TMT 51. évf. 2004. 5. sz.

Dezsényi Csaba – Varga Péter – Mészáros Tamás – Strausz György – Dobrowiecki Tadeusz
Budapesti Műszaki Egyetem méréstechnika és információs rendszerek tanszék

Tudásalapú információkinyerés: az IKF projekt

Az elektronikusan hozzáférhető hatalmas dokumentumgyűjtemények szövegeinek gépi feldolgozása, információkinyerése rendkívül fontos, de nagyon összetett probléma. A könyvtártudomány hagyományos módszereit kiegészítve ezen a téren a tudásalapú megoldások hozhatnak áttörést. Egy konkrét projekt bemutatásával ezt az új területet tekintjük át.

Rohanó világunk legfontosabb értéke a gyors és pontos információ, illetve az ezzel koherensen megalkotott tudás. Ehhez az internet mint információs média megfelelő alap, hiszen nagy mennyiségű információ folyamatosan hozzáférhető bárki számára. Azonban az óriási, heterogén és elosztott információs közegben nem könnyű feladat megtalálni egy-egy igényelt dokumentumot, és főképp nem könnyű egy-egy igényelt információdarabkát kibányászni belőle, amelyhez esetleg több forrás több részletét kell koherens módon megvizsgálnunk és elemeznünk. A hatékony megoldás támogatására számos szoftver jelent meg az elmúlt években, amelyek segítségével részben vagy teljesen automatizálni lehet bizonyos információkeresési és -kezelési folyamatokat. Ezek részben sikeresek, ám közel sem elegendőek ahhoz, hogy integrált intelligens információs és tudásmenedzsment-környezetet biztosítsanak egy-egy alkalmazás számára. A *BME méréstechnika és információs rendszerek tanszékén* folyó *IKF kutatási és fejlesztési projekt* egy komplett tudásalapú információkinyerő rendszer megalkotását tűzte ki célul, amely korszerű tudásintenzív technológiák segítségével képes emberi felhasználásra szánt információt feldolgozni. Jelen tanulmány a projekt célkitűzéseinek, a rendszer elméleti és technológiai felépítésének és néhány – a folyóirat témakörét érintő – innovatív megoldásnak a rövid áttekintése. Habár a téma folyóiratbeli viszonylagos újszerűsége miatt inkább a technológiai irányzatok bevezető jellegű leírásával adna átfogóbb képet, mi a projekt keretében megvalósított konkrét alkalmazással szeretnénk betekintést nyújtani a tudásalapú információfeldolgozás és tudásábrázolás témaköreibe.

Mi a tudás?

A hagyományos döntéstámogató rendszerek stratégiai szerepe az utóbbi években jelentős fejlődésen ment keresztül [1]. Ennek oka az internet elérhetőségének a kiszélesedése, ennek következtében a hozzáférhető információforrások ugrásszerűen megnövekedett típusválasztéka és száma. Az integráció növekvő mértéke (az adattárházat, az adatbányászatot, és egyéb hasonló technológiákat is beleértve) a döntéstámogató rendszerek olyan fejlődéséhez vezet, amely képes hasznosítani a különböző (külső és belső) forrásokból származó és különböző típusú – akár strukturált, akár strukturálatlan – adatokat. Így a döntéstámogató rendszerek legújabb generációja teljesebb funkcionalitást kínál, és felhasználóit versenyképesebb információkhoz, előnyhöz juttatja.

A következő néhány évben a Tudás Kinyerés, Tudás Menedzsment (TK, TM) és ezekkel rokon technológiák egyre nagyobb jelentőséghez jutnak, mivel az elérhető információforrások minél teljesebb ellenőrzését, és azok lehető legjobb kiaknázását célozzák meg. A tudásmenedzsment rendszerek a technológiák széles körét használják fel a dokumentummenedzsmenttől a szöveg- és adatfeldolgozáson át a megjelenítésig. Alapvető céljuk az üzleti folyamatok támogatása. A „tudás" és „intelligencia" kifejezések alkalmazása e rendszerek elnevezésében azonban jelenleg sokkal inkább a marketing által megkívánt fogalom, mint e rendszerek belső felépítéséből és képességeiből fakadó tulajdonság kifejezése. E rendszerek általában dokumentum- és adatmenedzsment, elemzési és riportgenerálási, szövegkereső, illetve adatbá-

nyászeszközök, melyek nem (vagy csak elvétve) tartalmaznak valódi tudásábrázolási mechanizmusokat. A „tudás" szót sokkal inkább „információ" jelentéssel használják, egy kereskedelmi tudásmenedzsment rendszer pedig inkább az emberek fejében lévő tudás menedzselésének a támogatását célozza meg. Dokumentumtárolásra és -elérésre példaként említhetnénk a Lotus Domino, az OpenText vagy a Filenet rendszereket. Az információhoz való hozzáférést könnyítő kereső, illetve portál rendszereket gyárt az IBM/Lotus (Raven), Fulcrum, Verity, Excalibur, illetve Autonomy.[1] Adatelemzésre és adatbányászatra alkalmas rendszereket gyártanak a nagyobb adatbázis-kezelő rendszerek fejlesztői.

Ezzel szemben a „tudás", „tudásalapú" és rokon szakkifejezések valódi információtechnológiai jelentése mást takar, ezért rendkívül fontos tisztázni a témakör kulcsfogalmainak pontos értelmezését. És mivel a legjobb építkezési mód az, ha az alapokat tesszük le először, mi is az elemi építőkockával, az adattal kezdjük a definíciót, és jutunk el egészen a tudás fogalmáig.

Az adattól az információn át a tudásig

*Adat*nak tekintünk általában mindent, amit információs rendszerekben fogadhatunk, tárolhatunk, illetve feldolgozhatunk. Önmagában a jelentése azonban nem több, mint a reprezentálására szolgáló szimbólum. Az *információ* ezzel szemben olyan adat, amelynek a jelentése túlmutat az őt ábrázoló szimbólumon, amivel a felhasználó információs igényét kielégíti egy probléma megoldásában. Egy konkrét információ értelmezését az adott feladat és felhasználó kontextusában tudjuk megadni, tehát egy adatelemnek többféle információs vetülete lehetséges, amit az aktuális felhasználás feltételei szabnak meg.

*Tudás*on a valóság egy darabjára vonatkozó információk koherens halmazát értjük. Ez egy adott probléma megoldásához szükséges összes olyan információt jelenti, amely a problémával kapcsolatos általános ismereteinket koherens módon írja le, tartalmazza a problémában adott jelenségek (rendszerek) viselkedését, belső felépítését stb. Míg az információ egy önmagában statikus ismeretanyag, tudás alatt (az ismeret mellett) a hozzá kapcsolódó intelligens cselekvési képességeket is feltételezzük. Egy tudásalapú informatikai rendszer így többet jelent egy hagyományos információtárnál, hiszen képes a meglévő információ és tudás segítségével intelligens és automatizált cselekvések elvégzésére.

Adatot keresni és megtalálni könnyű feladat lehet, legalábbis az elméleti problémák felől megközelítve. Erre számos kész és jól működő rendszer létezik manapság, kezdve az egyszerű adatbázis-kezelő rendszerektől egészen a komplex adattárházakig és különböző adatbányászati módszerekig. Egy ember által igényelt információ megtalálása már jóval összetettebb feladat. Míg az adatbázis-kezelő rendszerekben végzett keresés esetében a keresett információ egy konkrét adat, determinisztikus módszerrel előállítható egy teljesen specifikált lekérdező nyelv segítségével, addig az információkeresés esetében a keresett információ csak valószínűségi relációba hozható a tárolt dokumentumok egy halmazával. Nem véletlen tehát, hogy információkeresés és -kezelés tekintetében a mai napig óriási erőfeszítések folynak mind a kutatások, mind a technológiai fejlesztések terén.

Ezek után könnyű elképzelnünk, milyen nehézségekbe ütközünk, ha egy adott témával kapcsolatban az emberek számára értelmezhető és felhasználható tudást szeretnénk kinyerni a rendelkezésre álló információs forrásokból, és ennek segítségével egy koherens, gépileg is feldolgozható tudásbázist szeretnénk létrehozni. Nem titok, hogy az ilyen rendszerek még igencsak gyerekcipőben járnak, azonban a jövő mindenképpen ebbe az irányba mutat, rengeteg kutatás folyik, és ami a legfontosabb: óriási igény van rá mind a tudományos, mind az üzleti világ oldaláról.

A BME méréstechnika tanszék egy konkrét projekt keretében tűzte ki célul az előzőekben felvázolt, ígéretes témakörben történő kutatási és fejlesztési munkát. A következőkben a projektet és legfontosabb célkitűzéseit mutatjuk be röviden.

Az IKF projekt

A bemutatandó információelemzési és -kinyerési technikák, illetve az elkészült, tudásalapú információkinyerő rendszer fejlesztése az „Információ és Tudás Tárház" (Information and Knowledge Fusion = IKF) kutatási és fejlesztési projekt[2] keretében zajlik. A projekt része az Information and Knowledge Fusion EUREKA Applied Research Project-nek[3] [2]. A nemzetközi konzorcium fő célkitűzései újszerű Intelligens Tudástárház Környezetek (Intelligent Knowledge Warehousing) elemzése és

kifejlesztése, amely lehetővé teszi a korszerű Tudás Menedzsment és Üzleti Intelligencia (Knowledge Management and Business Intelligence) szolgáltatások megvalósítását. A nemzetközi projekt keretében a partnerek különböző alkalmazási területekre készítenek önálló IKF rendszereket. A magyar konzorcium tagjai az ML Tanácsadó és Informatikai Kft., a MorphoLogic Kft. és a BME méréstechnika és információs rendszerek tanszék.

Célkitűzések

A jelenleg elérhető kereskedelmi rendszerek több funkciója felhasználható egy intelligens rendszer kialakításához, de valódi tudásintenzív megoldások hiányában nem képesek teljes megoldást adni. A magyar Információ és Tudás Tárház projekt célja egy komplett tudásalapú döntéstámogató rendszer kidolgozása és kifejlesztése pénzügyi cégek és bankok részére. A rendszer fő tevékenysége az információ témaspecifikus, különböző típusú forrásokból (internet, intranet erőforrások, adattárházak stb.) történő keresése, és az információ strukturált szolgáltatása a felhasználóknak. A rendszer emelt szintű szolgáltatásokat nyújt a hazai felhasználók számára azáltal, hogy:

- az információszolgáltatás és -keresés folyamatát az információgyűjtés tárgyáról, forrásairól és felhasználójáról meglévő ismereteinket tároló tudásalapú modell felhasználásával vezérli; az információszolgáltatást egy jól definiált, hatékonyan modellezhető, szűk tárgyterületen végzi el;
- az információszolgáltatást a beépített modellek által automatikusan vezérelt tudásgyűjtéssel felállított és folyamatosan karbantartott tudástár alapján biztosítja;
- a strukturálatlan és részben strukturált szöveges információk feldolgozását a tárgyterület ontológiájának létrehozásával és alkalmazásával végzi el;
- a hazai információforrások elemzését jelenleg is alkalmazott magyar nyelvi elemző eszközök a rendszer céljaira továbbfejlesztett változatával támogatja.

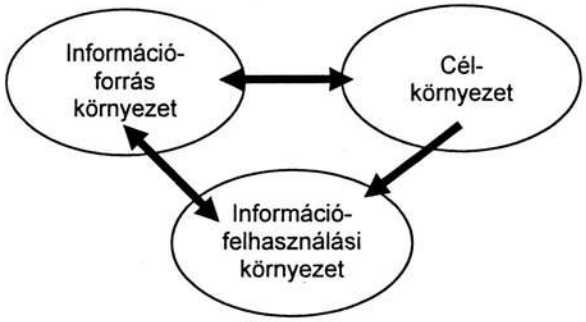
A projekt keretében kifejlesztendő prototípus rendszer és mintaalkalmazás célja pénzügyi cégek ügyfeleinek folyamatos monitorozása, és információszolgáltatás biztosítása a döntéshozatali folyamatok (pl. hitelkérelem elbírálása, ügyfélminősítés) támogatásához. A rendszer felhasználja és kiegészíti az elérhető, hatékony információkereső, -tároló és -feldolgozó szoftver- és hardvereszközöket, szabványokat.

Az IKF rendszer

A továbbiakban a projekt eddigi szakaszában létrejött IKF keretrendszert ismertetjük (erről részletesebben lásd [3] és [4]). Először a rendszer környezetét és magas szintű felépítését mutatjuk be, majd egyes fontosabb, innovatív szolgáltatásokat és a hozzájuk kapcsolódó elméleti és technológiai hátteret fogjuk részletesebben ismertetni. Ezen elméleti bevezetők és gyakorlati megvalósítások tárgyalásával szeretnénk bemutatni az információkinyerés és tudásábrázolás témakörök alapjait.

A rendszer környezetmodellje

Egy általunk elképzelt tudásintenzív információmenedzsment rendszerhez három különböző környezet kapcsolódik *(1. ábra)*. Ez a környezetmodell – mint később látni fogjuk – meghatározza a rendszer absztrakt felépítését is.



1. ábra **IKF környezetmodell**

A *célkörnyezet* a témához kapcsolódó tudás fizikai forrása, a valós világ objektumait tartalmazza: fogalmakat, eseményeket stb., illetve ezek közötti relációkat és összefüggéseket. A rendszer intelligens működéséhez szükséges háttértudás, tudásmodell a célkörnyezet elemzésével és modellezésével jöhet létre.

Az *információforrás környezetben* találhatók azok a dokumentumok, szöveges anyagok, amelyek egyrészt tükrözik a célkörnyezet tárgyát, másrészt tartalmazzák a szükséges információt a rendszer számára, és hozzáférhetőek digitális úton. Elsődleges forrásként az internetet nevezhetjük meg, amelynek nagy hátránya, hogy a dokumentumok tipikusan strukturálatlan, emberi felhasználásra szánt formában állnak rendelkezésre, illetve (ahogy a bevezető fejezetben már utaltunk rá) a heterogén, elosztott „dokumentumrengeteg" mélyéből igen nehéz kiszűrni a számunkra fontos információdarabkát. Ezenkívül természetesen megnevezhetünk más, strukturált forrásokat is, mint

például publikus adatbázisok, adattárházak. Egy fontos jellemzője még a forráskörnyezetnek, hogy a célkörnyezet által leírt információ, tudás csak erős hiányokkal, időben és térben is elszórtan jelenik meg, ami külön megnehezíti beszerzésüket és értelmezésüket.

Az *információfelhasználási környezetben* helyezkednek el azok a felhasználók (pl. banki menedzsment, személyzet), akik bizonyos tudást akarnak beszerezni a célkörnyezetről, hogy céljaikat elérjék. Ezt a forráskörnyezetből tudják kinyerni a közvetítő tudásmenedzsment rendszer segítségével.

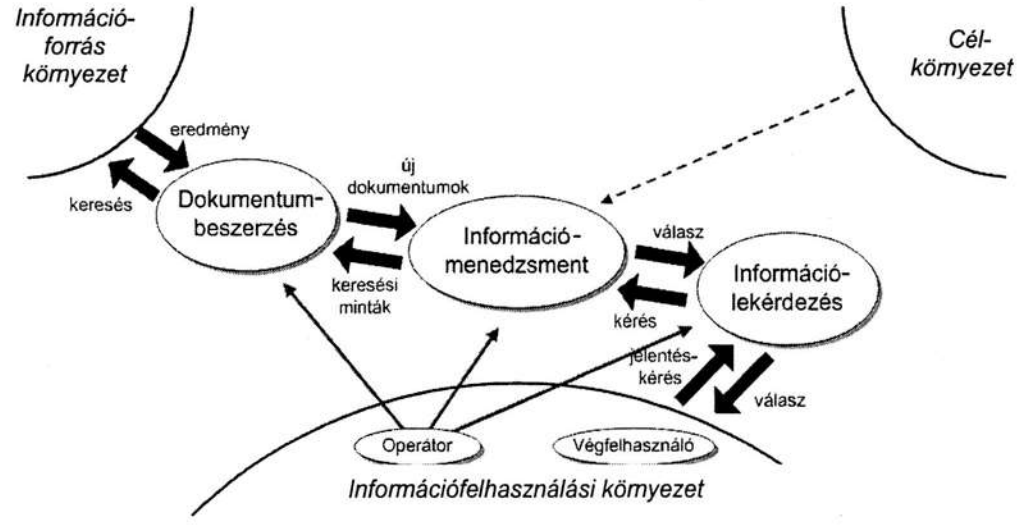
A rendszer magas szintű felépítése

Egy tudásalapú információkereső és -elemző rendszer általunk ajánlott magas szintű felépítése három fő komponensből áll: dokumentumbeszerzés, információmenedzsment, illetve információlekérdezés *(2. ábra)*.

*Dokumentumbeszerzés*en azt a tevékenységet értjük, amely során a rendszer beszerzi a forráskörnyezetről az információkinyeréshez szükséges forrásdokumentumokat. Feladata az összes, a rendszer számára hasznos (releváns) dokumentum felkutatása, letöltése és előelemzése. Ezt a rendelkezésre álló háttértudás, illetve különböző információkeresési és -kinyerési eszközök segítségével teszi meg. A háttértudás részei a menedzsmentmodultól kapott ún. keresési minták, amelyek a releváns dokumentumok kereséséhez szükséges tárgyterület-specifikus tudást írják le. A beszerző rendszer a megtalált és letöltött dokumentumokat elemzés után megfelelő strukturált, belső formára konvertálja (amely így tartalmazza az eredeti forráson kívül az összes kinyert információt is), majd továbbítja a menedzsernek.

Az *információmenedzsment* feladata, hogy a beszerzett és elemzett dokumentumokból az igényelt információt kinyerje, és a rendszer tudásbázisában tárolja gépileg értelmezhető, strukturált formában. Az így kialakított koherens tudástár segítségével válaszol a rendszer a beérkező kérdésekre, amelyek az információlekérdező modul felől érkeznek. A menedzsmentmodul közvetlenül egyik környezettel sincs kapcsolatban, azonban a célkörnyezet modelljét, azaz a rendszerben előzetesen létrehozott témaspecifikus háttértudást tartalmazza. Elsősorban nyelvi elemző módszerek (NLP) és tudásintenzív feldolgozás (ontológia és logika) segítségével valósítja meg a megfelelő témaspecifikus információ- és tudástár építését.

Az *információlekérdező* rendszer feladata az információfelhasználási környezettel való kapcsolattartás, azaz a felhasználói kérések, parancsok értelmezése, és azok továbbítása a menedzsmentmodulnak, majd az onnan visszakapott információ rendezett, átlátható formában történő visszaadása. Lehetőséget teremt a rendszerben lévő háttértárak (dokumentumtár, tudásbázis) böngészésére, visszakeresésre, bizonyos felhasználói lekérdezések megválaszolására, illetve előre definiált riportok automatikus generálására. Legfontosabb eleme a felhasználói interfész, amelynek jól áttekinthető hozzáférést kell nyújtania a kinyert információhoz.



2. ábra **Az IKF absztrakt architektúra és meghatározó információs folyama**

A teljes rendszer nagy szabadságfokú, tetszőleges tárgyterületre konfigurálható, és számos paraméter segítségével hangolható. Ezért külön hangsúlyt kapnak a különböző segédprogramok, grafikus felületek és eszközök, amelyek a konfigurálásban támogatják a rendszer operátorait. Ez feltétlenül szükséges, hogy hatékonyan és rugalmasan lehessen alkalmazni egy ilyen nagy komplexitású eszközt.

Az IKF rendszerben számos magas szintű szolgáltatás (modul szinten) kap szerepet, melyeknek szoros és konzisztens együttműködése szükséges a teljes feladat hatékony megoldásához. A különböző szolgáltatások típusaik szerint is csoportosíthatóak, mint például információkinyerő funkciók, tudásintenzív elemzők, tudásmodellezés, háttértár menedzsment szolgáltatások, felhasználói felületek stb. Ezek részletes ismertetésétől eltekintünk. A következőkben a tanulmány témáját érintő szolgáltatások és megoldások főbb jellemzőit, illetve a hozzájuk kapcsolódó elméleti háttereket mutatjuk be.

Dokumentumbeszerzés és elemzés

Az felhasználók által igényelt tudás a forráskörnyezetben lévő információforrásokban lelhető fel, de sajnos több nehézséggel is meg kell küzdeni, hogy a szükséges források gépileg értelmezhető formában rendelkezésre álljanak a tudásbázis felépítéséhez. Mivel a forráskörnyezet elsősorban az internet, az ebből fakadó buktatók ismertek: a megfelelő releváns dokumentumokat (amelyek hasznos információt tartalmaznak az igényelt tudásbázis építéséhez) először is meg kell találni, ami önmagában is nehéz feladat. Mivel az interneten lévő dokumentumok zömét emberi olvasásra, nem gépi feldolgozásra szánták, a következő lépcső a szükséges információ azonosítása és kinyerése a természetes nyelvű dokumentumokból. Ennél a lépésnél a strukturálatlan, csupán vizuális megjelenítésre formázott forrásdokumentumokat gépileg is értelmezhető, logikai (szemantikai) struktúrákba kell önteni. Az így átalakított források már alkalmasak a tudásbázis automatizált építéséhez, amely az információmenedzsment modul feladata lesz.

Az előzőekben említett két fő feladat két nagy elméleti témakörrel hozható kapcsolatba. Az első feladat az *információkeresés* (Information Retrieval = IR) témakörébe tartozik [5], amely releváns dokumentumok kollekcióban történő keresésével foglalkozik. A második problémát az *információkinyerés* (Information Extraction = IE) témaköre fedi le [6], amelynek célja a szöveges dokumentumokból történő információkinyerés megoldása. Mindkét elméleti témakör igen fontosnak számít a manapság nagy intenzitással folyó információs kutatások és fejlesztések terén, azonban ezek rövid ismertetése is meghaladja a jelenlegi tanulmány kereteit.

Webforrás modellezése

Mint említettük, az interneten található dokumentumok többsége emberi olvasásra szánt, csak vizuális megjelenítés céljára van strukturálva. Az oldalak általában HTML[4] formátumúak, amelyben olyan strukturális elemeket találhatunk, mint „bekezdés", „dőlt betű", „felsorolás" stb. A gépi feldolgozáshoz azonban nekünk olyasféle szemantikai strukturáltság kellene, mint például „cégleírás", „igazgató telefonszáma", „konkurens cég neve", és még sorolhatnánk különféléket az alkalmazástól függően. Habár a természetes nyelvű leírást és a vizuális jelölések szemantikai jelentését a szoftver értelmezni nem, vagy csak erősen korlátozva tudja, egy fontos tulajdonságot ki lehet használni: valamilyen szempontból összetartozó, hasonló dokumentumok esetén bizonyos logikai struktúrák ugyanolyan vagy hasonló vizuális struktúrával azonosíthatók. Egy webes hírportál cikkei például nagyjából ugyanolyanok, így a megfelelő logikai elemeket (szerző, dátum, cikkhasáb stb.) egy szoftver be tudja azonosítani az összes cikkben, miután valahogy leírtuk, hogyan találja meg. Összetettebb feladat a szoftver számára leírni általánosabb strukturális elemeket, amelyek már csak néhány jellegzetességükben hasonlítanak. Erre példa lehet személyek honlapjain lévő publikációs listák felismerése és kinyerése.

A webcsomagolók (webwrapper, webforrás-modellező) olyan speciális szoftvereszközök, amelyek a körülírt probléma megoldását célozzák meg [7]. Segítségükkel ismert struktúrájú internetes oldalakról automatikusan tudunk információt kinyerni, és megadott logikai formára konvertálni. A megfelelő szövegrészek kinyeréséhez szükségesek az ún. *forrásmodellek*, amelyek leírják, hogy a hasonló struktúrájú dokumentumokban hol találhatóak meg az igényelt részek. A modell leírása (modellező nyelvtan) tulajdonképpen hasonló dokumentumok strukturális jellemzőit próbálja megragadni, és ennek segítségével a releváns információt tartalmazó szöveges részeket azonosítani a kinyerés-

hez. Egy webcsomagoló szoftver a következő fontos tulajdonságokkal jellemezhető:

- Modellgenerálás: az a módszer, ahogy a különböző forrásokhoz a felhasználó a megfelelő forrásmodelleket elkészíti.
- Struktúra feldolgozása: a dokumentumok strukturális jellemzőinek feldolgozási módja, maga a modellező nyelv jellege. Ez meghatározza az eszköz által kezelhető strukturális elemek fajtáit, ezzel pedig a kinyerhető információtípusok skáláját.
- Kimeneti formátum: a kimeneti adatobjektumok formátuma az információ kinyerése után.

Az elmúlt években több kutatási projekt és szoftverfejlesztés irányult hatékony webforrás-modellező eszközök létrehozására. Ezek az eszközök különböző módszereken és technológiákon alapulnak, úgymint deklaratív vagy procedurális nyelvek, HTML struktúra elemzése, természetes nyelvű feldolgozás, gépi tanulás és adatobjektum-modellezés [8]. E szoftverek mind elsődlegesen a legegyszerűbb modellgenerálásra koncentrálnak, hogy egy átlagos felhasználó minél könnyebben tudjon megfelelő leírást készíteni forrásoldalakhoz. Ez alapján nagyjából két csoportba sorolhatjuk őket:

- Gépi tanulás alapú: a felhasználó néhány forrásoldalon „kézzel" bejelöli a számára igényelt adatrészeket, ezek alapján a program létrehozza (kikövetkezteti) a forrásmodellt, amit alkalmazni lehet hasonló felépítésű oldalakra az információkinyeréshez, pl. [9, 10].
- Leírónyelv alapú: a felhasználó közvetlenül a szoftver leírónyelvét használja fel a forrásmodellek elkészítéséhez. Itt általában a minél egyszerűbb nyelv és hozzá tartozó szerkesztőprogram kialakítása a cél, pl. [11, 12].

Mindkét csoportba tartozó eszközöknek megvannak az előnyeik és hátrányaik, azonban az összes eddig készült szoftvernek van néhány erősen hátrányos tulajdonsága. Elsődlegesen a modellgenerálás egyszerűségére törekszenek (elhanyagolva általános strukturális elemek széles skálájának feldolgozhatóságát). Ebből adódóan, és a megoldandó probléma komplexitása miatt tipikusan csak adatcentrikus forrásokat (pl. táblázatos jellegű, nagymértékben hasonló portáloldalak) vagy egyéb, a szoftvertől függő specifikus strukturális elemeket (mintákkal definiálható adatobjektumok – dátum, pénznem stb.) tudnak kezelni. Az ismeretlen vagy változó információforrások feldolgozását sem tudják megoldani. Annak ellenére, hogy a webcsomagoló szoftvereknél fontos szempont, hogy általánosan használható eszköz szülessen, még mindig heterogén a kínálat ezen a téren, minden megoldás specializált valamilyen szempontból.

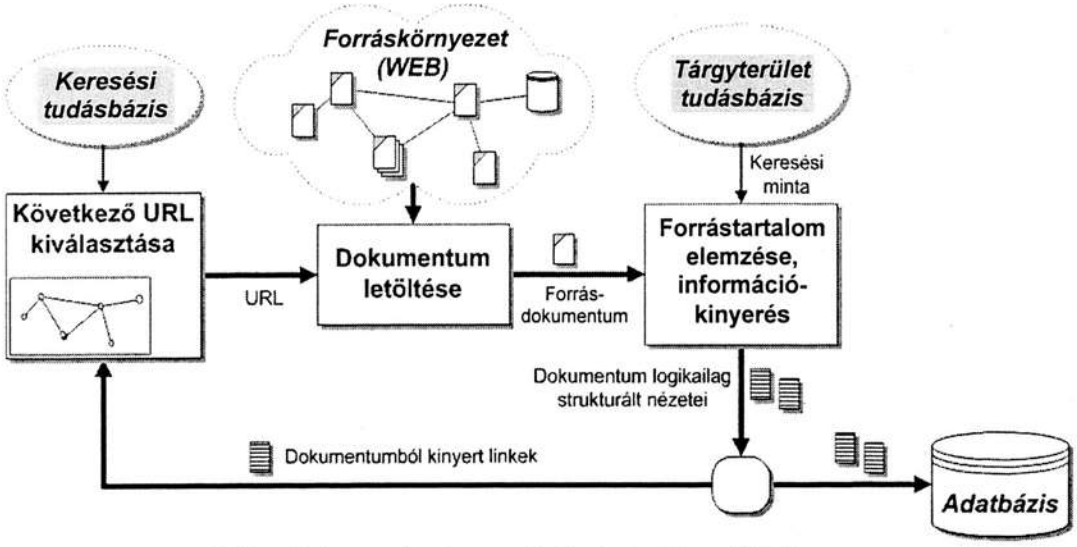
Az XML technológia

Míg az interneten található, vizuális megjelenítésre szánt dokumentumok kiválóan leírhatóak a HTML jelölőnyelv segítségével, az automatizált, gépi feldolgozáshoz más leírónyelvre van szükségünk, amelynek segítségével a tetszőleges logikai dokumentum struktúrája kialakítható. Ennek a megoldására fejlesztették ki az XML nyelvet (Extensible Markup Language = kiterjesztett jelölőnyelv) [13], amiért is rendkívül fontos szerepet tölt be az információ- és tudásmenedzsment területén belül.

Az XML egy dokumentum-jelölőnyelv, a W3C[5] konzorcium fejlesztéseként jött létre a HTML és SGML[6] nyelvek utódjaként. Segítségével dokumentumok strukturált leírása valósítható meg. Az XML tulajdonképpen olyan nyelv (ún. metanyelv), amelynek segítségével tetszőleges leírónyelvet tudunk definiálni (pl. az XHTML, amely XML alapú HTML), azaz nincsen előre rögzített elem- vagy struktúrakészlete, ez az adott alkalmazástól, dokumentumtípustól függ. Viszont azt előírja, hogy a struktúra hogyan épülhet fel, melyek az egyes szabályok a leírás helyességére vonatkozóan; számos szabványos és rendkívül hasznos eszközzel rendelkezik, amelyek XML dokumentumok feldolgozását támogatják.

Bár XML-lel tetszőleges jelölő nyelvtan létrehozható, mégis legtöbbször egy XML formátumú dokumentum nem tartalmaz megjelenítésre vonatkozó információt (mint például az XHTML-ben, ami kivétel), sokkal inkább a dokumentumok tartalmi leírását célozzák meg, vagyis az egyes logikai egységeket, amelyek segítségével felépül egy dokumentum. Ezzel elérhető, hogy az adatok, információk és dokumentumok önleíróak legyenek (nem pedig önformázóak) annak érdekében, hogy a különböző szoftveralkalmazások értelmezni tudják őket, ne csupán emberi olvasásra legyenek alkalmasak. Egy XML nyelven, tartalmilag strukturált dokumentum automatizált feldolgozása jóval egyszerűbb feladat, mint pl. egy HTML oldalé, mivel az egyes szövegelemek az információtartalom alapján vannak megjelölve.

Az XML nyelv szimbólumkészletét tekintve nagymértékben hasonlít az ismert HTML-re, bár a strukturális felépítés szabályai valamivel szigorúbbak, aminek viszont a következménye, hogy egy XML állományt igen egyszerű használni és feldolgozni.



3. ábra **Dokumentumbeszerzés funkcionális működése**

Egy XML dokumentum egyértelműen leképezhető egy fastruktúrába, mivel az egyes elemek (ún. tagek) nem lapolódhatnak át, csak a teljes tartalmazás megengedett (szemben a HTML-lel). Egy adott XML alkalmazás (azaz XML-lel definiált dokumentum-jelölőnyelv) elemeinek neveit, illetve a strukturális felépítés szabályait az ún. DTD-vel[7] (Document Type Declaration = dokumentumtípus-deklaráció) tudjuk rögzíteni. Segítségével ellenőrizni és érvényesíteni (validálni) tudjuk egy megszerkesztett dokumentum helyes felépítését.

Az XML hasznos szabványos eszköze az XSLT[8] (XML Style Sheet Transformation), amely különböző XML struktúrák közötti transzformációt valósít meg. Olyan mechanizmust ír le, amely segítségével egy adott DTD-vel rendelkező forrás XML dokumentumot egy másik DTD-vel rendelkező formára tudunk hozni. Az XSLT képes olyan műveletek elvégzésére, mint elemeket törölni, létrehozni, átsorolni, átnevezni és sorba rendezni, előtagokkal és utótagokkal kiegészíteni a tartalmat stb. Az átalakítás a megadott mintaillesztő szabályoknak (template) megfelelően történik. A forrásdokumentumban szereplő elemeket a feldolgozó bizonyos útvonal-kifejezések segítségével (aminek a formáját az XPath[9] szabvány rögzíti) összehasonlítja a mintákkal, ahol azok illeszkednek, ott végre lehet hajtani a kimeneti dokumentumra vonatkozó utasításokat.

Az IKF dokumentumbeszerző rendszer

A dokumentumbeszerzés feladata a megfelelő forrásdokumentumok megkeresése, és ezek átalakítása tartalmilag strukturált formára, amivel már az információmenedzsment rendszerben a tényleges tudáskinyerés és tudásbázis-építés megvalósulhat. Az IKF rendszerben ezt a feladatot egy autonóm ágens látja el (az ágens technológiáról bővebben lásd: [14]), ún. webrobot, amely az internetet bejárva kutat releváns dokumentumok után [3], [15]. A rendszer vázlatos működési mechanizmusa a *3. ábrán* látható.

Az intelligens viselkedést támogató háttértudás két részre bontható: a *tárgyterület tudásbázis* az éppen aktuális, alkalmazástól függő témaspecifikus háttértudás, amely nagyrészt a keresési minták formájában érkezik az információmenedzsertől. Ez az elemzésre vonatkozó információt tartalmaz, például kulcsszólistákat statisztikai relevancia vizsgálathoz, vagy forrásmodelleket dokumentumok strukturális elemzéséhez és információkinyeréshez.

A *keresési tudásbázis* előre rögzített tudást tartalmaz. Ez a keresés általános módszertanát írja le, vagyis azt, hogy milyen eszközökkel és hogyan érdemes a weben adott témájú dokumentumok után kutatni. Ezek lehetnek például algoritmusok a hatékony URL-választási mechanizmushoz, általános internetes keresők használatának módszerei és szükséges paraméterei stb.

A rendszer nagy vonalakban a következőképpen működik: első lépésként ki kell választani annak a forrásnak a címét (URL[10]-jét), amelyről a dokumentumot szeretnénk letölteni és elemezni. Hogy a választás hatékony legyen, azaz ne véletlenszerűen vizsgáljunk meg az interneten egy dokumentumot, szükség van bizonyos háttértudásra. Ennek

egy része a már megismert keresési tudásbázis, de ezenkívül hasznos felhasználni a megelőző keresések eredményeit is, mint például a HTML oldalakról kinyert linkeket, melyik oldal volt releváns stb. Ennek a támogatására az ágens működés közben a forráskörnyezetről épít egy belső gráf alapú modellt. Ezzel megvalósulhat, hogy a webrobot ne csak közvetlen környezetét érzékelje lokálisan, hanem globális képe legyen a már megismert forráskörnyezetről. A belső modell segítségével hatékony gráf alapú algoritmusok implementálhatóak, amelyek az URL-kiválasztási mechanizmust vezérlik.

A kiválasztott URL-en lévő dokumentum letöltése után a következő lépés a forrás tartalmi elemzése, a bejövő dokumentumok logikai struktúrájának felismerése. Az elemző a forrásdokumentum bizonyos tartalmi *nézeteit* állítja elő, amelyek az elemzést követően strukturált formában fogják tartalmazni a különböző típusú kinyert információrészleteket (részletesebben lásd a következő alfejezetben). Egy-egy ilyen nézet az eredeti dokumentum bizonyos információs vetületének feleltethető meg, szemantikailag strukturált formára alakítva. A nézetek tipikusan kinyert szövegrészleteket foglalnak magukba, azonban ezek a töredék szövegek tartalmazzák az alkalmazás számára lényeges információt, amelyen majd a tudásintenzív elemzők dolgoznak. A bejövő dokumentumokon ezenkívül hagyományos statisztikai szövegelemzésre is sor kerül, a létrejövő index és statisztikai relevancia információ a nézetekhez lesz csatolva.

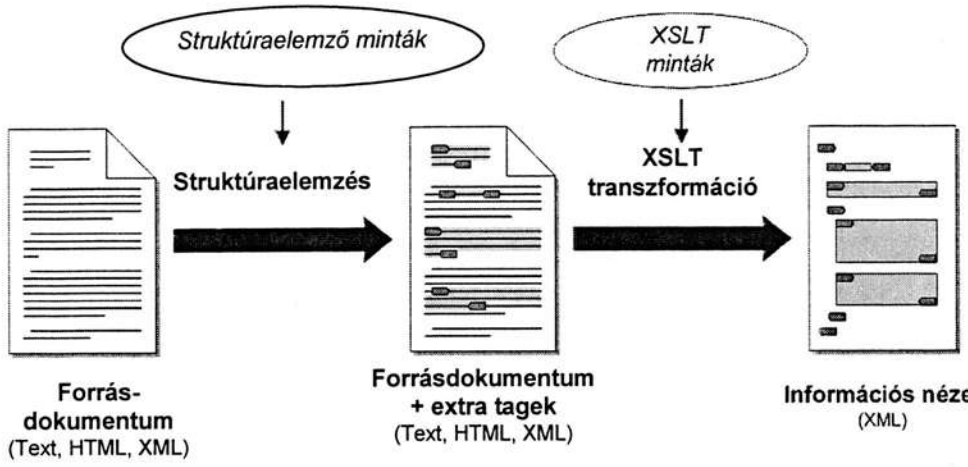
A létrejött nézetek egy részére az URL-kiválasztási mechanizmusnak is szüksége van (visszacsatolás), hiszen ezzel tovább tudja építeni a belső forráskörnyezet-modellt, és információt szerez a további sikeres kereséshez. Végül a teljes dokumentum a létrehozott nézetekkel együtt a rendszerben lévő dokumentumtárba kerül, ahol a további IKF modulok hozzáférhetnek.

Forrásdokumentumok strukturális elemzése

A beszerző rendszer a keresés során letöltött forrásdokumentumokat elemzi, és releváns információt próbál kinyerni belőlük. A kinyert információt egy vagy több kimeneti XML állományba, a már röviden ismertetett nézetekbe konvertálja. Egy ilyen nézet hordozza a forrásdokumentumból kinyert információ egy meghatározott részletét, az eredeti tartalom bizonyos „vetületét" strukturált formában. Két fontos jellemzője van: a *típusa*, amely meghatározza, hogy milyenfajta információt tartalmaz (pl. egy egyszerű nézet tartalmazhatja a HTML oldalból kinyert linkeket, egy összetettebb pedig az oldalon előforduló cégneveket és elérhetőségeket). A másik a rögzített *struktúrája*, amely leírja a benne lévő típusos információ felépítését. Mivel a nézet XML formátumú, ezért a struktúráját DTD-vel tudjuk definiálni. Tetszőleges nézettípust és hozzá tartozó DTD-t definiálhatunk az IKF rendszerben az alkalmazási területtől függően.

A forrásdokumentum tartalmi elemzése során a hagyományos indexelés és statisztikai relevancia analízis mellett helyet kapott egy forrásmodell alapú struktúraelemző eszköz (webcsomagoló) is, amely a megfelelő XML nézeteket hozza létre. A megközelítés azonban különbözik az eddigiektől, a hagyományos webcsomagolóktól (lásd a „Webforrás modellezése" c. fejezetben). Mi – az egyes módszerek és a szoftver tervezésekor – elsődlegesen a strukturális feldolgozásra koncentráltunk. A fő szempont egy olyan általános és kellőképpen rugalmas eszköz létrehozása volt, amely a forrásdokumentumokban fellelhető strukturális elemek lehető legszélesebb skáláját tudja kezelni, az egészen általánostól kezdve a teljesen specializáltig bezárólag. Egy olyan leírónyelv és hozzá tartozó elemzési technika fejlesztése a cél, amely bár komplexitását tekintve felülmúlhatja az eddigieket, alapja lehet egy olyan rendszernek, amely segítségével a forrásdokumentumok (akár ismeretlen, akár előre ismert) tetszőleges strukturális és egyéb jellemzői jól kezelhetőek.

Ezek alapján a forrásból egy bizonyos típusú információ kinyerése és a megfelelő XML nézet előállítása két fázisban történik *(4. ábra)*. Az első fázisban a forrásdokumentum szignifikáns részleteit jelölik meg. Ezt egy XML alapon működő elemző végzi, amely az eredeti szövegben a számunkra fontos részeket megfelelő XML címkékkel látja el. Ezt a műveletet az ún. *struktúraelemző illesztési minták* vezérlik. Ez tulajdonképpen a forrásmodellek leírónyelve, amelynek segítségével a dokumentumokban lévő strukturális sajátosságokat tudjuk megragadni. A leírónyelv szemantikája, illetve a mintaillesztés működésének alapjai egy speciális technikával lettek megoldva, melyben paraméterekkel ellátott, reguláris kifejezés[11] alapú mintaelemek sorozatos illesztésével tudja az elemző meghatározni a leírt részek helyét a dokumentumokban. Ezenkívül külső, speciális elemző modulok is beilleszthetők, amivel egészen speciális heurisztikákat is el lehet készíteni. A nyelv sza-

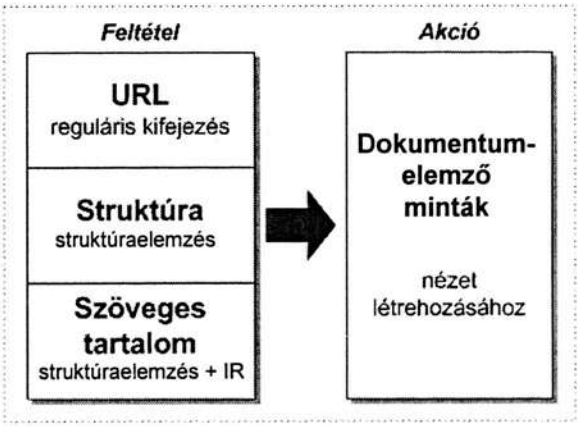


4. ábra **Dokumentum strukturális elemzése**

badságfoka elég nagy, így sokféle strukturális felépítés leírható. Ennek megfelelően viszont kissé komplexnek tűnhet, azonban a feltevésünk az, hogy ezeket a mintákat nem általános „desktop" felhasználók, hanem szakértő operátorok fogják létrehozni. Emellett a későbbiekben grafikus felhasználói felülettel rendelkező szerkesztő környezet kialakítása is cél. Az illesztési mintákat leíró konfigurációs állomány formátuma szintén XML.

Az első elemzési lépés eredményeképpen egy ideiglenes XML dokumentum jön létre, amely az eredeti dokumentum szövegét és a kiegészítő XML címkéket tartalmazza. A második fázis az így megjelölt releváns információ kiemelése, és strukturális átalakítása előre definiált nézetekké (mivel azok struktúrája rögzített). Mivel teljes mértékben XML alapú dokumentumokon dolgozunk, ezért ezt szabványos XSLT transzformáció segítségével megtehetjük. A transzformáció vezérléséhez csupán az XSLT illesztési minták megírására van szükségünk.

Egy dokumentum egyfajta elemzéséhez tehát két XML konfigurációs állományt kell létrehoznunk: a szignifikáns szövegrészletek megjelölését vezérlő illesztési mintákat, és az XML struktúra átalakításához szükséges XSLT illesztési mintákat. Az így megvalósított dokumentumelemzési technika az általunk megvalósított szabály alapú forrásmodellezésnek az alapja. Az információbeszerző rendszer a keresési folyamat során dokumentumokat tölt le a forráskörnyezetről (alapvetően az internetről), és megfelelő elemzési szabályokat rendel hozzájuk. A hozzárendelés a letöltött dokumentum bizonyos sajátosságai alapján történik. Egy ilyen szabály sematikus felépítését láthatjuk az *5. ábrán.*



5. ábra **Dokumentumelemzési szabály**

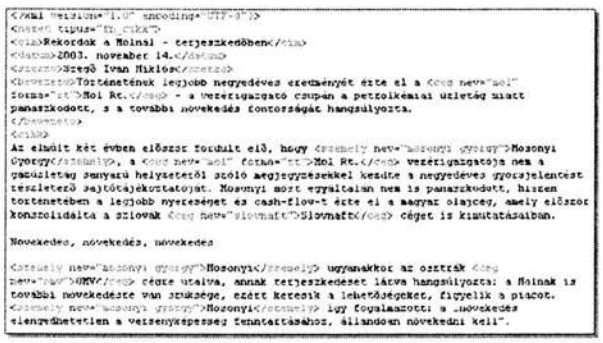
A szabály egy feltétel- és egy akciórészből áll. A feltételrészt próbálja illeszteni a rendszer az aktuálisan bejövő dokumentumra, ez a lépés felelős a dokumentum felismeréséért. Egy dokumentum háromféle sajátossága: a címe (URL), a struktúrája és szöveges tartalma alapján jellemezhető. Mind a három (és tetszőleges logikai kombinációjuk is) lehet a felismerés alapja. Az URL-t egyszerű reguláris kifejezés illesztéssel oldhatjuk meg, különböző strukturális elemek azonosítását az előzőekben bemutatott struktúraelemző segítségével, míg a szöveges tartalmat a struktúraelemző és egyszerű statisztikai módszerek (IR) alkalmazásával ellenőrizhetjük.

Miután a rendszer kiválasztotta a megfelelő szabályt a bejövő dokumentum elemzése alapján, a szabály akciórészében lévő dokumentumelemzési minták segítségével létrehozza a minták által meghatározott nézetekhez a már korábban leírt módon (4. ábra). A rendszer további moduljai,

illetve más elemző rendszerek már ezeken a típussal ellátott, szemantikailag strukturált XML állományokon dolgoznak.

Hagyományos webcsomagoló nyelvek és eszközök csupán előre ismert portáloldalakat képesek modellezni. A mi szabály alapú megközelítésünk segítségével a felhasználók általános forrásmodelleket készíthetnek előre nem ismert vagy részben ismert dokumentumokhoz is, de (az eddigiekhez hasonló) specializált modelleket is létrehozhatunk.

6. ábra **Eredeti HTML dokumentum**



7. ábra **Kinyert információ az XML nézetben**

A *6.* és *7. ábrán* egyszerű példát láthatunk arra, hogy a rendszer milyen formában vágja ki a szükséges információt egy portál cikkeiből. A 6. ábrán található az eredeti portálcikk.[12] A cikket magába foglaló oldal számos zavaró elemet is tartalmaz (hírek, menük, linkek stb.), amelyek nem kívánatosak az alkalmazás számára. A 7. ábrán a kinyert XML nézet látható, amely az elemzés során létrejött. Ebben az egyszerű példában a cikk címe, dátuma, szerzője, bevezetője és szöveges tartalma, illetve azon belül a cégek és személyek nevei lettek kinyerve.

A portálon lévő cikkekhez egyszer kell elkészíteni a megfelelő forrásmodellt, ezután az összes régi és jövőben megjelenő cikk letölthető az ábrán látott szemantikus struktúrával. Természetesen a személy- és cégnevek nem a portálon lévő cikkek sajátosságai, ezek felismeréséhez általános heurisztikákat lehet alkalmazni (mint pl. cégnévnél a nagybetűs szót követő „Rt.", „Kft." vagy „cég" azonosítása, személyneveknél lexikon alkalmazása).

Az eredményül kapott XML nézet már jó hatékonysággal használható fel egyrészt további elemzők bemeneteként (pl. statisztikai elemzés), mivel számos zavaró tényező (reklámok, menü stb.) el lett távolítva. Másrészt a tudástár építéséhez is, hiszen az információmenedzsment modulban lévő nyelvi elemző segítségével (lásd a következő fejezetben) olyan tudásra tehet szert a rendszer, mint:

- A Mol Rt. egy cég.
- Mosonyi György egy személy.
- Mosonyi György a Mol Rt. vezérigazgatója.

Ezután olyan kérdéseket tehetünk fel a rendszernek, hogy például „Mi a Mol cégformája?" vagy „Ki a Mol vezérigazgatója?" Ez már valódi tudás, hiszen az előkészített tárgyterületi modell segítségével a rendszer tényleges szemantikai jelentéseket és összefüggéseket tud felismerni és tárolni.

Ontológiára épülő szolgáltatások

Mi az ontológia?

Az IKF projekt a magas szintű szolgáltatások megvalósításához az ontológiákat használó tudásreprezentációt vezeti be. Mielőtt ezeket a szolgáltatásokat ismertetnénk, nem lesz talán haszontalan röviden áttekinteni, mit is jelent az ontológiákra épülő tudásreprezentáció. Mindenekelőtt azt szeretnénk tisztázni, hogy ebben a kontextusban mit jelent az „ontológia" szó. Félreértésre adhat okot ugyanis, hogy ezzel a szóval különböző tudományterületeken más és más, nem azonos, de azért nem is teljesen különböző fogalmakat jelölnek. A szó görög eredetű, már régóta egy filozófiai diszciplínát jelöl, amely – hagyományos felosztás szerint – a létezőkkel és magával a léttel foglalkozik. A mesterséges intelligenciában a kilencvenes évek elejétől jelent meg ez a fogalom, és vált egyre inkább elterjedtté. Az ontológiák előzményeinek a

tudásbázisok felsőszintű része (az ún. TBox), az adatbázisok sémainformációja, a szemantikus hálók egyes kezdeményezései, és néhány független tudásreprezentációs projekt (pl. Cyc) tekinthetők. A kilencvenes évektől ezeken az egymástól addig független területeken integratív fogalomként jelent meg az ontológia, összekötve addig még kevéssé ismert területeket is (elektronikus kereskedelem, szemantikus web).

Az első különbség a szó ezen új jelentésében az, hogy a mesterséges intelligenciában nem egy diszciplínát jelent, hanem konkrét produktumokat jelöl, és ennek megfelelően többes számban is használják. Az ontológiák ugyanis arra szolgálnak, hogy a számítógépes rendszerek felhasználóinak fejében lévő fogalmi sémát (az ún. konceptualizációt) leképezzék a számítógépes rendszer nyelvére. Most már érthető a kapcsolat a filozófiai diszciplínával: a fogalmi séma feltérképezésénél sok megállapítás vehető át, sőt egyes ontológiákkal foglalkozó és analitikus filozófiai műhelyek között élénk együttműködés is folyik (pl. a mereológia területén).[13]

Minden interdiszciplináris kapcsolata ellenére az ontológia azonban a mesterséges intelligenciában eszköz egy konkrét tudásreprezentációs probléma megoldására. Nézzünk egy példát! Tegyük fel (egy bevett példa nyomán), hogy két gépi rendszer (ágens) borokkal kapcsolatos elektronikus kereskedést szeretne. Az ágenseknek szót kell érteniük egymással abban az értelemben is, hogy melyikük mit ért a különböző borfajtákon, hogyan fejezi ki a borok különböző tulajdonságait stb. Elég kínos lenne ugyanis, ha a rendszer a leadott rendeléstől eltérő, vagy más tulajdonságú borokat szerezne be a fogalmi különbségek révén.

Az ontológiákat először hasonló, ún. sémaegyeztetési feladatokra tartották igazán alkalmasnak, valamint a klasszikus tudásreprezentációs feladatok megoldására gondolták felhasználhatónak.[14] Létrejött néhány nagy kezdeményezés, amely átfogó, felsőszintű ontológia építését tűzte ki céljául. Ilyen a Standard Upper Ontology,[15] amely az IEEE szabvány-előkészítő bizottságaként működik, és ide sorolható *J. F. Sowa* elképzelése is [17], aki sajátos egyéni szintézist hozott létre a koncepcionális hálókra építve, és ezeket a hálókat egy másik szabványügyi szervezetnél, az ANSI-nál próbálja szabványosíttatni.

Ebbe a sajátos szabványosítási „versenybe" beszállt a nagy múltú DARPA szervezet is[16] (amelynek nevéhez fűződik az Internet alapjainak, a DARPANET-nek lerakása). A „versenyben" más szabványügyi testületek is részt vettek, de számunkra most nem ez a fontos, hanem az, hogy – szerencsés módon – egyfajta konvergencia figyelhető meg a különböző kezdeményezések között. Ezt a közeledést nem utolsósorban az ontológiák újabb, egyre nagyobb teret hódító felhasználási területe, a szemantikus web motiválja.

A World Wide Web alapítójaként is emlegetett *T. Berners-Lee* újabb elképzelése szerint a szemantikus web[17] egy olyan új generációs internetes tartalom lenne, amely a gépi ágensek (köztük intelligens keresőprogramok) számára is feldolgozható. Berners-Lee megfogalmazta a szemantikus webet alkotó szolgáltatások egy ún. rétegmodelljét is, és ma egyre több kutató, illetve alkalmazásban érintett szakember előtt tűnik úgy, hogy az ennek felsőbb szintjein megfogalmazott szolgáltatásokat az ontológiák segítségével lehet megvalósítani. A World Wide Web Consortium (W3C), amelyet az Internet *de facto* szabványosító testületének tekintenek, megfogalmazta a Web Ontology Language (OWL) szabványt-javaslatot.[18] A javaslatot a korábban említett DARPA szervezet is támogatja, jelenleg a szabványosítás előtti utolsó szakaszban áll, és januárban a W3C vezető testülete várhatóan el is fogadja.[19]

Az ontológiákat mint tudásreprezentációs eszközt tehát több területen is lehet használni, már egy általánosan elfogadott ontológianyelv szabvány is alakulóban van. Felmerül azonban a kérdés, hogy miként is történik maga a tudásreprezentáció, és hogyan viszonyul az ontológia néhány jól ismert formalizmushoz (tezaurusz, taxonómia stb.). A könyvtári világban ugyanis komoly erőfeszítésekkel kifinomult tezaurusz- és taxonómiarendszerek is létrejöttek, amelyeket – úgy tűnik – az ontológiákkal foglalkozók mintha nem vennének észre, vagy – ami még rosszabb – ellenségesen viszonyulnak hozzá. Ez a magatartás teljesen indokolatlan, és talán el lehet oszlatni a fogalmak tisztázásával. Valójában az ontológiák abban különböznek a taxonómiáktól, tezauruszoktól, szemantikus hálóktól (amelyek mind a tudás reprezentációját szolgálják), hogy logikai háttérrel, formális szemantikával rendelkeznek. Amikor az ontológiákat „ténylegesen működésbe kell hozni," akkor az ontológiában lévő állításokat (közvetlenül vagy közvetve) át kell fordítani ún. leíró logikai állításokká.

A leíró logika (description logics) az elsőrendű formális logika egy rendszere. Tárgyalási univer-

zuma fogalmakból, relációkból (amelyeket itt szerepeknek neveznek) és individuumokból áll. A fogalmak neveiből a szokásos módon (logikai operátorokkal, mint az „és", „vagy" stb.) összetett fogalmak képezhetők, de – és ebben különbözik a leíró logika más, ismertebb logikai rendszerektől – fogalmak között a relációkkal (szerepekkel) kapcsolat létesíthető, és ezek az összetett fogalmak részét képezhetik. A fogalmaknak az individuumok lehetnek a példányai. A leíró logikai rendszerekben olyan kérdések válaszolhatók meg (matematikailag megalapozott algoritmusokkal), amelyek a fogalmak egymás közti tartalmazási viszonyaira és a példányokra vonatkoznak. A leíró logikáknak is több válfaja létezik, annak megfelelően, hogy milyen bonyolultabb nyelvi konstrukciókat (pl. különböző kvantorokat) engedünk meg. A logikában jártasabb olvasóink már talán hasonlónak érzik a leíró logikát az intenzionális (pl. modális) logikákhoz, és megérzésükben nem is tévednek: a leíró logika egyik válfaja éppen a multimodális logikával egyezik meg (más válfajai azonban bonyolultabbak). Ennek a megegyezésnek a felismerése sokat lendített előre a leíró logikákkal kapcsolatos kutatásokon, amelyek az ontológiákkal párhuzamosan, a kilencvenes évektől kezdődően zajlottak. A leíró logikákkal kapcsolatos ismereteket jól összefoglalja a nemrég megjelent kézikönyv [21].

A leíró logikákra alapozott formális szemantika nem öncél, hanem gazdagabb (jobban strukturált) leírást tesz lehetővé. A korábban említett ontológianyelvek (például az OWL) olyan leírásra adnak lehetőséget, amely a tárgyterület fogalmait, a fogalmak attribútumait és relációit rögzíti. Az attribútumok és relációk esetén különböző kikötéseket, megszorításokat tehetünk, a fogalmakat nemcsak tartalmazási hierarchiába szervezhetjük, de (halmaz) logikai műveleteket (pl. két fogalom kizárja egymást, vagy egy fogalom két másik metszete) is használhatunk. Logikai axiómákat is megfogalmazhatunk. Ezután ki lehet számolni a fogalmak egymás közti viszonyait, és ellenőrizni lehet, hogy az individuumállítások konzisztensek-e.

Ebből a rövid ismertetőből is látszik talán, hogy mit is jelent az, hogy az ontológiákra épülő tudásreprezentáció gazdagabb leírást tesz lehetővé. Az is világossá válhat egyben, hogy a bonyolultságnak ára van: az ontológiákat kezelő eszközöket nehezebb létrehozni, és a számítási idők is lényegesen nagyobbak. A korábban ismert tudásreprezentációs eszközöket tehát nem leváltani, hanem kiegészíteni hivatott az ontológia (a „minden feladatra a megfelelő eszközt" elv alapján). Arról nem is beszélve, hogy az ontológianyelvek, a leíró logika és a leíró logikai következtetéseket végrehajtó ún. következtetőgépek csak egy formalizmust definiálnak, amelyet a tényleges tartalommal még fel kell tölteni, és a feltöltöttség szempontjából pedig különösen nagy tisztelettel kell tekinteni a könyvtári világban eddig létrejött produktumokra.

Az ontológiák használata tárgyterület modellezésében

Az IKF projekt célja mind a funkcionalitásról szóló általános jellegű, mind a tárgyterületről (célkörnyezetről) szóló specifikus jellegű tudás beépítése az IKF rendszerbe. Ezt a célt tölti be a tárgyterületmodellező egység, amely az információmenedzsment alrendszer szerves részét képezi. Nyilván a tárgyterületi tudás és a funkcionalitás általános tudása csak tárgyában válik el, tárolásának technológiája azonos. Erre a technológiai feladatra az IKF projekt – a fentiek után talán érthető módon – a tudást tároló ontológiák alkalmazása mellett kötelezte el magát.

Az ontológiák választását a tudásreprezentáció szerepére az is motiválta, hogy az IKF projekt megcélozta gazdasági tárgyterület és az azt leíró gazdasági nyelv egy elméleti diszciplína, a közgazdaság-tudomány hatására formálódik, tehát – várhatóan és részben beigazoltan – logikailag feltárhatóak fogalmi viszonyai. Hosszú távon lehetővé teszi az IKF alkalmazás és a szemantikus web rendszerei közti könnyebb átjárhatóságot, a jelenben azonban megoldandó feladatot jelent, mivel az IKF rendszer forráskörnyezetének dokumentumai jelentős részben gépi feldolgozásra előkészítetlenek (lévén csak embereknek íródtak), tehát az ontológiákkal kapcsolatos eddigi eredmények közvetlenül nem vehetők át. Mindez az IKF projekt saját ontológiaelképzelésének kialakítását tette szükségessé.

Az IKF rendszer ezen alrendszerét tényleges használatbavétele előtt tehát még paraméterezni kell, azaz fel kell tölteni a feladat- és intézményspecifikus tárgyterületi tudással. Ugyanakkor az IKF projekt célja ezen paraméterezés megkönnyítése mind a tárgyterületi modellépítő komponenssel, mind a tudástárban már előzetesen meglévő részlegesen elegendő tudással.

Az ontológiára épülő szolgáltatások

Mi a haszna a tudás modellezésének az IKF projekt céljainak szempontjából? Erre a kérdésre az

ontológiára épülő szolgáltatások adják meg a választ. Ezeket a szolgáltatásokat az IKF projekt során folyamatosan fejlesztjük.

A keresőkérdésekkel kapcsolatos szolgáltatás

Ez a szolgáltatás az ontológiának már egy viszonylag kezdetleges stádiumban is hasznát tudja venni, ugyanakkor megoldást jelent az információkinyerő rendszer tervezése során felmerülő általános problémára. A természetes nyelv és a dokumentumtár indexelt dokumentumainak indexnyelve között ugyanis komoly különbségek lehetnek (poliszémia, szinonímia stb. miatt). Ezenfelül egy általános, index alapú keresés sikerességét sokban javítja egy gondosan kiválasztott, több összetevős keresőszólista.

A funkcionalitás során tehát a természetes nyelven megfogalmazott keresőkérdést úgy alakítja át a rendszer a dokumentumtár indexnyelvén megfogalmazott keresőkérdéssé, hogy nemcsak a keresőkérdés szavainak indexnyelvi megfelelőjét tartalmazza, hanem a háttértudás által vonatkozónak tartott indexnyelvi szavakat is. Ez a kibővítési eljárás bővítési operátorok használatával történik. Először meg kell keresni a természetes nyelvi szavak által jelölt fogalmak ontológiabeli megfelelőjét, mert a bővítési operátorok az ontológián értelmezettek.

Minden bővítési operátor egy adott fogalomból kiindulva három fogalomlistát eredményez: a tartalmazó, az azonos és a tartalmazott fogalmak listáját. Ehhez a három fogalomlistához három különböző súlytényező is tartozik (az eddigi tapasztalatok alapján a legkisebb súllyal a tartalmazott fogalmakat kell figyelembe venni, míg az azonos fogalmak súlytényezője természetesen egységnyi). A konkrét bővítési operátorok ennek a sémának a kitöltésével származtathatók: a kiinduló fogalom lehet a keresőkérdés fogalma (a tapasztalat alapján a bővítési operációnál vagy-szemantikát kell alkalmazni), annak negáltja, és fogalomközti viszonyok által implikált fogalmak. A bővítési operátorok konkretizálása során ismét megjelenik egy súlytényező (pl. a negált esetben negatív egységnyi, a közvetve származtatott fogalmaknál egy diszkontáló jellegű tényező), amely a másik súlytényezővel összeszorzódik. Ezután a fogalomból az indexnyelvi szót kell származtatni. Mivel egy fogalomhoz több indexszó is tartozhat, amelyek közül egyesek kevésbé jellemzőek, ezért itt ismét fellép egy súlytényező. Az összevont funkcionalitás kimenetén ennek a konverziónak az eredménye jelenik meg.

Vizsgálataink alapján ez a funkcionalitás jelentősen javítja a találatok relevanciáját, és segít a releváns találatok kiemelésében is [22].

Természetes nyelvű szövegek elemzése

A keresőkérdés kiegészítésével segít a releváns dokumentumok (avagy dokumentumrészletek) megtalálásában, azonban az információigény kielégítéséhez még mindig a rendszer emberi felhasználójának kell a megfelelő információt kiemelnie a szövegből. Ez a feladat, az írásos szöveg értelmezése általános esetben rendkívül bonyolult (beszélnek például a hermeneutikáról, az értelmezés tudományáról, vagy inkább az értelmezés művészetéről). A mindennapi keresési gyakorlatban felmerülő információs igények azonban sokkal egyszerűbben nyerhetők ki (azonban még mindig szükség van ehhez a nyelvi kompetenciára). A projekt az egyszerűbb ilyen természetű információs igények kinyerésének automatizálást is céljául tűzte ki.

Ennek a képességnek a megteremtéséhez két részfeladatot kell megoldani: létre kell hozni egy természetesnyelv-elemző eszközt (NLP), amely a humán nyelvi kompetencia megfelelője; valamint modellezni kell a háttértudást, vagyis azt a tudásdarabot, amely a szöveg (szükséges mértékben történő) értelmezéséhez és az információdarabok összeállításához szükséges.

A projekt keretében először a megfelelő NLP-eszközt kellett létrehozni. Egy mondattani szintű nyelvtani elemző készült, amely a Morphologic Kft. morfoszintaktikai elemzőjére támaszkodik. Az elemzéshez az MTA Nyelvtudományi Intézete által felállított igei vonzatkeret-gyűjteményt használjuk. Az elemző első változata csak a mondatok nagy részének gerincét alkotó predikatív szerkezeteket (alany–állítmány–tárgy hármas) és annak néhány bővítményét tudta felismerni, azonban a projekt jelenlegi szakaszában készül az elemző újabb változata, amelytől nagyobb hatékonyságú mondatelemzést várunk el (különösen az összetett mondatok terén).

A mondatok elemzése során több problémával kell megküzdeni. Mindjárt a szavak alaktani elemzésénél gondot jelent, hogy olyan szóalakokat is fel kell ismerni, amelyek szótöve nincsen benne a magyar nyelv még legteljesebb szótárában sem. Ezek többnyire ragozott tulajdonnevek (pl. cégnevek, terméknevek) vagy tudományos terminusok. Az eddigi alaktani elemzők rögzített szótárral dolgoz-

tak, ezért kiegészítésükre készítenünk kellett egy ún. heurisztikus alaktani elemzőt, amely ismeretlen szótövek esetén is képes elemzési javaslatokat szolgáltatni. Nehézséget jelent a többféle elemzési variáns megjelenése. Ez a szavak szintjén kezdődik, de a mondatelemzési szabályoknál is felbukkanhat. Egy másik, sokkal mélyrehatóbb probléma abból ered, hogy a különálló mondatok nem azonos szintű kifejezéssel referálnak ugyanarra a dologra. Nézzünk erre egy kisebb példát:
„Az IKF-prototípus alkalmazás több részből áll. A rendszert ezért lehet modulárisnak is nevezni."

A második mondat tárgya azonos az első mondat alanyával (pontosabban szólva ugyanaz a jelölete a két szónak). Ámde a második szó egy általános kifejezés („rendszer"), amely azonban nem az összes rendszerre vonatkozik (mint ezt a határozott névelő is jelzi). Meg kell tehát találni azt a (korábban előfordult) valamit, ami rendszernek is mondható (azaz egy felsőbb fogalomként érvényes rá az a predikátum, hogy rendszer). Az ilyen típusú feladatokat nevezik anafóra-feloldásnak, és – véleményünk szerint – ez hosszabb távon csak ontológia felhasználásával lesz megoldható (amely pl. tárolja azt a tudást, hogy egy számítógépes alkalmazás egy rendszer). Térjünk azonban vissza a prototípus szintjén is megvalósított funkciókhoz.

Amint azonban már korábban említettük, az információkinyerés nem feltételez tökéletes NLP-eszközt, így már a fenti mondatelemzővel is eredményeket lehet elérni. A továbblépéshez azonban a fent említett második részfeladat megoldására, a háttértudás modellezésére is szükség volt. Mint az eddigiek fényében már sejthető, ezt a feladatot az ontológiák felhasználása hivatott megoldani. Ez egyrészt a tárgyterületi tudást tartalmazó ontológia felépítését, másrészt az ontológiát kezelő eszközöket igényli. Ezek az eszközök egy leíró logikai következtetőgépen alapulnak, és az ontológia is a leíró logika nyelvén lett megfogalmazva. Szükség van azonban egy közvetítő rétegre a tárgyterületi tudás és az NLP-elemzés kimenete között. Ezért a kidolgozott ontológiába a nyelvtani elemzés logikai modellje is bekerült.

A kijelölt szövegrészek mondatait elemezzük, majd az eredmény az ontológiához kötődő tudásbázisba kerül. A keresőkérdés hasonló feldolgozása után pedig egy algoritmus szerint kinyerjük a tudásbázisból azokat az információkat, amelyek a keresőkérdés kijelölte ontológiai bejegyzésekhez tartoznak. A felhasznált logikai apparátus kifejezőereje lehetővé teszi, hogy akár olyan származtatott fogalmakat keressünk, amelyek közvetlenül nem is fordulnak elő a forrásszövegekben.

Az információkinyerő funkcionalitás fejlesztése jelenleg még kísérleti fázisában van, azonban a projekt következő szakaszában szeretnénk beépíteni a prototípus-alkalmazásba. Nézzünk azonban egy példát a működésére! Adva vannak rövid gazdasági hírek, amelyek cégek teljesítményéről szólnak. A feladatunk ennek alapján eldönteni, hogy a hírek a cég helyzetének javulásáról vagy romlásáról szólnak (azaz minősíteni kell a cégeket). Ehhez az információkinyerő alkalmazás számára kiépítettünk egy ontológiát, amely a minősítéshez szükséges szabályokat, és a prosperál (jelölése felfelé nyíl), avagy rosszul teljesít (jelölése lefelé nyíl) fogalmakat tartalmazta egységes logikai formátumban. (A szabályok tulajdonképpen a prosperál, és a rosszul teljesít fogalmak jelentését írják le.) A kísérleti rendszer teljesítményét a *8. ábra* mutatja.

Goodyear: a rosszalkodó hetedik 03.4.30 14:46

A Goodyear Tire & Rubber Co., a világ egyik legnagyobb gumiaroncs-gyártója hiába javított eredményein az elmúlt negyedévben hét ágazata közül hatban, a hetedik, az észak-amerikai gumi-szektor mindent elrontott.

>>>Tovább

AstraZeneca: túl pesszimisták voltak az elemzők 03.4.30 14:05

Az első negyedévben a generikus termékek versenye visszavetette kissé az AstraZeneca nyereségét, ami így jobb lett a vártnál.

>>>Tovább

Cégnév	Minősítés
AstraZeneca	▽
Goodyear Tire & Rubber Co.	△
Eurotunnel	▽
Novo Nordisk	△
SSL International Plc.	△
Adidas	△
Hugo Boss	▽
Solvay	△

8. ábra **A kiindulási hírek és a gépi minősítés eredménye**

Mint említettük, az információkinyerő funkcionalitás még korántsem befejezett, azonban a felmerülő problémák (pl. a nyelvtani elemző tökéletlensége, az ontológia hiányossága) nem a lényeget érintik, hanem csak az eddigi munka folytatását igénylik. Végleges formájában ez a funkcionalitás nagyot segíthet a tudásalapú információkinyerésben. (Természetesen az információkinyerés ilyen automatizálása csak az emberi szempontból könnyen értelmezhető szövegek esetén jöhet szóba, ámde az ilyenek alkotják a mindennapi információkeresési gyakorlat jelentősebb részét.)

A készülő prototípusrendszer

A kész IKF prototípusrendszer elsősorban demonstrációs célokra és kutatási eredmények bemutatására szolgál, azonban – ahogy ez szándékaink szerint kiderülhetett – valós kereskedelmi körülmények között is bevetésre szánjuk. Bár a projekttervezet szerint az elsődleges célterület a gazdasági szféra, egy ilyen rendszernek számos egyéb alkalmazási területe lehet. A kitűzött célok között erős hangsúllyal szerepel az, hogy nagymértékben hangolható, parametrizálható és moduláris eszközt fejlesszünk ki. A tervezett nagy szabadságfokú hangolhatóságot három dimenzió mentén lehetne ábrázolni, amelyek megfelelnek a tanulmány első részében ismertetett környezeti modelleknek is.

Az *első* tengely az alkalmazási területtel (célkörnyezettel) azonosítható, azaz a rendszernek alkalmasnak kell lennie tetszőleges tárgyterületi háttértudás befogadására. A jelenlegi prototípusban az üzleti szférának megfelelő tudásmodell van kialakítva, de más területekhez is készülhet alkalmazás, mint például tudomány, szórakoztatóipar, oktatás, egészségügy stb.

A *második* dimenzió a felhasználók mentén helyezkedik. A rendszernek képesnek kell lennie arra, hogy különböző felhasználói igényeket kezeljen, akár egy alkalmazáson belül is. Más információt és más jellegű szolgáltatásokat igényel egy ügyvezető igazgató, mást egy brókerügynök, és mást egy hitelbíráló szakember. A rendszer felhasználói felülete kulcsfontosságú lehet, ha egy konkrét alkalmazást valós kereskedelmi körülmények között szeretnénk működtetni, így ennek a fejlesztése a jövőben nagyobb hangsúlyt fog kapni.

Végül fontos annak a szem előtt tartása is, hogy egy igazán sokoldalú és teljes megoldást kínáló alkalmazásnak tudnia kell kezelni többféle információforrást, ami a *harmadik* szabadsági fok a hangolhatóság terében. A jelenlegi kutatások az internetre és strukturálatlan dokumentumokra koncentrálnak, de számolni kell egyéb forrásokkal is, mint például belső adatbázisok, adattárházak, szemantikailag strukturált, de ismeretlen dokumentumok. A rendszernek integrálnia kell a heterogén forráskörnyezeteket, és univerzális interfészt rendelni hozzájuk.

Ami a prototípus technológiai kialakítását illeti, szintén cél volt, hogy a rendszer fejlesztése során kipróbáljuk és felhasználjuk a jelen lévő korszerű eszközöket és megoldásokat. Ehhez kapcsolódik, hogy minél több licenc nélkül használható, ingyenesen hozzáférhető eszközt próbáltunk bevonni a fejlesztésbe.

További információ a projekttel kapcsolatban a *http://ikf.mit.bme.hu* webcímen található.

Jegyzetek

[1] A szoftverekről bővebben lásd: IBM Corp., *http://www.lotus.com*; Hummingbird Ltd., *http://www.fulcrum.com*; Verity, Inc., *http://www.verity.com*; Excalibur Search Corp., *http://www.excaliburseach.com*; Autonomy Corp., *http://www.autonomy.com*

[2] IKTA 181/2000 – Információ és tudás tárház.

[3] A nemzetközi projektpályázat 2000 áprilisában a nemzeti képviselők támogatásával megkapta az EUREKA státust; EUREKA projektszám: 2235.

[4] A HTML szabványról bővebben lásd: *http://www.w3.org/MarkUp*

[5] World Wide Web konzorcium: *http://www.w3c.org*

[6] Az SQML szabványról bővebben: *http://www.w3.org/MarkUp/SGML*

[7] A DTD szabványról bővebben: *http://www.w3c.org/DTD*

[8] Az XSLT szabványról bővebben: *http://www.w3c.org/Style/XSL/*

[9] Az Xpath szabványról bővebben: *http://www.w3c.org/TR/xpath*

[10] Unified Resource Locator (URL), bővebben lásd: *http://www.w3.org/Addressing*

[11] A reguláris kifejezésekről bővebben lásd pl.: *http://www.wikipedia.org/wiki/Regular_expression*

[12] A példacikk forrása a FigyelőNet online hírportál, *http://www.fn.hu*

[13] Ilyen kezdeményezésekről pl. a [16] konferenciakötetben olvashatunk.

[14] Ebben a megközelítésben tárgyalja az ontológiákat egy magyar nyelven is megjelent, átfogó mesterséges intelligencia könyv [18] is.

[15] A projektről bővebb tájékoztatás a *http://suo.ieee.org* címen található.

[16] Elképzelésük, a DAML+OIL (DARPA Agent Markup Language + Ontology Interchange Language), amelyről a *http://www.daml.org* oldalon található bővebb leírás.

[17] Az eredeti elképzelés a [19]-ben fogalmazódott meg, jelenlegi állásáról áttekintést ad pl. [20].

[18] Hivatalos weboldala a *http://www.w3.org/2001/sw/WebOnt/*

[19] Ezek az ontológianyelvek a szemantikus web ún. RDF(S) szabványcsomagjára épülnek, annak kiterjesztései. Az RDF(S) ugyanis, mint ez a kezdeti alkalmazások során kiderült, önmagában túl rugalmas ahhoz, hogy ontológia tárolására lehessen használni. (A legfőbb problémát az okozza, hogy az önhivatkozás mint nyelvi elem bevezetése miatt a nyelv sze-

mantikája logikailag nagyon nehezen kezelhetővé válik. Ezért az ontológianyelvek kibővítik a nyelv szintaktikáját [új primitíveket adnak hozzá], viszont korlátozzák a nyelv szemantikáját.)

Irodalom

[1] TURBAN, E.–ARONSON, J. E.: Decision Support Systems and Intelligent Systems. 6th Edition, Prentice Hall, 2000.

[2] EUREKA PROJECT: IKF – Information and Knowledge Fusion. Institute of Cognitive Sciences and Technology, Laboratory for Applied Ontology, Italy, March 2000.

[3] The IKF Architecture, IKF project report. Budapest University of Technology and Economics, Department of Measurement and Information Systems, Budapest, Hungary, August, 2002.

[4] MÉSZÁROS T.–BARCZIKAY Zs.–BODON F.–DOBROWIECKI T.–STRAUSZ Gy.: Building an Information and Knowledge Fusion System. IEA/AIE-2001 The Fourteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Budapest, Hungary, June 4–7, 2001.

[5] BAEZA-YATES, R.–RIBEIRO-NETO, B.: Modern Information Retrieval. Addison-Wesley Pub Co, 1999.

[6] PAZIENZA, M. T.–SIEKMANN, J.–CARBONELL, J. G.: Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer Verlag, 1997.

[7] KUHLINS, S.–TREDWELL, R.: Toolkits for Generating Wrappers (A Survey of Software Toolkits for Automated Data Extraction from Websites). Net. ObjectDays–2002, Erfurt, Germany, October 2002.

[8] LAENDER, A.–RIBEIRO-NETO, B.–SILVA, A.–TEIXEIRA, J.: A Brief Survey of Web Data Extraction Tools. SIGMOD Record, 31. köt. 2. sz. 2002.

[9] CHIDLOVSKI, B.–RAGETLI, J.–RIJKE, M.: Wrapper Generation via Grammar Induction. = Ramon Lopez de Mantaras and Enric Plaza, editors, Proceedings ECML 2000, 1810. sz. LNAI, Springer, 2000. p. 96–108.

[10] KNOBLOCK, C. A.–LERMAN, K.–MINTON, S.–MUSLEA, I.: Accurately and reliably extracting data from the web: A machine learning approach. = Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2000.

[11] AROCENA, G. O.–MENDELZON, A. O.: WebOQL: Restructuring Documents, Databases and Webs. = Proceedings of the 14th International Conference of Data Engineering, Orlando, Florida, 1998. p. 24–33.

[12] SAHUGUET, A.–AZAVANT, F.: Web Ecology: Recycling HTML pages as XML documents using W4F. = Proceedings of the Second International Workshop on the Web and Databases, Philadelphia, Pennsylvania, 1999. p. 26–31.

[13] HAROLD, E. R.–MEANS, W. S.: XML in a Nutshell. 2nd Edition, O'Reilly & Associates, 2002.

[14] BROADSHOW, J. M.: Software Agents. The MIT Press, 1997.

[15] DEZSÉNYI Cs.–MÉSZÁROS T.: Domain Knowledge Based Document Retrieval. IEEE-TTTC – International Conference on Automation, Quality and Testing, Robotics, Cluj-Napolca, Romania, May, 2002.

[16] GUARINO, N. (ed. Formal Ontology in Information Systems.): Proceedings of FOIS'98, Trento, Italy, 6–8 June 1998. IOS Press, Amsterdam.

[17] SOWA, J. F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

[18] RUSSEL, J. S.–NORVIG, P.: Mesterséges intelligencia – modern megközelítésben. Panem Kft., 2000.

[19] BERNERS-LEE, T.–HENDLER, J.–LASSILA, O.: The SemanticWeb (A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities). = Scientific American, May 17, 2001.

[20] FENSEL, D.–WAHLSTER, W.–LIEBERMAN, H.–HENDLER, J.: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.

[21] BAADER, F.–CALVANESE, D.–McGUINNESS, D.–NARDI, D.–PATEL-SCHNEIDER, P.: The Description Logic Handbook (Theory, Implementation and Applications). Cambridge University Press, January, 2003.

[22] VARGA P.–MÉSZÁROS T.–DEZSÉNYI Cs.–DOBROWIECKI T.: An Ontology-based Information Retrieval System. Developments in Applied Artificial Intelligence, 16th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE 2003, Laughborough, UK, June 23–26, 2003. Proceedings Lecture Notes in Computer Science, 2718. köt. Springer-Verlag.

Beérkezett: 2004. II. 2-án.