Magyar internetes gazdasági tematikájú tartalmak keresése

Tikk Domonkos[1], Biró György[2], Szidarovszky P. Ferenc[1,3], Kardkovács Zsolt T.[1], Héder Mihály[1], Lemák Gábor[4]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.
{tikk, szidarovszky, kardkovacs}@tmit.bme.hu, merlin@sch.bme.hu
[2] TextMiner Bt.
H-1029 Budapest, Gyulai P. u. 37.
george.biro@gmail.com
[3] Szidarovszky Kft.
H-1392 Budapest, Pf. 283.
ferenc.szidarovszky@szidarovszky.com
[4] GKI Gazdaságkutató Zrt.
H-1092 Ráday u. 42–44.
lemakg@gki.hu

Kivonat: A projektünk célja egy olyan keresőszolgáltatás kiépítése, amely az Interneten magyar nyelven elérhető gazdasági tematikájú tartalmak lehető legteljesebb körét egy helyen kereshetővé, és – amennyiben a tartalomszolgáltató, ill. jogtulajdonos részéről ennek nincs akadálya – elérhetővé is teszi a felhasználók számára. Jelen munkánk ismerteti a szolgáltatás funkcióit, felépítését, és megvalósítását. A komponensek közül részletesen foglalkozunk a nyelvtechnológiái módszereket alkalmazó szövegfeldolgozó és webszüretelő modulokkal.

1 Bevezetés

A projektünk célja egy tematikus, szemantikus elveket és újfajta vizualizációt alkalmazó keresőszolgáltatás létrehozása, amelyen keresztül a felhasználók az Interneten magyar nyelven elérhető gazdasági tematikájú tartalmak lehető legteljesebb körében kereshetnek, és amely – amennyiben a tartalomszolgáltató, ill. jogtulajdonos részéről ennek nincs akadálya – tartalmakhoz való hozzáférést is biztosítja a felhasználók számára. A keresőszolgáltatók egyre bővülő piacán egy olyan szegmenst célzunk meg, amely jól körülhatárolható, de korántsem elhanyagolható jelentőségű felhasználói kör. A tágabb értelemben vett gazdasági tartalmak érdekelhetik mind az átlagfelhasználót (pl. kisbefektetői kör, laikus érdeklődők), mind a vállalatvezetői, tanácsadói, döntéshozói pozícióban lévőket, mind pedig a szakmai felhasználókat – oktatók, kutatók, hallgatók.

Projektünk eredményétől azt várjuk, hogy a keresési kérésnek megfelelő dokumentumok pontosabban kielégítik a felhasználói igényeket, mint a jelenlegi alkalmazások, illetve az újfaja vizualizáció lerövidíti az információfeldolgozás idejét. Emellett a projektmegvalósítás során egy olyan know-how is létre jött, amelynek segítségével újabb tematikákkal, tudományterületekkel bővíthetjük a szemantikus keresést biztosító keresőszolgáltatásunkat, ezáltal hozzájárulva a magyar nyelvű világháló jelentés-orientálttá válásához.

Cikkünk felépítése a következő. Először a 2. szakaszban ismertetjük a kitűzött funkcionalitásokat, keresési formákat, majd a 3. szakaszban bemutatjuk a rendszer felépítését és az egyes komponenseket. A 4. szakaszban a működés során jellemző folyamatok vizsgálata következik hangsúlyozottan kiemelve a nyelvtechnológiai eljárásokat alkalmazó komponensek vonatkozó részleteit, míg az 5. szakasz a hasonló, elsősorban hazai vonatkozású kezdeményezéseket veszi számba. Végül a 6. szakaszban röviden összegzést adunk.

2 Támogatott keresési formák

A keresőszolgáltatás keresési funkcióinak meghatározása során célunk az volt, hogy a szokásos keresési lehetőségeknél fejlettebb szolgáltatásokat nyújtsunk, és támogassuk a felhasználóknak a keresési eredmények böngészése, azokon való navigálás során felmerülő továbbkeresési igényeit.

A keresőmotorok hatékonyságának növelésére egyik lehetőség, ha a felhasználó meghatározhatja a keresett tartalom tematikáját. Ez segíti a keresőmotort a keresési igény pontos meghatározásában, pl. több értelmű keresőkifejezések esetén, és leszűkíti a találat lista méretét csökkentve ezáltal az irreleváns találatok számát. A felhasználók tematikus navigációjának, illetve keresésük orientálásának támogatása a tartalmak tematikus rendszerezésével érhető el, ennek előfeltétele, hogy rendelkezésre álljon a megcélzott tematikát lefedő megfelelő részletezettségű hierarchikus kategóriarendszer (*taxonómia*). Az általános, nagy nemzetközi keresőszolgáltatások is rendelkeznek hasonló keresési lehetőséggel, (ld. pl. Google Directory, Yahoo Directory, Zeal kereső[1] stb.), de egy ilyen opció jelentősége egy tematikájában és nyelvében eleve korlátozott tartalomgyűjteményt összefogni kívánó szolgáltatás esetén sokkal számottevőbb. Az általános keresők esetén ugyanis sokkal nehezebb egy mindenre kiterjedő, kellően részletes taxonómia megalkotása és karbantartása, valamint szintén nagy kihívást jelent a taxonómia megfelelő minőségű tartalommal való feltöltése. A projektünk által megcélzott szűkebb tematika és rögzített nyelv viszont a tartalmak sokféleségéből és a témák dinamikus változásából eredő taxonómia-karbantartási feladatok bonyolultságát jelentősen csökkenti.

Esetenként a felhasználó számára rendelkezésre áll a kereséséhez egy teljes kiindulási – akár saját készítésű – dokumentum, amelyhez hasonlókat kíván megtekinteni. Az általános keresők nem támogatják a bizonyos szószámot meghaladó[2], hosszabb keresőkifejezéseket, ezért – amennyiben a dokumentum nincs indexelve – nem képesek a feladatot végrehajtani.

[1] http://www.google.com/dirhp, http://search.yahoo.com/dir, http://www.zeal.com

[2] A Google legfeljebb 32 szavas keresőkifejezéseket értelmez.

Egy keresés találati listája és annak elemei gyakran szintén fontos kiindulási pontot jelenthetnek további keresések kezdeményezésére, a keresett tartalom pontosítására, a keresés finomítására. A felhasználó számára azonban korántsem egyértelmű – még a találatok rövid átfutása után sem –, hogy milyen módon tudja leghatékonyabban bővíteni, vagy módosítani a keresését. Ezt a tevékenységet a találati listában lévő dokumentumok kulcsszavainak felkínálásával eredményesen lehet támogatni.

A felsorolt megfontolások alapján a következő keresési funkciókat határoztuk meg:

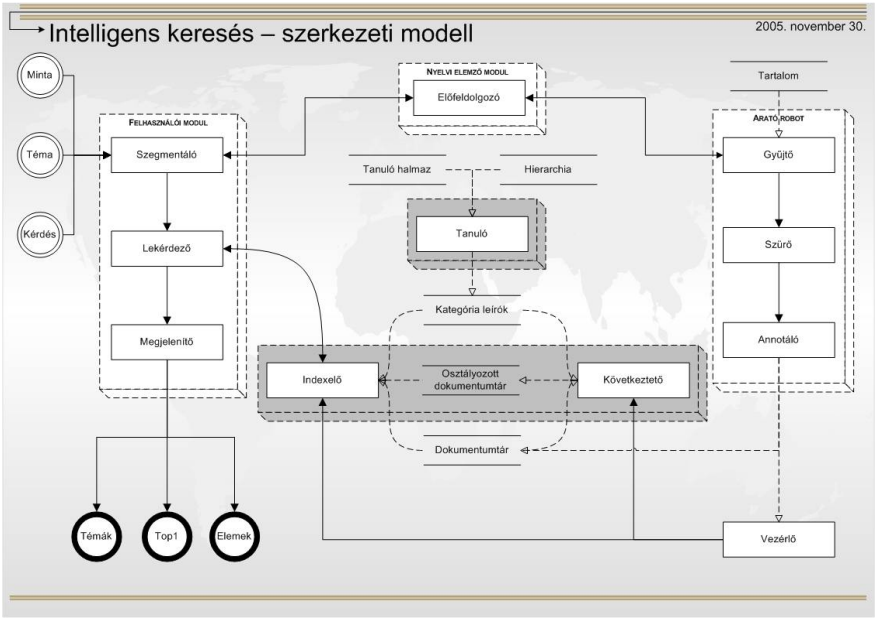
– szabadszavas kérdés-vezérelt keresést,
– mintadokumentum alapú keresést,
– tematikus böngészési lehetőséget rögzített témastruktúrában,
– keresés finomítási lehetőség a találatok kulcsszavai alapján.

3 A rendszer felépítése

A keresőszolgáltatás mögött egy négy fő komponensből álló, összetett rendszer hivatott kiszolgálni a felhasználók igényeit (ld. 1. ábra). A rendszer fő komponensei az alábbiak:

– aratórobot,
– nyelvi feldolgozó modul,
– indexelő és osztályozó motor,
– felhasználói felület.

Az alábbiakban bemutatjuk az egyes komponensek feladatát és vázlatos működését.



1. ábra. A rendszer szerkezeti modellje

Aratórobot

Az aratórobot feladata a kiválasztott, gazdasági témájú híreket (is) közlő magyar oldalakról való tartalomgyűjtés és a rendszer által használt XML formátumra való átalakítás. A cikk írásakor a tesztrendszer mintegy 50 forrásból gyűjti a dokumentumokat, melyek közt túlnyomó részben országos gazdasági tartalomszolgáltatók anyagai szerepelnek, de vannak köztük általános portálok témába vágó cikkei, és regionális tartalomszolgáltatók is.

Nyelvi feldolgozó modul

A nyelvi feldolgozó modul feladata, hogy a különböző forrásokból beérkezett dokumentumokat egységes formátumra hozza. A feldolgozás folyamatát úgy határoztuk meg, hogy akár különböző alapú szövegreprezentációs technikák (pl. szó-alapú vagy karakter n-gram alapú) is megvalósíthatóak legyenek, a feldolgozás során, pedig tetszőleges nyelvtechnológiai eszközök alkalmazásának eredményeit is integrálni lehessen.

Indexelő és osztályozó motor

Az indexelő motor feladata a keresés végrehajtásához szükséges indexállomány létrehozása, karbantartása és a keresések kiszolgálása. Az osztályozó motor a taxonómiával bővített kulcsszó alapú kereséshez szükséges kategóriainformációk nyilvántartását, ill. meghatározását támogatja. Az osztályozó modul felügyelt gépi tanulást végez, azaz tanítódokumentumok alapján megtanulja a taxonómia kategóriáinak jellemző szavait, ill. kifejezéseit. Ennek megvalósítására a HITEC osztályozóalgoritmusát integráltuk a rendszerbe [1, 2]. Az osztályozó motor segítségével tehát egyrészt lehetőség van az egyes kategóriák jellemző szavainak, ill. kifejezéseinek meghatározására, azaz ún. kategóriaprofilok kiépítésére, másrészt a rendszerbe kategóriacímke nélkül bekerülő dokumentumok kategóriáinak automatikus becslésére. Ezek a motorok korábbi fejlesztések eredményeiként álltak elő, ezt az 1. ábrán szürke alapszínnel jelöltük.

Felhasználói felület

A felhasználói modul biztosítja a keresési felületet a felhasználók felé, a lekérdezések továbbítását a keresőmotorhoz, illetve a keresőmotortól kapott eredmények megjelenítését és feldolgozását.

Taxonómia kiépítése és feltöltése

A keresőszolgáltatás hatékonysága és a keresés minőségének biztosítása szempontjából kiemelt fontosságú, hogy a taxonómia jól reprezentálja a tématerületet, kellően részletes finomítását adja a legfontosabb fő témaköröknek, ugyanakkor a kapcsolódó, ill. peremterületeket érintő tematikát is lefedje. Ezért a gazdasági témájú szövegeket tematikus osztályozásának alapját jelentő taxonómiát egy szakkönyvtár, a Budapesti Corvinus Egyetem Központi Könyvtárának tárgyszórendszere alapján alakítottuk a könyvtár szakértőinek segítségével. A kiindulási tárgyszórendszer különböző kapcsolattípusokat tartott nyilván (szűkebb/bővebb terminus, használt/nem használt terminus, kapcsolódó fogalom), valamint köröket is tartalmazott, ezért közvetlenül nem volt alkalmas egy hierarchikus, csak generatív/partitív relációkat tartalmazó taxonómia megalkotására. Az átdolgozást a könyvtár munkatársai végezték el az informatikus szakértők útmutatásai alapján. Ennek során elsődleges szempont a megfelelő struktúra kialakítása volt, úgy hogy a tárgyszórendszer élő elemei a taxo-

nómiába is átkerüljenek. A megfelelő struktúra kialakítására néhány új, korábbi tárgyszavakat egy csomópontba összekapcsoló kategóriát is létrehoztunk. Az így kialakított fastruktúrájú taxonómia 16 legfelső szintű kategóriából kiindulva összesen 2397 kategóriát tartalmaz, legnagyobb mélységében hat szintes. Amennyiben a rendszer tesztelése során kapott felhasználói visszajelzések szükségessé teszik, a taxonómia még módosulhat.

A taxonómia osztályozáshoz való felhasználására feltétlenül szükség van tanulódokumentumokra, azaz olyan mintákra, amelyek jól reprezentálják az egyes kategóriákat. Ehhez természetesen a BCE Könyvtár tárgyszórendszerét használtuk fel, mivel így számos, a könyvtár eredeti tárgyszórendszere segítségével annotált elektronikus dokumentum azonnal a rendelkezésünkre állt. A tanulókörnyezet teljessé tételét, azaz hogy minden lényeges csomóponthoz megfelelő számú tanulóadat legyen, úgy valósítottuk meg, hogy lehetőség szerint beszereztük könyvtári katalógusrendszerben elektronikus formában nem szereplő dokumentumok elektronikus verzióját, illetve újonnan annotált elektronikus dokumentumokkal bővítettük a rendszert.

4 A rendszer működése

4.1 Dokumentumok feldolgozása és tárolása

A rendszerbe való bekerülés módjától függően két dokumentumtípust különböztetünk meg: tanuló- és szüretelt dokumentumokat. Az egyedüli különbség, hogy a tanulódokumentumok rendelkeznek kategóriainformációval, míg a szüreteltek nem. (A felhasználó által megadott keresőkifejezéseket a feldolgozás szempontjából a szüretelt dokumentumokkal analóg módon kezeljük, csak ezeket nem tároljuk el.) A rendszerbe kerülő dokumentumok eredeti formátuma több féle lehet, HTML, PDF, DOC, RTF, illetve szöveges (TXT), ezeket a rendszer által használt XML alapú reprezentációra kell megfelelő konverziós eljárások alkalmazásával átalakítani. A dokumentumok tárolására egy olyan egyszerű, de a dokumentumfeldolgozás bármely lépését tárolni képes struktúradefiníciót (DTD) hoztunk létre[3], amely elsősorban a szövegbányászati feladatok elvégzésére optimalizált, de egyszersmind könnyen átalakítható bármely más szabványos XML formátumra (pl. NewsML, TEI[4], stb.).

A DTD létrehozásakor fontos szempontot volt, hogy

1. a meghatározott információk kódolására képes legyen az XML struktúra,
2. az XML formátumú szöveg tárigényének minimalizálása. Ennek kiemelt jelentősége van egyrészt a dokumentumgyűjtemény mérete, másrészt a feldolgozó algoritmusok működési sebessége és memóriaigénye miatt.

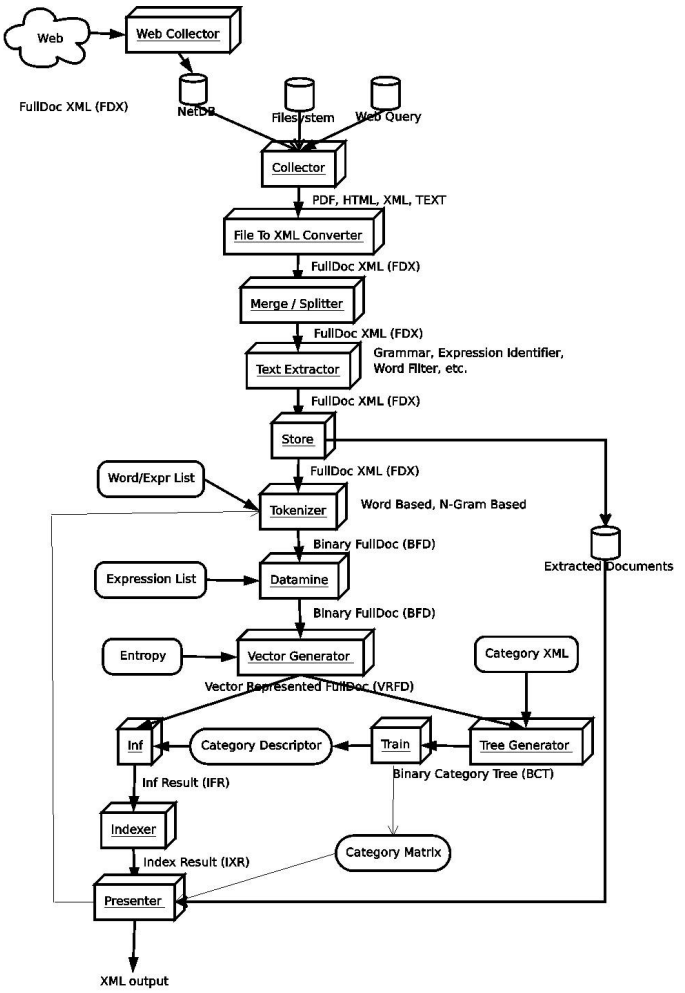
Az első követelményt a viszonylag rugalmas szerkezeti felépítés segítségével értük el, míg a másodikat a gyakran ismétlődő szerkezeti elemek rövid elnevezésével, és csak a feltétlenül szükséges attribútumok kötelező megadásával valósítottuk meg. A DTD tervezése során kiemelt figyelmet fordítottunk arra, hogy az összes szöveges mező elemek tartalmaként kerüljön feldolgozásra, és kizárólag az egyéb metainformációk kerüljenek az egyes elemek attribútumaiba. Ezzel egyrészt a konvertáló programok számára kívántuk feldolgozási konvenciót előírni, s ezáltal a kon-

[3] fulldoc.dtd – Jelenlegi elérhetősége: http://dodona.tmit.bme.hu/~tikk/fulldoc.dtd

[4] http://www.newsml.org/pages/spec_main.php, http://www.tei-c.org/P4X/ST.html

verziót megkönnyíteni, másrészt ily módon oldottuk meg a szövegjellegű információk és a metaadatok keveredésének kiküszöbölését. A szöveges adatok képezik az indexelő modul primer bemenetét, míg az attribútumokban előforduló metaadatok az intelligens szövegbányászati és nyelvtechnológiai módszerek bemeneteként szolgálnak.

A szövegfeldolgozási folyamat lépéseit a 2. ábra ismerteti. Ez alapján látszik, hogy minden dokumentum esetén ugyanaz a feldolgozás folyamata. Az eredeti dokumentumok XML konverziója után (FDX formátum) a Merger/Splitter modul a dokumentumok összefűzésén kívül a karakterkódolás egységesítését is elvégzi. A Text Extractor komponens nyelvtechnológiai eljárások alkalmazását végzi. Ennek során az alábbiakat valósítja meg:



2. ábra. A feldolgozási folyamat lépései

- **Szótövező**: A rendszer két alternatív lehetőséget kínál a feladat elvégzésére. Egyrészt tartalmazza a szabályalapú, ún. óvatos szótövező algoritmust [4], másrészt pedig integrálja a Szószablya projekt[5] HunMorph csomagjának HunStem szótövesítő eljárását [5]. A DTD lehetőséget ad különböző elemzési alternatívák kezelésére (ld. g[rammar] elem), így több szótő megadására is, amit a g elem stem attribútuma tárol.
- **Szófaj meghatározása**: Ezt az eljárást szintén a HunMorph csomag morfológiai elemzőjének segítségével valósítjuk meg. Az indexállományok általános megvalósítása szótő szerinti nyilvántartást végez, ezért homonim szótövek esetén a különböző előfordulások egybevonódnak. Ennek elkerülésére rendszer az indexállományban [szótő, szófaj] párokat tárolunk. A szófaj információt a g elem pos attribútuma tárolja.
- **Szószűrő**: Ahhoz, hogy hatékonyan lehessen a keresések finomításához kulcsszavakat javasolni, elengedhetetlen az általános értelmű ún. funkció- v. stopszavak szűrése. Ezt egyrészt egy előre megadott szótár, illetve minták alapján valósítjuk meg. A szűrőn fennakadó szavak esetén a g elem sw attribútumát igazra állítjuk, ugyanis az indexelésnél ezekre a szavakra is szükség van, így nem törölhetőek.
- **Szótári névelemek felismerése:** Szótári névelemeknek nevezzük azokat a rögzített formájú kifejezéseket (többnyire tulajdonnevek), amelyek alapalakja a szövegben változatlan formában fordul elő. A névelemeknek lehetnek különböző előfordulási alakjaik (pl. Petőfi Sándor és Petőfi vagy Orléans-i szűz és Jeanne d'Arc), amelyek közül egyet kanonikus alaknak jelölünk ki, a többit pedig a kanonikus alak szinonimájaként kezeljük. Ezeket, ahogy elnevezésük is utal rá, egy szótárban tároljuk. A felismerésükhöz a HunMorph csomag morfológiai elemzőjét is felhasználó eljárást alkalmazzuk [6]. A névelemként felismert kifejezéseket e[xpression] címkével látjuk el.
- **Mondathatár-detektáló**: A modul a szövegek mondatszintű szegmentálását végzi, eredményét a keresési eredmények rövid legjellemzőbb részletének meghatározásánál alkalmazzuk. Működése szabályrendszer alapú: mondathatároló jelek előfordulásánál a szabályok alapján eldöntjük, hogy az adott jel ténylegesen mondathatárt jelöl, vagy sem. A szabályokhoz előjeles súlyértékeket rendelünk. Amennyiben egy adott mondathatár-környezetre több szabály illeszkedik, akkor a szabályok súlyának aggregálásával határozzuk meg a végső értéket. A feldolgozás során a szabálytár mellett rövidítéstárat is alkalmazunk. A detektált mondatokat s[entence] címkék közé tesszük.

A felsoroltakon kívül a DTD lehetőséget nyújt tetszőleges nyelvtechnológiai alkalmazás, pl. teljes morfológiai elemzés kimenetének felhasználására is. A rendszer továbbfejlesztése során ezen eljárásokat a keresés támogatásában nyújtott hatékonyságuk alapján integráljuk a rendszerbe.

A dokumentumoknak három különböző mértékben feldolgozott verzióját tároljuk a rendszerben. Az eredeti formátumú dokumentum mellett, a nyers XML dokumentumot is tároljuk, majd a Store modul a feldolgozott XML-t tárolja el, és amennyiben rendelkezésre áll, kategóriainformációt rendel hozzá. A dokumentumok különböző verzióinak elérési útvonalát a document elem megfelelő attribútumaiban tároljuk.

[5] http://mokk.bme.hu/projektek/szoszablya

A dokumentumok feldolgozása ezek után már numerikus alakban történik, az átalakítást a Tokenizer modul végzi el. A Datamine modul már ebben a formátumban keres gyakran ismétlődő tokensorozatokat, amelyeket egyedi indexszel lát el. Végül a belső reprezentációs alakot a Vector Generator komponens látja el, amely az irodalomban leggyakrabban használt vektortér alapú szózsák (*bag of words*) modellel[6] a dokumentumokból két vektort állít elő, egyet az indexeléshez, egyet pedig az osztályozáshoz. Az indexeléshez létrehozott vektor TF-IDF súlyozást alkalmaz és tartalmazza a stopszavakat is; míg az osztályozáshoz készített vektor entrópia-alapú súlyozást használ, és a stopszavakat nem tartalmazza [7].

Ezen a ponton válik el a különböző dokumentumok feldolgozási folyamata, hiszen a tanulóadatokat az osztályozó motor tanítására használjuk (Train), a többi dokumentum kategóriáját pedig a tanulóadatok alapján felépített osztályozási modellel (Inf) határozzuk meg. Ezután kerülnek a dokumentumok indexelésre, majd a felhasználói felület felé a Presenter modul jeleníti meg a szükséges kulcsszó- és kategóriainformációkat, immár nem numerikus (tokenizált), hanem szöveges formában.

4.2 Az aratórobot működése

Az arató modul feladata a célirányos, előre specifikált, illetve keret-megállapodással rendelkező partnerek portáloldalainak (összefoglalóan: gyűjtési tartomány) folyamatos követése, archiválása és címkézése. Az aratásnak jellegzetesen két fő funkciót kell kielégíteni:

1. Az oldalak folyamatos gyűjtését és háttértárra mentését (röviden szüretelés).
2. Az elmentett oldalak előfeldolgozását és szerkezeti címkézését.

A gyűjtés során adott, jól meghatározott forrásokat kell üzemszerűen meglátogatni. A gyűjtést egy ún. dæmon kell végezze – nevezzük a továbbiakban Aratónak –, amelyet indítani, azonnali gyűjtésre ösztönözni, valamint leállítani és késleltetni lehet.

Az Arató elindítja a letöltési folyamatot, amelynek bemenete a specifikált, gyakran meglátogatandó URL – tipikusan egy portál főoldala, vagy egy RSS-csatorna[7]. Az URL-t meglátogatva, az oldal tartalmát letöltve, rövid analizálás és címkézés után az oldal új dokumentumait le kell töltenie, és dokumentumarchívumba el kell helyeznie, majd a letöltött dokumentumból el kell távolítania a nem releváns részeket.

A megoldást nem kívántuk egyetlen tématerületre limitálni, így a specifikációban a legáltalánosabb megoldást választottuk. Ugyanakkor látni kell, hogy a releváns szövegek tartalmi elválasztása a dokumentum többi részétől nem biztosítható egyetlen univerzális algoritmus segítségével. A tartalmilag összefüggő, a tényleges információt hordozó szöveg kiválasztását legfeljebb nagyon mélyreható szemantikai elemzéssel lehetne a 100%-os pontosság közelébe juttatni. (Pontosság alatt értve azt, hogy a gazdasági hír, mint tartalom, teljes anyaga szerepel a kiválasztott szövegrészletben, és kizárólag az szerepel benne.) E tekintetben a statisztikai megoldások sem lehetnek segítségünkre, hiszen az összefüggő szövegek kiválasztására ma még nem ismert statisztikai alapú módszer.

---

[6] A modell a dokumentumokat a bennük szereplő szavak, illetve kifejezések (általánosan: tokenek) halmazának tekintik, ez tehát figyelmen kívül hagyja a tokenek pozícióját és sorrendjét a szövegben.

[7] Real Simple Syndication – http://blogs.law.harvard.edu/tech/rss

Felfigyeltünk ugyanakkor arra, hogy a cikk megjelenített és tényleges címének azonossága, illetve a cím ismerete esetén, valamint öt kulcsjellemző (dátum, szerző, cím, kivonat, szövegtörzs) egymáshoz viszonyított elhelyezkedésének ismeretében a ténylegesen releváns szöveg, mintegy 90%-os pontossággal azonosítható.

A gyakorlatban a különböző portáloldalak szerkezeti címkézését oldalanként egy-egy kis segédprogram – *plugin* – elkészítésével oldottuk meg. Ezek a segédprogramok az adott oldal szerkezeti jellegzetességeit figyelembe véve a HTML forrást fulldoc sémára illeszkedő XML-lé alakítják.

Mivel a segédprogramok elkészítésénél csak az egyes portálok aktuális jellegzetességeit ismertük, fel kellett készülnünk arra, hogy egy esetleges portál-motor váltáskor, vagy az oldal szerkezetének nagyobb léptékű változásakor a régi szerkezet figyelembe vételével készített segédprogram rossz kimenetet kezd produkálni. Ezért minden, az aratórobot által előállított XML-t megvizsgálunk, szintaktikailag ellenőrzünk. Egy oldal változása miatt elavult segédprogrammal előállított XML-ből többnyire hiányoznak a legfontosabb, kötelezően kitöltendő mezők, (pl. cím, szövegtörzs), ezért a fájl a szintaktikai ellenőrzésen fennakad. Az egyes portálok tartalmából előállított XML dokumentumok ilyen módon vizsgált tulajdonságairól statisztikát vezetünk, ami lehetővé teszi, hogy a figyelt portálok szerkezeti változtatásairól értesüljünk.

Ugyanakkor elképzelhető, hogy egy portál szerkezete úgy változik, hogy a szintaktikai ellenőrzés helyes marad, de a tartalom nem, pl. nem gazdasági témájú cikket gyűjtünk be. Az aratórobot az ilyen problémákat egyelőre nem tudja automatikusan kiküszöbölni.

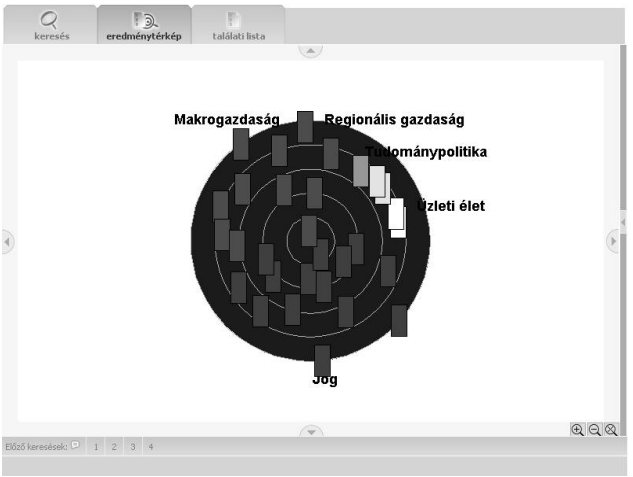
4.3 A felhasználói felület

A felhasználói felület a tervezésénél a funkcionális szempontok mellett a vizuális megjelenítésre is nagy hangsúlyt fektettünk. Terjedelmi okokból csak a találatok egyik megjelenítése formáját, az eredménytérképet tudjuk megmutatni. A szolgáltatás prototípusának beüzemelése és a szolgáltatás publikussá tétele 2006 év végére fog megvalósulni.

5 Hasonló kezdeményezések

Az elmúlt 5 évben az otthonukból internetezők tábora 6%-ról közel 30%-ra emelkedett és a havi legalább 1 órát internetezők száma elérte a 1,5 milliót. A felhasználók számának növekedése a tartalomipar expanzióját vonta maga után, amelyhez napjainkban az üzleti oldalú tartalom-előállítás mellett – a technológiai fejlődés és támogatás eredményeként – a felhasználó oldali tartalom-létrehozás is hozzájárul. Ennek a bővülésnek köszönhetően megnőtt az igény a keresőszolgáltatások iránt, amelyek kiépítésére a szolgáltatásból adódó üzleti lehetőség, a hazai online hirdetési piac dinamikus növekedése is serkentően hatott. Az üzleti oldalt megelőzve a tudományos szféra hamarabb felismerte az internetes keresésben rejlő tudományos kihívásokat és 2000-től – elsősorban az NKFP IKTA program finanszírozásában – tudományos kutatóműhelyek kezdtek különböző keresőalgoritmusok és -intelligenciák fejlesztésében. Az elmúlt 6 évben mind a tudományos, mind az üzleti szférában voltak törekvések olyan újfajta keresőeljárások kidolgozására, amelyekkel a szövegek gépi meg-

értésén keresztül próbálták a keresést pontosabbá és hatékonyabbá tenni, ám e kezdeményezések gyakorlati hasznosulása és hasznosítása jellemzően nem történt meg, így indokoltnak tartottuk olyan kutatás-fejlesztési projekt megvalósítását, amely nemcsak új eredményeket képes felmutatni a keresőintelligencia-kutatás területén, de az üzleti hasznosítást is képes biztosítani. A következőkben összefoglaljuk a hasonló hazai kezdeményezések.



3. ábra. A keresőszolgáltatás eredménytérkép oldala

Az *Információ és Tudás Tárház* (IKTA3-181/2000) projekt fő célkitűzése új intelligens tudás tárházak analízise, tervezése és megvalósítása volt, melyek lehetővé teszik a fejlett tudás- és üzletiinformáció-menedzsmentet [8, 9]. A projekt tudásalapú információ visszakeresési rendszert fejlesztett ki a pénzügyi szféra cégei számára, amely különböző forrásokból (Internet, belső adatbázisok, külső adattárházak, stb.) merít információt az alkalmazási környezet igényeinek megfelelően, majd ezt strukturált formában tárja a felhasználó elé.

A szavak hálójában (NKFP 0019/2002) projekt célja egy komplex internetes kereső/kérdező eszköz létrehozása volt, amely mind az Interneten elérhető online adatbázisok szöveges tartalmaiban – azaz a *mélyhálón,* a hagyományos keresőkkel nem indexelhető tartalmak összességén –, mind képek közti keresések terén új technológiákat tartalmaz [10, 11]. A képi keresés támogatására egy vizuális tezaurusz került kifejlesztésre, ami a képi tartalmak jellemzésére és indexelésére használható szöveges leírások, mint tartalmi kategóriák rendszere, strukturált szótára. A mélyhálótartalmakban történő keresésnél a rendszer támogatja magyar nyelven megadott teljes mondatok keresőkifejezésként történő használatát.

A *Szemantikailag szervezett lexikai hálózat és internetes tartalomkeresés* (IKTA5-123/02) projekt célja egy szemantikai szerveződésű, lexikai hálózat kifejlesztésére épülő, internetes tartalomkeresésre alkalmazható, újfajta technológia létrehozása volt. A projekt a célját a lexikai hálózat alapegységének tekintett, már kifejlesztett, ún. jelentésközpontok egy lehetséges kapcsolódásainak kutatásával és a kapcsolóelemek kiépítésével kívánta elérni (a jelentésközpont az azonos jelentés köré szerveződő, értelmezett természetes nyelvi kijelölők – szavak, szószerkezetek, mondatértékű kifejezések – egy struktúrában összefogott és kezelt egysége). A jelentés-

központok egymással való összekötésével, ún. linkek létrehozásával létrejövő, szemantikailag szervezett, kommunikatív lexikai hálót a projekt olyan kutatási szempontok alapján fejlesztett ki, hogy az képes legyen nyelvtechnológiai alkalmazásokban (természetes nyelvi szövegfeldolgozó-rendszerek, értelmezett információ-keresés elektronikus szövegekben és strukturált szövegtestekben, tartalomfigyelés, gépi fordítás, kontextus- és stílusérzékeny helyesírás-ellenőrző) értelmezetten és hatékonyan működni.

Az Országos Baleseti és Sürgősségi Intézet vezetésével valósult meg a *Tudásalapú magyar nyelvű szemantikus kereső rendszer kifejlesztése és alkalmazása a sürgősségi betegellátásban* (IKTA 00148/2002) projekt. A projekt tartalmazza az adatok statisztikai kontrollja mellett az adatok fogalmilag rokon csoportjainak (klaszterezés), valamint a logikai kapcsolatok extenzionális összefüggéseinek megállapítását. Ehhez a kidolgozott technológia a „tudásfeltárás" gépi tanulási és neuron-hálós eljárásokon alapuló módszereit ajánlja a „klasszikus" adatbányászati módszerekkel (drilling-up, drilling-down, stb.) együtt. A rendszer éles kipróbálása az Országos Traumatológia Intézet Információs rendszerébe ágyazva történt meg, ahol a szükséges orvosi ontológia rendelkezésre áll, és a megfelelő dokumentumok gyors megtalálása életbevágóan fontos. A kidolgozott tudásalapú keresési technológia teljesen általános, és széleskörűen használható könyvtárak, archívumok, orvosi, jogi és vállalati adat- és ismeretbázisok keresőmotorjaként, és mindazoknál a kereskedelmi alkalmazásoknál, amelyekben a célorientált keresés fontos szerepet játszik.

A *WebKat*[8] az első magyar fejlesztésű tématérképen alapuló modell, amelyet a Neumann-ház egy pályázat keretében hozott létre 2002-ben. Ez a tezaurusz a kereséseket a meglévő tárgyszórendszer relációinak tématérkép alapú vizuális megjelenítésével támogatja. A szolgáltatás nem az internetes tartalmakban, hanem a saját adatbázisában keres.

A *PolyMeta*[9] egy általános célú metakereső, amely lehetőséget nyújt tetszőleges számú Interneten keresztül elérhető kereső (adatbázis, forrás) egyidejű keresésére. Az eredményekből közös találati lista készül, amelyben az elemek fontossági sorrendbe rendezettek. Megjelenítésre kerül egy „tartalomjegyzék" is, ami segítséget ad a felhasználónak a témához tartozó résztémák, kapcsolódó fogalmak azonosításában, az azokhoz tartozó találatok megjelenítésében.

A Vipkereső[10] nevű legújabb kezdeményezés jelenleg még teljes funkcionalitásában nem érhető el, de az előzetes információk alapján a szabadszavas webkereső mellett képkereső, hírkereső és blogkereső funkciókat is nyújt majd. Várhatóan az Index szolgáltatásaként jelenik majd meg.

6 Összefoglalás

Cikkünkben ismertetjük egy tematikus keresőmotor felépítését és megvalósításának fő lépéseit. A megvalósított keresőszolgáltatás prototípusa az Interneten fellelhető magyar nyelvű gazdasági tartalmakat gyűjti, indexeli és teszi egy helyen kereshetővé.

[8] http://www.webkat.hu

[9] http://www.polymeta.hu/polymeta/meta.html

[10] http://www.vipkereso.hu/ – az alkalmazás a cikk írásakor még nem érhető el.

A találati dokumentumokat, amennyiben a tartalomszolgáltató ezt engedélyezi, a keresőfelületen keresztül is elérhetővé válik.

Köszönetnyilvánítás

A cikk a Gazdasági Versenyképesség Operatív Program GVOP-3.1.1.-2004-05-0130/3.0 jelű projektjének támogatásával készült.

Bibliográfia

[1] D. Tikk, Gy. Biró, and J. D. Yang. Experiments with a hierarchical text categorization method on WIPO patent collections. In N. O. Attok-Okine and B. M. Ayyub, editors, *Applied Research in Uncertainty Modelling and Analysis*, number 20 in Int. Series in Intelligent Technologies, pages 283–302. Springer, 2005.

[2] D. Tikk, J. D. Yang, and S. L. Bang. Hierarchical text categorization using fuzzy relational thesaurus. *Kybernetika*, **39**(5):583–600, 2003.

[3] Z. Alexin and D. Csendes, editors, *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY'05)*, Szeged, Hungary, December 8–9, 2005. SZTE, Informatikai Tsz.csoport.

[4] D. Tikk, A. Törcsvári, Gy. Biró, and Z. Bánsághi. Szótövező eljárások hatása magyar szövegek automatikus kategorizálásánál. In [3], pages 430–434.

[5] V. Trón, P. Halácsy, P Rebrus, A. Rung, E. Simon, E, and P. Vajda: morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In [3], pages 169–179.

[6] D. Tikk, F. P. Szidarovszky, Zs. T. Kardkovács, and G. Magyar. Ismert névelemek felismerése és morfológiai annotálása szabad szövegben. In [3], pages 190–199.

[7] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management, **24**(5):513–523, 1998.

[8] Cs. Dezsényi, P. Varga, T. Mészáros, Gy. Stratusz, T. Dobrowiecki: Ontológia-alapú Tudástárház Rendszerek. http://nws.iif.hu/ncd2003/docs/ahu/AHU-118.htm

[9] Cs. Dezsényi et. al: Tudásalapú információk kinyerése: az IKF projekt. In: Tudományos és Műszaki Tájékoztatás, 2004/5. http://www.neumann-haz.hu/tei/publikaciok/2004/biro_ref_ikf_hu.html

[10] D. Tikk, Zs. T. Kardkovács, et al: Natural language question processing for Hungarian deep web searcher. In *Proc. of the IEEE Int. Conf. on Computational Cybernetics (ICCC'04)*, pages 303–308, Vienna, Austria, Aug 30–Sept 1.

[11] D. Tikk, Zs. T. Kardkovács, and G. Magyar. Searching the deep web: the WOW project. In *Proc. of the 15th Int. Conf. on Intelligent Systems Development (ISD'06)*, Budapest, Hungary, Aug 31–Sept 2, 2006.