az, hogy ezeket a döntően bölcsészeti jellegű szövegeket (szótárak, történeti, irodalmi szövegek, folyóiratok) egyben lehet látni, egységes szempontok szerint lehet keresni (pl. a Kossuthot ábrázoló képeket, vagy a Kossuth írta dokumentumokat). A szövegmennyiségből számos tanulság is adódik pl. az összességében több millió szóalak jelentős részét a mai magyar nyelvi értelmezőprogramok nem ismerik fel, nem beszélve arról, hogy az elérhető magyar ontológiák (pl. OSzK tezaurusz) néhány tízezres fogalomkészlete ennek az adathalmaznak csak töredékét fedik le. A szemantikus web szempontjából pedig a program egy jó kísérleti alap: számos olyan robusztus komponens áll rendelkezésre, amivel a szabványos formákat is kezelni lehet, illetve adott egy jól definiált adathalmaz, ami alkalmas a triviálison túlmenő problémákon való nyelvi kisérletezésekre.