Horváth György – Horváth Zoltánné – Papp Attila

ADVISE adaptív automatikus kereső – miért más belül, mint kívül?

Az ADVISE[1] innovatív keresőeszközt vállalati, államigazgatási, ipari és kulturális intézmények információs vagyonának automatikus keresésére, összefüggéseinek felismerésére, analitikus statisztikák előállítására, a kereséshez használható keresőnyelvek automatikus előállítására, fogalmi vizualizációra terveztük. Továbbfejlesztésében hangsúlyt kap a taxonómiák automatikus előállítása, tématérképek létrehozása, valamint a portálintegráció, hiszen az információs források, a vállalati adatvagyonok áttekintése és intelligens felhasználása sem kellően hatékony, sem kellően pontos nem lehet ezen eszközök alkalmazása nélkül.

Az információkereséshez alkalmazható fogalmi rendszerező eszközök fejlesztése és karbantartása általános problémaként jelentkezik a tudásalapú társadalmi környezetben. Az információmenedzsmenttel foglalkozók körében komoly szakmai erőfeszítések folynak az automatikus osztályozás, automatikus fogalomalkotás és az automatikus keresés megvalósítása, a szükséges leírónyelvek és szabványok készítése érdekében. Az információforrások rohamosan növekvő tömegéhez rendező elveket és nyelveket létrehozni, alkalmazni és szinten tartani nem kellően hatékony csak manuális módon és eszközökkel, ugyanakkor jogosan vitatott, hogy ez a tevékenység teljesen megoldható lenne az emberi intelligencia értékelő és elemző beavatkozása nélkül. Könyvtári környezetben alapvető probléma a különböző, heterogén adatforrások integrált keresésének és áttekinthetőségének megoldása. A feladatot tovább nehezíti az adatforrások egy részének távoli elérhetősége, valamint az alkalmazható adatlekérés és -áttöltés vegyesen szinkron és aszinkron lehetősége.

A nem információszolgáltatással foglalkozó intézmények, üzleti vállalkozások körében sem más a helyzet. Idővel ugyan megszületett a felismerés és a szándék az információtárolás szabályosságának és a metaadatok egységesítésének erősítésére, viszont a felismerést követően létrejöttek azok a robusztus információtechnológiai megoldások, amelyek az adatok tárolásának és kinyerésének hatékony, biztonságos megoldásait kínálják (adattárház, middleware-eszközök), valamint az utóbbi évek slágere, az üzleti intelligencia-rendszerek. Nem véletlen az „intelligencia” megnevezés, ugyanis a kulcsszó alapján történő keresés csak felszíni eredményeket hoz, a mélyben rejlő és nem indexelt adatokat, az adatoknak a következtetéshez és a döntéshez szükséges összefüggéseit nem tárja fel. Ha kézi erővel történik a keresés, akkor a kulcsszó alapján kinyert információkat is ki kell egészíteni a találatok értékelésével (mennyiségi és tartalmi szűrés), a bennük rejlő információk esetleges szemlézésével, analízisével, szintézisével – speciális esetekben a döntés-előkészítési szintig. Ezt a feladatot információkutatók, piackutatók, tudásmenedzserek, speciálisan felkészült könyvtárosok végzik, és a tevékenységnél rendkívül fontos az elemzőképesség, a számítástechnikai felhasználói szakértelem, valamint a kiszolgált terület ismerete és mindezen képességekhez szükséges ismeretek folyamatos fejlesztése. Az általánosan rendelkezésre álló szolgáltatásoknál azonban nem kívánhatjuk meg a felhasználóktól az ilyen szintű képességet és háttérismereteket, ezért ezeket egyre inkább a számítástechnikai rendszerektől várjuk. Az ún. „intelligens” kereső- és elemzőeszközök a keresést az adattárak „mély” rétegeiben is végzik, és az eredményeket táblázatos, grafikonos összeállításban, például egy portálon mutatják akár szinkron megjelenítéssel, változáskezeléssel, kiemelve a „veszélyes” mutatókat. A portálon megjelenő automatikus monitorozó és jelentéskészítő eszközök között vannak már üzleti szimulációs szoftverek is (pl. Oracle Essbase), amelyekben lehet kísérletezni esemény és következmény vizsgálatával.

Mindezek mellett mégis hiányzik egy „front-end” típusú keresőeszköz, amely az összes létező forrásban keres, legyen az e-mail, adatbázis, adattárház, fájlszerver, internet vagy bármi más infor-

mációs vagyon, és igényünk az, hogy a rendszer bonyolultságának megfelelő technológia a háttérben intézze a mély rétegek keresését, és az eredményeket felhasználóbarát környezetben kapjuk meg.

Az ADVISE kialakításánál a fentieket kiemelten kezeltük: az automatikus, szemantikai és vizualizációs módszerek alkalmazása során egyszerű felhasználói felület áll rendelkezésre, a program adaptív, tanuló rendszerként támogatja a keresést több (integrálható) forrásból, különböző adattárakból az együttes információkinyerés céljából. A háttérben zajlik a rendszerek integrációján alapuló tranzakciók sorozata, amelyet a felhasználó nem érzékel.

Az ADVISE ismertetése előtt teszünk egy kis kirándulást a keresés, a vállalati intelligens keresők, az internetes keresés és az internetes szemantikus keresők területén, amelyek értékeit és tapasztalatait a fejlesztés során felhasználtuk.

A döntések általában nem ott születnek, ahol az információ rendelkezésre áll

Az információk jelentős része (egyes becslések szerint 80%-a) nem strukturált adatbázisokban jelenik meg, hanem különböző fájlokban (.doc; .ppt; .xls; .pdf; .mpp; .jpg; .html stb.), és bizonyos részük metabázisokban lévő strukturálatlan adat, amelyekre jellemző, hogy nincs egységes megjelenési felületük és közvetlen hozzáférésük (pl. adattárházak*).* A jelen tudásalapú és innovációvezérelt gazdasági környezetben a teljesítmény hatékonysága erősen függ az információk keresésétől, ezért flexibilis megoldásokra van szükség, alkalmazásuknál pedig kreativitásra. Az információkereséssel eltöltött – egyes felmérések szerint kb. 30% – munkaidő-hatékonysági tényező szempontjából rendkívül fontos, hogy ennyi idő alatt milyen eredményt tudunk felmutatni. Az üzleti intelligencia-eszközrendszer, technológia és eljárás a vállalatoknál az adattárakban, adatforrásokban, azok mély rétegeiben lévő információtartalom magas szintű kinyerését szolgálja, amelyek alapján következtetések vonhatók le – mára már automatikus, vizualizációs, adatbányászati, vagy értékelő, szintetizáló és analizáló megoldásokkal. Az ADVISE ezen eszközök körébe tartozik, mindamellett a hagyományos információszolgáltató intézmények számára is megoldást kínál.

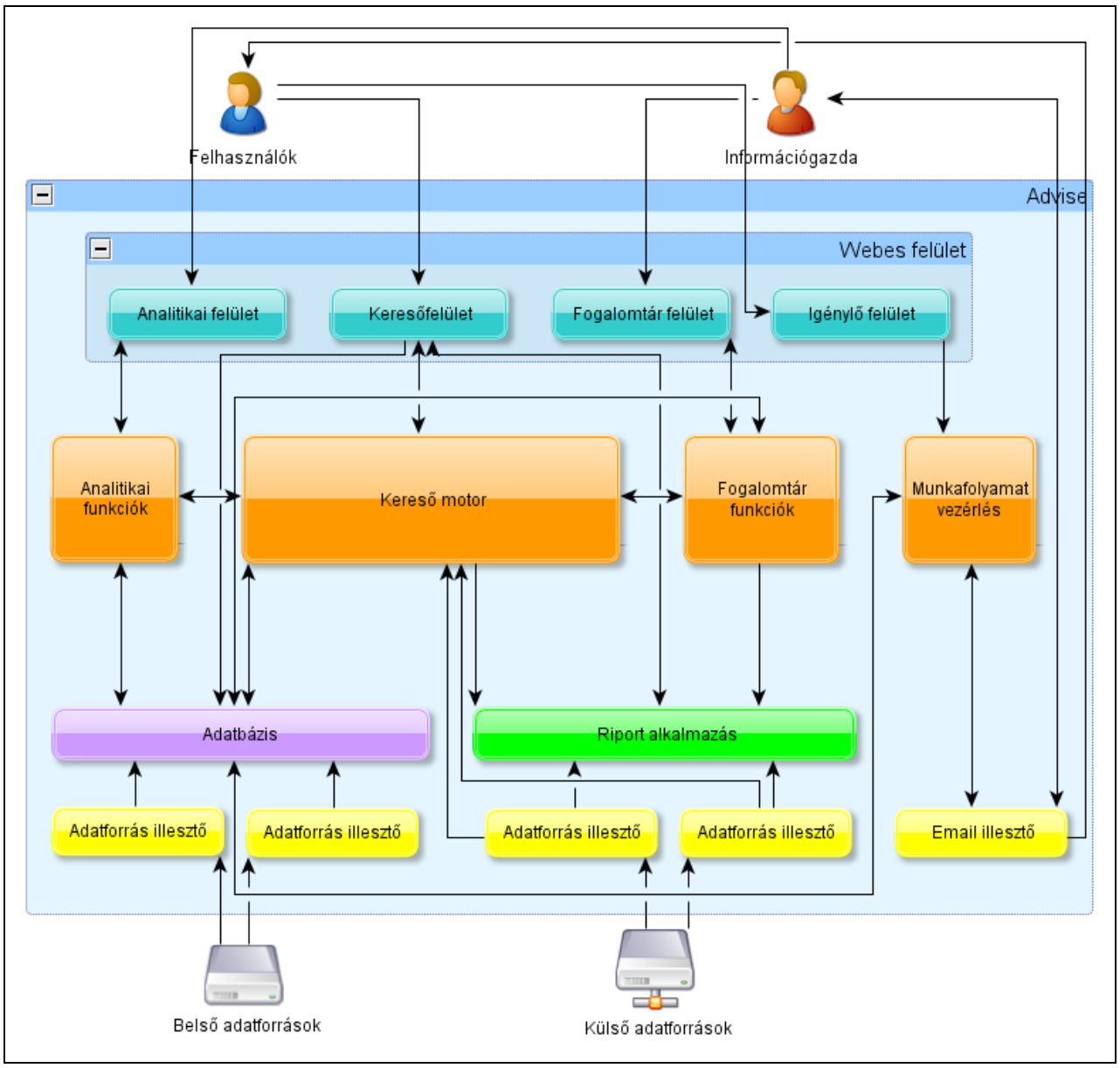
A teljesítménykényszer és az információmenedzsment

A teljesítményünk szervezése, növelése forradalmáról beszélnek egyes szakírók[2], akik szerint ezért fog egyre növekedni az elemző információszolgáltatók, „knowledge workerek" száma, akik lefedik majd a pénzügyi, az egészségügyi, a média és egyéb frekventált szakterületek munkavállalóinak 25%-át. Ha netán kétkednénk ebben a jóslatban, akkor is valószínű, hogy ezzel párhuzamosan, egymást erősítve egyre nagyobb figyelem hárul az innovatív információkereső megoldásokra. Az információforrások növekedését nem lehet megállítani, ezért az információs vagyon jobb kihasználása egyre több erőforrást fog lekötni a kutatásfejlesztés, és az emberi intelligencia szempontjából is. Nyilván nemcsak az elérés, hanem a források alapján végezhető műveletek lehetősége (analízis, szintézis, következtetések, transzformációk, döntés-előkészítés) jelentik azt az elméleti, tudományos, vagy üzleti előnyt, amelyből egyéb eredmények fakadnak.

A könyvtárak számára az üzleti intelligenciaeszközök a robusztus technológia miatt igen drágák, amelyek elterjedése a nagyvállalatokat követően csak néhány éve jellemző a közép- és kisvállalatoknál, azonban az ADVISE kereső reális lehetőséget kínál a könyvtári szektornak is arra, hogy a keresési metodika készségeinek a birtokában talán a könyvtárosok használhassák ki a leginkább a rendszer adottságait (*1. ábra*).

Keresés az automatikus és vizualizációs módszerek igénybevételével

Az internetkultúra környezetében a n*agy tömegű elektronikus információk keresésénél a tartalomszolgáltatók és a felhasználók oldaláról egyaránt jelentkezik az automatikus keresés és az automatikus fogalomalkotás lehetőségének, az elkészült kereső- vagy tartalomosztályozó rendszerek gyors karbantartásának, és a rugalmas szerkezeti megoldásoknak, módosításoknak az igénye.* Az egyes szakterületeken különböző fogalmi, osztályozási rendszerek készülnek és állnak rendelkezésre a hierarchia és a tudományos megalapozottság különböző szintjein. Megfigyelhető, hogy vállalati környezetben erős a pragmatikus megközelítés, amely a tudományos nyelvi elemzés és rendezettség helyett a teljes vállalat gyakorlati, közérthető szaknyelvi megközelítésére alapoz, és a fogalmi



1. ábra **Az ADVISE igénylési munkafolyamatok támogatása**

rokonságokat alacsony szintű struktúrában tükrözi, amelyet könnyebb átfordítani a számítástechnika nyelvére. Nagy jelentősége van a kombinált megoldásoknak, a rendszerek átjárhatóságának, a szemantikai keresőknek, és az összefüggéseket grafikusan is megjelenítő fogalmi vizualizációnak. A fogalmi vizualizációnak az összefüggések tükrözése, elemzése terén különös előnye, hogy a grafikus megjelenítés mellett grafikus eszközökkel módosíthatók az élek és a csomópontok, ezért a szerkezet könnyen karbantartható a használat során, ahogy ezt az ADVISE kereső esetében is tapasztaljuk.

A szövegösszefüggés szerepe – a szemantikai tér

Az információkeresésnél nem szavakat, hanem témákat keresünk, amelyek valamilyen szövegösszefüggésben, szerkezetben, nem „string”-ként jelennek meg, és „szemantikai térnek” is nevezhetők[3]. Az információkereső segédeszközök kialakítása során ezért megnőtt a szerepe a szövegösszefüggések tükrözésének, azonban ez a módszer nem lehet annyira alapos, hogy ne legyen közérthető, veszélyeztetve a használat gyorsaságát vagy könnyű elsajátítását. A szemantikai webhez kap-

csolódó kutatások, leírónyelvek is a szövegekben lévő szemantikai és szintaktikai összefüggések leírására törekednek, mert ezeket az összefüggéseket egy gépi rendszer számára formalizálni kell ahhoz, hogy az automatikusan felismerje. A leírónyelvek alkalmazása azonban messze áll jelenleg az általános használatbavétel lehetőségétől a szintaktikai és szemantikai formális nyelvi elemzés nehézsége, a leírás számítástechnikai átfordításának bonyolultsága miatt. A széles körű elterjedés megértésbeli gátjai rávilágítanak arra, hogy menynyire csodálatos az emberi intellektus, amely könnyedén mozog ebben a térben, ám nyelvezetünk formális visszatükrözése eléggé megoldhatatlannak látszik még akkor is, ha például a humor, a gúny, a metafora stb. tükrözésétől eltekintünk.

Az ADVISE mint szemantikus kereső újdonságértéke

A szemantikus keresőknek nevezett eszközök nem régen jelentek meg a piacon, és elsősorban az interneten megjelenő információk intelligens feltárását szolgálják. A *Google* és a *Microsoft* egymással versengve törekednek a keresés finomítására, a keresés vertikális lehetőségeinek kiterjesztésére. (L. *Google Universal Search, Google Analytics, Google Squared, Bing*). A keresőkben előre beállított kategóriákat adnak meg, amelyek a keresés típusa, az információ megjelenési formája, időbeli megoszlása és egyéb szempontok szerint szűrik a találatokat. A Google Squared táblázatokba tömöríti a találatokat, módosítható sorokkal és oszlopokkal, a táblázat elmentési lehetőségével. A kereső igen jól működik az ún. webmarketing-szolgáltatásokban, azonban a hazai tematikájú információk esetében meglepő és nehezen igazolható összefüggéseket mutathat a táblázatba foglalt információegyüttes.

Vannak statisztikai, vagy szövegkörnyezeti, jelentéstani, tudásalapú, vagy ontológiai összefüggésekre épülő keresők (*TextWise, Radar Networks Twine, Hakia, Wolfram Alpha, Jebol, WOWD* stb.). A Wolfram Alpha több beépített modellt használ, amelyek számos valós területhez kapcsolódnak, és nem a webmarketing, hanem a tényinformációk kereséséhez ajánlják.

Kutatások folynak különböző nemzetközi digitális könyvtári projektekben, ám ezek inkább az ontológiai nyelv szintjét célozták meg a keresőkategóriák és keresőrendszerek fejlesztésével, ontológiai leíró nyelv alkalmazásával. Külön meg kell említeni az *Autonomy* rendszert[4], amelyhez a leginkább hasonlít az ADVISE. Mindkettő ún. „tanuló” rendszer, tudásportál-funkciókat támogat, megismertet az egyes témák gazdáival, automatikusan hozza létre a témacsoportokat, olyan adattárakban végzi a kereséseket, amelyek egyébként nem kommunikálnak egymással, jól kezeli a többnyelvű megoldásokat, feltárja a rendezett és strukturálatlan információk kapcsolatát, és automatikus taxonómia előállítását is lehetővé teszi.

Megemlítendő, hogy a szemantikus keresők és a szemantikus web kapcsolata nem következik a jelen keresők természetéből, így az ADVISE megoldása sem a szemantikus web eszközrendszerébe tartozik jelenleg. A rendszer filozófiájában a rugalmas és gyakorlati alkalmazás fejlesztését tartjuk követendőnek távlatilag is, és kiemelten fontosnak tartjuk az egyes szektorokhoz való alkalmazkodást.

Összességében megállapítható, hogy az újabban megjelent szemantikai alapú keresőrendszerek stratégiában és megoldásban is jelentősen eltérnek az ADVISE rendszerétől, amely nem csupán a web keresésére, hanem több, különböző típusú információforrás egyidejű keresésére szolgál. Az eddig megismert hasonló célú keresők egyes elemeihez kimutathatók hasonlóságok (asszociációs, automatikus keresés, analitikus statisztika készítése stb.), azonban egyik említett keresővel sem rokonítható sem célját, sem az alkalmazott technológiát, sem a keresés módszereit tekintve.

Tématérkép előállítása automatikusan – az ADVISE innovatív eszközrendszerrel

A tématérképnél minden téma valójában egy szinonimacsoportot képez, amelyet egyetlen megnevezés képvisel. A témahely tárgyának formális deklarálására és azonosítására a tárgyi osztályozás technikáját használja, de a tématérkép megfordítja az általunk ismert információfeltáró folyamatot; itt nem a dokumentumból, hanem a témától jutunk el az objektumhoz, az információforráshoz.

A megnevezés ad egy helyet (scope), ahol megjelenik a téma egy halmaza, amely egy tartalmat képvisel. A megnevezések lehetnek azonos alakúak is, azonban a típus, az előfordulás és a kapcsolat pontos értelmezést ad a megnevezésnek, például: Paris (mitológiai alak); Paris (város). A tématérkép a többnyelvű információk szolgáltatását is támogatja, a felhasználó saját nyelvén választhatja

ki a megnevezést, és a rendszer nem arra figyelmezteti, hogy egy másik, preferált kifejezést alkalmazzon, hanem belső kapcsolatai alapján „érti" a kérdést. A tématérkép hátterében a tudásintegrációra fejlesztett ISO-szabványt használunk.

Az ADVISE eszközzel célunk, hogy automatikusan állítsuk elő néhány szakmai terület tématérképét és megvizsgáljuk az azonos vagy hasonló elemek importját és integrációját más rendszerekhez.

Különböző keresési módszerek és eszközök

Mikor van szükség keresésre? Az egyszerű válaszban: „amikor nem találok valamit" összetett feladatcsoportok határozhatók meg, amelyeknek minden összetevőjére figyelnünk kell. (Válaszidőcsökkentés, adatszerkezet, adatforrás-indexelés, időskálán való elhelyezés, elavulás, kérdés időpontja, kulcsszókezelés, szinonimák, ragozott formák, információkereső nyelvek stb.)

A keresés során a válasz minősége növelhető azzal, ha tudjuk, ki kérdez. Egy keresőt a keresési szokásai alapján tudunk leírni. Ezt az információt használják ki az *adaptív keresők*, amelyek csoportjához az ADVISE is tartozik[5]. A következőkben a hagyományos szöveges alapú keresőktől a szemantikus keresőkön át az internetes keresők specialitásait érintve jutunk el a vállalati keresők világához. Mindegyik területnél áttekintést adunk az adott terület főbb kihívásairól, jellemzőiről, illetve kapcsolatáról az egyéb területekkel – amely tulajdonságokat az ADVISE innovációnál figyelembe vettük.

Szöveges keresés

A szöveges keresők a keresési problematikát a *hol* kérdésre összpontosítják. A keresés tárgyát szövegrészletek alkotják, amelyek előfordulását a rendelkezésre álló adatforrásokban nagy hatékonysággal meg tudják mondani. Ennek a megközelítésnek előnye az egyszerűség, a nagy teljesítmény, valamint a kiforrottság. Ugyanakkor kétségtelen hátrányként kell megemlítenünk a keresés többi tényezőjét, miszerint a fogalmi kapcsolatok hiányában, az idő és a kérdező ismerete nélkül a válaszok sok esetben irrelevánsak vagy pontatlanok lesznek. A szöveges keresőrendszerek döntő többségében a következő architekturális modulokat tartalmazzák:

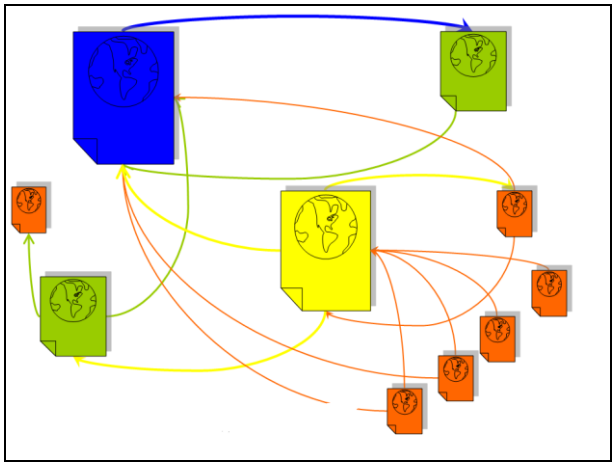
- *Pásztázás* (ún. *crawling*)[6]: a rendelkezésre álló adatforrások bejárását vezérelni szükséges. Ennek oka, hogy figyelembe kell venni az adatforrások hasznosságát, redundanciáját, valamint azt, hogy sok esetben nem is járható be a teljes halmaz, ezért szükséges algoritmizálni a bejárható szelet meghatározását.
- *Elemzés*: a bejárás során érintett adatforrásokat elemezni szükséges, hogy olyan reprezentációt készítsünk, amelyet egységesen és hatékonyan előkereshetően tudunk ábrázolni. Tipikus feladatok a formátumkonverzió, kis/nagybetűk kezelése, stopszavak kiküszöbölése, szótövezés, nyelvfelismerés, kivonatkészítés.
- *Tárolás*: az indexált adatok hatékony tárolása kulcsfontosságú, hiszen ez határozza meg döntő részben a keresés sebességét. Itt a relációs adatbázisok mellett nagy szerepet kapnak a speciális kívánalmakat is kezelő egyedi implementációk.
- *Keresés*: a keresés során a felépített adatszerkezet funkcióit használva meg kell határozni a találatokat, azok értékét-sorrendjét, valamint tipikusan valamilyen kivonatolt tartalmát.

A pásztázó algoritmusok feladata, hogy adott számítási kapacitás mellett biztosítsák a megoldandó feladat által meghatározott optimumot a következő paraméterek esetében: mennyiség, aktualitás, pontosság. A fenti architekturális határok rugalmasak, a keresési funkcionalitás szempontjából kategorizálják egy rendszer komponenseit. Példaként gondoljunk arra, hogy egy adatbázisban történő keresés során is csak akkor tudunk hatékonyan lekérdezni, ha található index a kérdezett információt tartalmazó oszlopokhoz. Ez esetben az indexelés természetesen nem ütemezetten, bejárás által vezérelve történik, hanem automatikusan az adatbázis-műveletek közben a háttérben.

PageRank algoritmus

A *PageRank* algoritmus[7] (*2. ábra*) az egyik legismertebb módszer az internetes keresők körében. Alapötlete az, hogy rendeljünk minden oldalhoz egy rangot, amely azt tükrözi, hogy az adott oldal menynyire fontos. Ennek alapján már tudunk szelektálni a beláthatatlan mennyiségű oldal között, hogy melyeket érdemes indexálni. A kérdés csupán az, hogy az oldal fontosságát hogyan lehet megállapítani. A *PageRank* válasza az, hogy egy oldal annál fontosabb, minél több fontos oldal mutat rá. Formálisabban megfogalmazva: egy oldal rangja a rá mutató oldalak rangjának súlyozott összege.
A fenti definíciót alkalmazva egy *N* darab oldalból álló webrészletre meg lehet határozni az egyes oldalak rangját. A valós implementációkban

ugyanakkor iteratív módszereket szükséges alkalmazni a rang meghatározására, hiszen az oldalak száma nagyobb annál, mint hogy direkt megoldó algoritmust lehetne alkalmazni. Ennek módszeréről az irodalom bőven ad tájékoztatást.



2. ábra **PageRank működési séma**

Elemzés

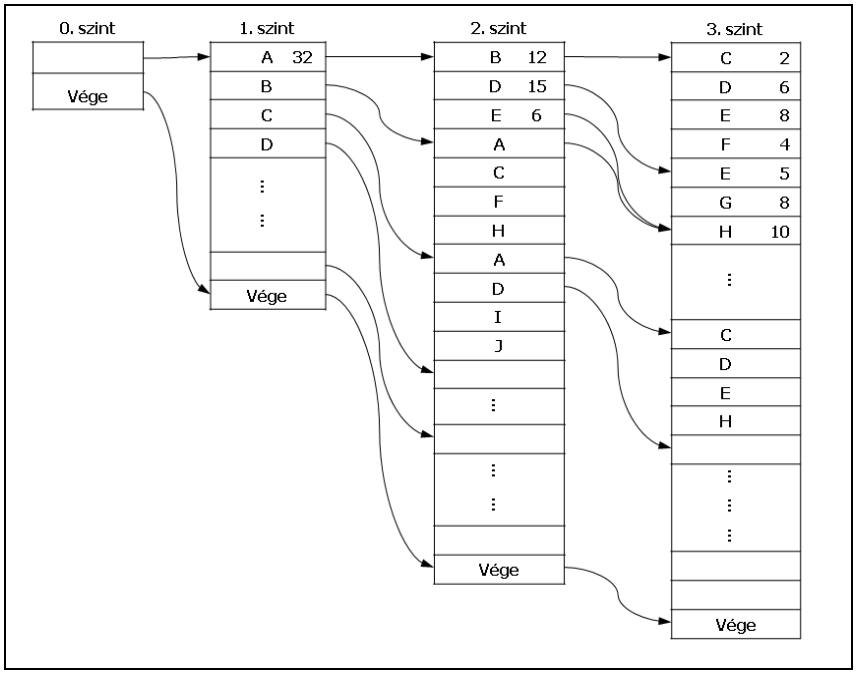
Az adatforrások elemzését az eredmény tekintetében a következő két csoportra bonthatjuk:

- *statikus elemzés*: az adatforrás tartalma a bejárás pillanatában lekérdezésre kerül, majd az elemzést ezen az információhalmazon végezzük el;
- *dinamikus elemzés*: az adatforráson keresztül elérhető információk leírása – ún. metaadatok – a bejárás pillanatában lekérdezésre kerül, azonban a tényleges információk lekérdezése és elemzése keresési időben történik.

Statikus elemzés

A lekérdezett adatokat több lépésben szükséges feldolgozni, hogy jól kereshető reprezentációhoz jussunk. A leggyakoribb feldolgozási lépések a teljesség igénye nélkül: normalizálás, stopszavak kiiktatása, szótövezés, nyelvi felismerés, szöveghasonlóság-elemzés, képi feldolgozás. A nyelvi detekció nem mindig történhet dokumentum-metainformációk alapján, mert azok sok esetben hiányosak vagy hibásak. Ezért szükséges magát a szöveges tartalmat alapul venni.

Az egyik legelterjedtebb módszer erre a *trigram-statisztika* készítése. A trigram egy betűhármas, a trigramstatisztika pedig ezen betűhármasok előfordulásának gyakorisága egy szövegben (*3. ábra*).



3. ábra **Trigramstatisztika készítése**

Számos esetben ütközünk keresés során abba a nehézségbe, hogy adathibákból, elgépelésekből vagy akár pusztán marginális dokumentummódosításokból kifolyólag egy kérdésre helyes válasznak tekinthető dokumentum semmilyen formában nem tartalmaz szavakat a kérdésből, még szótő szintjén sem. Ilyenkor ad segítséget a szövegek hasonlóságának elemzése. Ezek a módszerek nagyrészt ki tudják küszöbölni a fenti okokból keletkező kismértékű eltéréseket a szövegekben.

Szövegek távolságának meghatározására számos gyors módszer ismeretes, ezek közül az egyik a *Levensthein-távolság*[8]. A *4. ábra* ennek számítását illusztrálja:

		B	A	N	Y	A	B	Á	N	Y	Á	S	Z	A	T
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13
D	2	2	2	2	3	4	5	6	7	8	9	10	11	12	13
A	3	3	2	3	3	3	4	5	6	7	8	9	10	11	12
T	4	4	3	3	4	4	4	5	6	7	8	9	10	11	11
B	5	4	4	4	4	5	4	5	6	7	8	9	10	11	12
Á	6	5	5	5	5	5	5	4	5	6	7	8	9	10	11
N	7	6	6	5	6	6	6	5	4	5	6	7	8	9	10
Y	8	7	7	6	5	6	7	6	5	4	5	6	7	8	9
Á	9	8	8	7	6	6	7	7	6	5	4	5	6	7	8
S	10	9	9	8	7	7	7	8	7	6	5	4	5	6	7
Z	11	10	10	9	8	8	8	8	8	7	6	5	4	5	6
A	12	11	10	10	9	8	9	9	9	8	7	6	5	4	5
T	13	12	11	11	10	9	9	10	10	9	8	7	6	5	4

B A N Y A B Á N Y Á S Z A T

A D A T B Á N Y Á S Z A T

\+ = o o o = = = = = = = =

4. ábra **Levensthein-távolság számítása**

Mint látható az ábráról, az algoritmus lineáris időben futtatható és kiküszöböli a fajlagosan kis eltéréseket két szövegrészlet között.

A szöveges dokumentumokban sok esetben hordoz kulcsfontosságú információt a kép. Természetesen a kép tartalmának általános meghatározása nem reális feladat, ugyanakkor számos alkalommal nyílik lehetőség hasznos információk felderítésére. Ehhez elegendő pusztán két kép hasonlóságának felismerése – melyre már léteznek hatékony algoritmusok.

Dinamikus elemzés

Dinamikus elemzést szükséges alkalmazni akkor, ha az elemzendő információhalmaz mérete irreálisan nagy és/vagy gyorsan változó. Ez tipikusan fennáll adatbázis-tartalmakra, melyekre például az internetes keresés témakörében a „mély web” terminológiát szokás alkalmazni, utalva arra, hogy az információ felszínre hozható ugyan az internetes felületen keresztül, de ehhez kéréseket kell specifikálni az adatbázis felé – legtöbbször valamilyen űrlap formájában. Felmerül a kérdés, hogy ha az információ közvetlenül nem indexálható, akkor mit lehet kezdeni az ilyen adatforrásokkal. A választ a metaadatokban találjuk, azokban az adatokban, amelyek az üzleti értéket hordozó adatokat írják le. Ezeket értelmezve és indexálva tudjuk megállapítani, hogy egy adott kérdést érdemes-e feltenni az adott adatforrásnak – keresési időben – vagy nem.

Keresési mátrix

A keresési feladatot a következőképpen lehet a legegyszerűbben matematikailag szemléltetni. Képzeljünk el egy nagy mátrixot – ezt a következőkben keresési mátrixnak fogjuk nevezni –, amelynek sorai a kérdések, oszlopai pedig a válaszok. A mátrix egyes celláiban egy mérőszám áll, amely azt fejezi ki, hogy az adott kérdésre az adott válasz mennyire jó. Könnyen belátható, hogy a mátrix általában igen nagy, ugyanakkor igen ritka. Előbbi adja a keresés egyik technikai nehézségét, utóbbi pedig a megoldást. A mátrixot a ritka mátrixreprezentációnak megfelelően célszerű tárolni, azaz nem tárolunk le minden elemet, hanem minden sorból/oszlopból csak a nem nulla elemeket jegyezzük meg, pozíció szerint.

A feladatot nehezíti, hogy az indexálás során mindig egy oszlopban található adatok jelennek meg egyszerre, a keresés során pedig egy sor adataira vagyunk kíváncsiak. Mivel a célfüggvény az, hogy a keresés gyors legyen, az elemeket soronként csoportosítva kell tárolni, ami indexálási időben pontosan annyi egység módosítását jelenti, ahány kérdésre releváns választ találtunk. Az implementációkban ezért fontos szerepet kap a sorok elérési idejének minimalizálása.

Relációs modell

A keresési mátrix elemei egyszerűen betölthetők relációs adatbázisba például egy – *kérdés, válasz, relevancia* – adatszerkezetben (*5. ábra*). Indexet téve a *kérdés* oszlopra a lekérdezések hatékonyak lesznek.

Kérdés/Válasz	http://iqwiki	file:///IQSYS-Mukodes/BIT	https://y-iqsyssps
üzleti intelligencia	0,5	1	0,3
RDBMS	0,8	0,5	0,4
yEd	0,9	0,6	0
BI	0,6	0,9	0,7

5. ábra **Relációs modell**

A megközelítés kiválóan alkalmazható kisméretű keresőrendszerek esetén. Nagyméretű rendszereknél az index mérete igen nagy lehet, ami performanciaproblémákhoz vezethet. Ezt orvosolják a következőkben említésre kerülő módszerek:
A hashmap[9] egy olyan adatszerkezet, amely kulcsértékpárok között definiál hatékony leképezést a kulcsok alapján képzett ún. *hash*értékek felhasználásával. Hatékony működésének feltételei a következők:

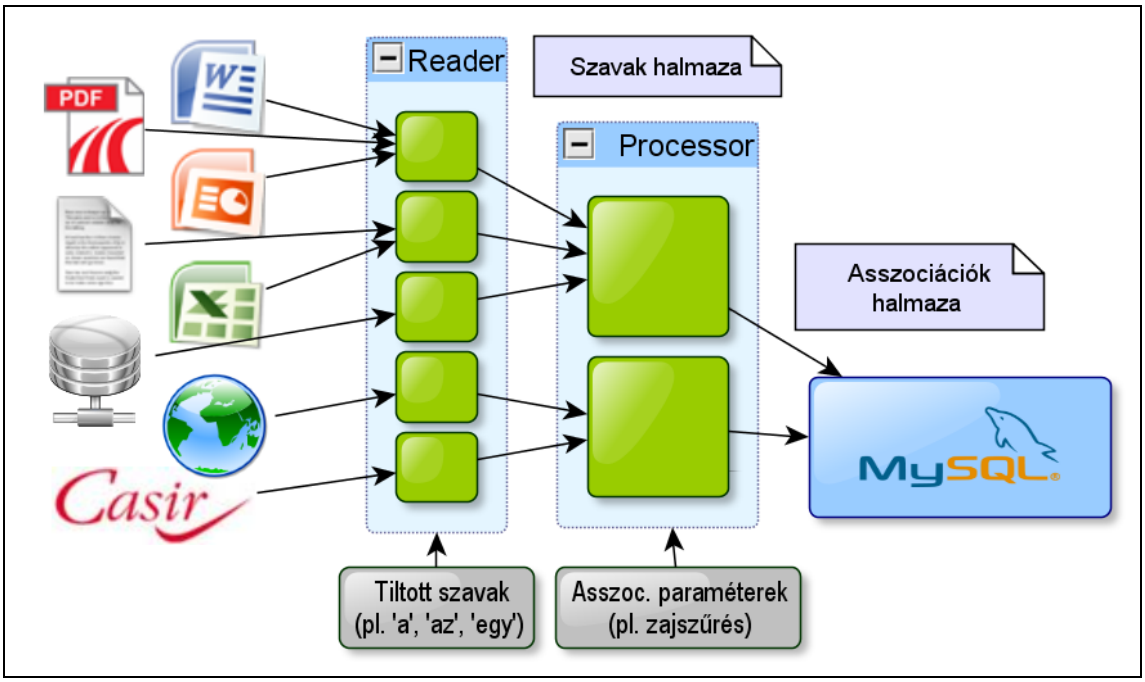
- a *hash*értékek képzésére szolgáló *hash*függvény kellően homogén módon szórja szét a kulcsokat az értékkészletben,
- a *hash*területhez rendelkezésre álló tárhely összemérhető legyen a várható elemárok számával.

A fentieket biztosítva elmondható, hogy ez az adatszerkezet konstans időben tud választ adni a kérdésekre, egy kérdés-válasz kulcsértékpár mentén történő előzetes felépítés esetén. Alkalmazásának a *hash*terület nagysága tud határt szabni, amelyre az *elosztott hashmap*technika[10] nyújt megoldást, amelynél az adatok több számítógépen vannak elosztva abból a célból, hogy a teljesítmény növelhető legyen.

Index a keresési mátrixhoz

A keresés alapja a keresési mátrixra felépített indexállomány. Ez teszi lehetővé, hogy a keresőrendszer méretezése során figyelembe vett adatforrás-mérettartományban a válaszidők egy előre meghatározott konstans alatt maradjanak. A keresőarchitektúra feladata az indexállományok karbantartása (*6. ábra*). Az új adatforrások frissítése, az elavultak öregítése. Adaptív rendszerek esetén itt szükséges figyelembe venni a megtanult információkat.

Adott adatforrás-mennyiség és hardverkapacitás mellett a keresési válaszidő tovább növelhető, ha a gyakori kérdésekre adott választ *gyorsítótárba* helyezzük. Ezzel átlagos válaszidő-követelmény esetén erőforrást tudunk felszabadítani a rendszerben más feladatokra – például mélyebb elemzés, differenciáltabb keresés.



6. ábra **Indexálás az ADVISE-ban**

Szemantikus keresők

A szemantikus keresők működése a rendelkezésre álló kereshető tartalmak értelmezésén, jelentésének felderítésén alapul. Könnyű belátni, hogy ez a koncepció relevánsabb találatokhoz és gyorsabb keresési ciklusokhoz vezet, ha a háttérben álló tartalomértelmezés adekvátnak tekinthető. A fentiek következményeképpen a szemantikus keresők legfontosabb tulajdonsága a taxonómiaépítés módszere, amely alapvetően meghatározza a keresőrendszer használhatóságát.

Taxonómiaépítés

Az információk értelmezésének alapját az ún. taxonómiák adják[11]. A taxonómia egy *fogalomrendszer*, amelyben a fogalmak között *relációk* vezetnek, ezzel hozva létre a szükséges kapcsolati rendszert a kereséshez. A fogalmak között húzódó kapcsolatok attribútumait a taxonómiaépítő módszertan határozza meg. Tipikus kapcsolatok a *szinonima, kategória,* illetve *tulajdonság*. Így könnyedén megfogalmazható, hogy például a „*google"* fogalom „*kategóriája"* a „*kereső"* fogalom.

A fogalmak és relációik meghatározása számos módon történhet a manuális – ember által végrehajtott – taxonómiaépítéstől kezdve a hibrid megoldásokon át a tisztán gépi hálózat kialakításáig. A manuális és az automatikus taxonómiaépítés tulajdonságainak összevetését mutatja az 1. táblázat.

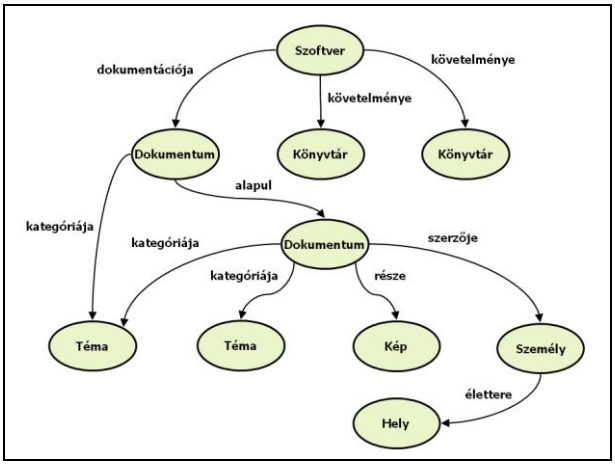
1. táblázat
A manuális és az automatikus taxonómiaépítés tulajdonságai

	Manuális taxonómia-építés	Automatikus taxonómia-építés
Sebesség	Lassú	Gyors
Minőség	Magas	Közepes
Erőforrásigény	Nagy	Kicsi/közepes
Karbantarthatóság	Nehézkes	Triviális
Felhasználhatóság	Univerzális	Speciális

A manuális és az automatikus taxonómiaépítést összevetve elmondható, hogy mind a mai napig kiegyensúlyozott a verseny és nincs általánosan kiválasztható „jó" irány. A megoldás általában a két módszer vegyítése, melynek módja erősen alkalmazás- illetve területfüggő. Példaként említhetjük, hogy a honlapokhoz kapcsolt metaadatok és a weboldalak kapcsolatai alapján könnyen lehet gépi módszerrel taxonómiát építeni, azonban az adatok hiányossága miatt ezt sokszor további intelligenciával kell kiegészíteni: címek megállapítása, szótövezés, illetve végső esetben manuális korrekció segítségével.

Szemantikus web[12]

Tim Berners-Lee, a világháló atyja meg van győződve arról, hogy a jövő világhálója szemantikai alapokon fog működni. A jövőkép szerint a naponta több millió új oldal megjelenéséhez a későbbiekben ezzel összemérhető mennyiségű szemantikai információ fog társulni. Ahogy a világháló természetes nyelve a HTML (*Hyper Text Meta Language*), úgy a szemantikus információé az RDF (*Resource Description Framework*), illetve az OWL (*Web Ontology Language*). Az RDF erőforrások – esetünkben tartalmak – egyedi és relációs leírására alkalmas nyelv. Az OWL pedig ezt egészíti ki magasabb szintű osztályozási és relációs információk leírásával. Alkalmazásukról az irodalomban bőségesen találunk leírást[13], itt csak példaként említjük meg egy személy nevének és a hozzá kapcsolódó információknak a kapcsolását elérhetőségek, naptár, honlap, illetve referenciákkal (*7. ábra*):



7. ábra **Metaadatok szemantikus kapcsolatai**

Visszautalva a manuális és az automatikus taxonómiaépítés összevetésére: a világháló méretéből kifolyólag a manuális taxonómiaépítés erős hátrányban van az automatikus módszerekkel szemben. Egyelőre nem látszik kellő mértékűnek az RDF és az OWL elterjedése ahhoz, hogy a szemantikus web elképzelése ilyen módon megvalósulhasson. Éppen ennek köszönhető, hogy olyan éles a verseny, és dinamikus a fejlődés az auto-

matikus taxonómiaépítő eszközök és a szemantikus keresők piacán.

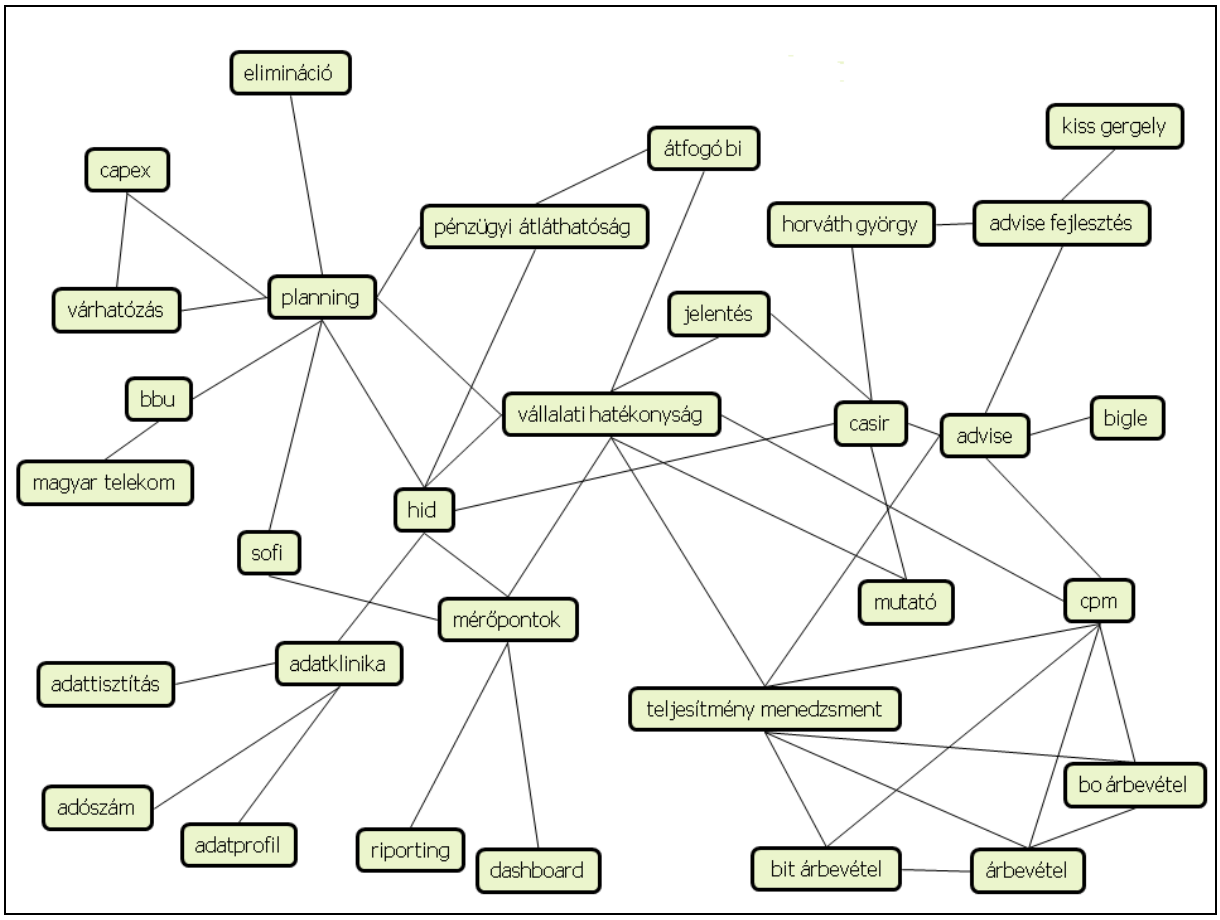
Szemantikus keresési metodika

A szemantikus keresők jelentős része nemcsak adatforrásokat kínál fel találatként, hanem kapcsolódó kereséseket, fogalmakat és témaköröket egyaránt. Ezzel segítik, orientálják a felhasználót a kívánt eredmény irányába. Azaz ilyenkor a keresőmotor nem a szöveges egyezések alapján ad csak találatokat, hanem megpróbálja felderíteni azt, hogy a felhasználó mire gondolhatott és az hogyan, milyen formában található meg az adatforrásokban.

A felajánlások algoritmikus alapja sokféle lehet. Felépített taxonómiával rendelkező rendszer esetében természetesen a taxonómia adja a kapcsolódó fogalmakat és erőforrásokat, azokat pusztán rangsorolni és megjeleníteni kell. Taxonómia hiányában a keresési szokások tanulása valamint a statisztikai, illetve szövegbányászati algoritmusok tudnak segítséget nyújtani.

Vizualizáció

Egy taxonómia igen jelentős méretű lehet, megjelenítésének módja és minősége már egy speciális szakterület esetén is kritikussá válhat, hiszen adott esetben ezen múlik, hogy a felhasználó kellő időben észreveszi-e a számára szükséges információt. Az XML-nek mint technológiafüggetlen információhordozó-formátumnak kiemelt jelentősége van az egyes tudásreprezentációs formák közötti átvitel szempontjából. Így nyílik lehetőség például a 8. ábrán egy diagramkészítő eszköz (yED)[14] vizualizációs technikájának alkalmazására egy tetszőleges XML alapú taxonómialeírás esetében.



8. ábra **Fogalmi vizualizáció yED eszközzel**

A weben egyre gyakrabban megjelenő szemantikus kereső megoldások döntő része – mint ahogyan az ADVISE is – rendelkezik valamilyen vizualizációs technikával, melyben pozícióval, mérettel, színekkel és egyéb eszközökkel vezetik a felhasználó tekintetét – a rendszer által elképzelt – optimális irányba.

Klasszikus internetes keresők és a mély web

A klasszikus internetes keresők alapvetően szöveges alapú keresést végeznek. Ez tömören nem más, mint a keresőkifejezésben szereplő szavak előfordulásainak megkeresése az adatforrásokban – részlegesen és teljesen egyaránt. Emellett minden keresőnek szüksége van egy rangsorolási modellre, amelynek alapján sorba rendezik azokat a dokumentumokat, amelyekben a keresett kifejezések szerepelnek. Itt legtöbbször az előfordulás gyakorisága, illetve helye a döntő. Tekintettel arra, hogy a legnagyobb keresők sem képesek teljes mértékben lefedni a webes tartalmak teljes egészét, valamint a tartalmak egy része meglehetősen gyorsan változik, kulcskérdés a „fontos" oldalak meghatározása (l. PageRank algoritmus), azaz, hogy mely oldalakat érdemes indexálni, hogy a legtöbb kérdésre releváns választ tudjunk adni.

Az interneten keresztül elérhető tartalmak döntő része láthatatlan marad a keresők előtt, mert űrlapok kitöltésével érhetők el. Az esetek többségében adatbázisból lekérdezett adatokról van szó. A web ezen – gép számára „láthatatlan" – részét nevezi az irodalom *mély web*nek. A mély web méretének becslése gyakorlatilag reménytelen feladat, hiszen a háttérben található adatbázisok szerkezete, mérete többnyire nem publikus, így azzal globális szinten nem lehetséges számolni.

Vannak azonban esetek, amikor a keresők – még ha kis számban is – fel tudják használni a mély web tartalmát. Ehhez speciális illesztőprogramokra, illetve metaadatokra van szükség az érintett adatforrásokhoz. Így lehetséges például, hogy a legnépszerűbb keresők az időjárást és a devizaárfolyamokat gond nélkül szolgáltatják – holott ezek az információk nyilvánvalóan nem HTML oldalak indexálásával álltak elő, hanem speciális adatbázis-hozzáférések által.

Vállalati keresés – miért más belül, mint kívül?

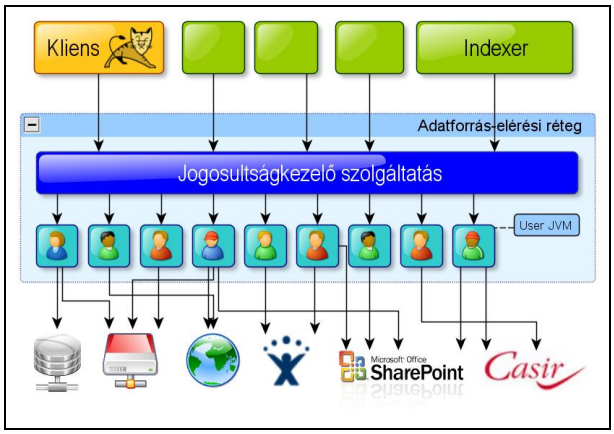
A vállalati keresők egészen más piacot képviselnek, mint az internetes keresők. Ennek oka az eltérő üzleti modell, technológiai háttér és a felhasználói kultúra. A vállalati keresőrendszereknek kisebb létszámú, egyértelműen azonosítható, hasonló érdeklődési körű, illetve általában kvalifikáltabb felhasználót kell kiszolgálni. Az azonosíthatóság nemcsak az információbiztonság szülte szükség, hanem előny is egyben a keresés szempontjából, hiszen a keresési szokások tipizálhatóak, az eredmények testre szabhatóak. Az adatforrások esetében is jelentős differencia mutatkozik. A vállalati rendszerek esetében rengeteg strukturált információ is rendelkezésre áll a strukturálatlan adatok mellett. A struktúrákhoz pedig az esetek többségében metaadatok is tartoznak, melyek segítik a keresést akár automatikus taxonómiát szolgáltatva. Vállalatok esetében a tevékenységi kör sok esetben jól körülhatárolható, ezzel specializálhatóvá téve a keresést és a találatok megjelenítését.

Mérhetőség, kontrollálhatóság, jogosultságkezelés

Vállalati keresőrendszer esetében az adatforrások birtoklása és a felhasználók azonosíthatósága révén a keresőrendszer az alap funkcionalitásához jelentős hozzáadott értéket tud előállítani a vállalat adatvagyonának, illetve a felhasználók, alkalmazottak munkaszokásainak feltérképezésével. Egy vállalati kereső üzemeltetése esetén képet kaphatunk az adatvagyon minőségéről, a hiányosságokról és a feleslegekről egyaránt. Ez visszacsatolást nyújthat a vezetésnek a fejlesztendő, vagy racionalizálandó adatterületek, illetve kompetenciák tekintetében.

A jogosultságkezelés alapvető kérdés vállalati közegben (*9. ábra*). Ennek megfelelően a vállalati keresőknek is igazodniuk kell ehhez. Ez adott esetben igen komplex feladatot is jelenthet, hiszen heterogén rendszerek esetében heterogén jogosultság-ellenőrzéssel állunk szemben, amelyet hibátlanul kell kezelni. További kihívást jelent a vállalati keresők számára, hogy a jogosultságokat valós időben – a keresés közben – szükséges vizsgálni, hiszen bármilyen gyorstárazási módszerrel biztonsági lyuk létrehozását kockáztatjuk.

A vállalati keresők jellemzését követően áttérünk az ADVISE bemutatására, amely vállalati keresőként indult, az IQPortál innovációs fejlesztését követően azonban könyvtári, illetve egyéb információmenedzsment-feladatokra is megkezdtük alkalmazását.



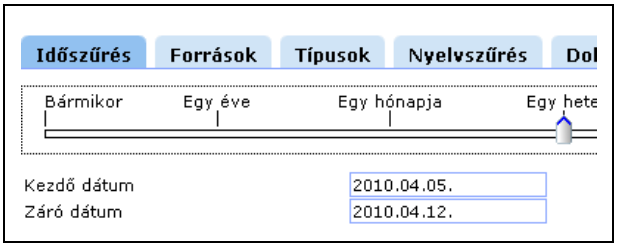
9. ábra **Jogosultságkezelés az ADVISE-ban**

Mi is tulajdonképpen az ADVISE?

Az elnevezés az „Adaptive DataWarehouse Search Engine" játékos rövidítése alapján született. A megnevezés talán félrevezető lehet abból a szempontból, hogy azt sugallja: ez a keresőmotor nemcsak a dokumentumokban, hanem az „adattárházszerű" rendszerekben is keres. Miközben ez egyébként igaz, a következőkben láthatjuk, hogy sokkal többről van szó.

Az ADVISE innovációs termék eleinte elsősorban vállalati igények kielégítésére született, mert a vállalatoknál olyan ütemezett információ- és jelentéskényszer van, amely közvetlenül befolyásolja a gazdasági eredményt, vagyis kimutatható az eszköz közvetlen haszna. Az adattárházakban tényszerű és számszerű adatok vannak, amelyek kinyerése tartalmi, minőségi, pontossági, teljességi és hatékonysági mutatószámokat eredményez. Összetett rendszerek keresése során az adattartalomnak legalább logikai szintű ismerete szükséges ahhoz, hogy a felhasználó meg tudja fogalmazni kérdéseit, illetve értelmezni tudja a kapott válaszokat. Szélesebb körben (pl. internethasználók vagy könyvtárhasználók esetében) már nem várhatók el a megfelelő szintű háttérismeretek. Az internetes keresőkhöz szokott felhasználók egy bonyolult háttérvilág egyszerű keresőjét használják az ADVISE alkalmazása során; az igen összetett informatikai háttér egyetlen felszíni (front-end) megoldásban integrálja a strukturált és a strukturálatlan adatok kereshetőségét.

Az ADVISE egy keresésre tervezett webes felülettel rendelkezik, amely adaptív képessége révén a felhasználók keresési szokásai szerint javítja a találatok súlyozását. A felület és a rendszer „tanulóképessége" a felhasználó igényeihez történő alkalmazkodást és felhasználóbarát megoldást szolgálja. A felhasználók kereséseikhez és a találatokhoz egyaránt könyvjelzőket rendelhetnek, amelyek könnyedén megoszthatók más felhasználókkal. A rendszer tárolja a keresések történetét, a leggyakoribb kérdések egy gombnyomásra lekérdezhetők, és lehetőség van a felhasználók csoportosítására, amit a rendszer automatikusan képes szinkronizálni az elterjedt szolgáltatásokból (LDAP, Active Directory). A keresésnél látjuk, melyik forrásrendszerre várunk, módunk van a kiválasztott találatok rendezésére, értékelésére, jegyzetelésére. A keresés pontosságát fejlett idő- és típusszűrési funkciók támogatják (*10. ábra*).



10. ábra **Idő- és típusszűrési funkciók**

Az ADVISE számos konceptuális elemet örökölt az alapvetően internetes keresésre kifejlesztett motoroktól – például tanulási képesség, adaptív logika alkalmazása, vagy a fogalmak közötti asszociációk építése és karbantartása –, számos területen viszont új megközelítést kellett kialakítani. Ilyen például a vállalati rendszerek esetében természetes *jogosultságkezelés*, és ezzel szoros összefüggésben a *szerepkörvezérelt tanulási algoritmus*. De az ADVISE szakít a hagyományos szekvenciális találati listával is, egy újszerűnek mondható, a fogalmak közötti *szemantikai összefüggéseket hálószerűen ábrázoló megjelenítő felület bevezetésével (11. ábra).* Nemcsak a belső hálózaton található strukturálatlan, szöveges tartalmakat kezeli, hanem a különböző rendszerekben, adatbázisokban és a kapcsolódó metabázisokban található információkat is összegyűjti. A megoldás rugalmas adatforrás-illesztéssel bír, amely – igény esetén –

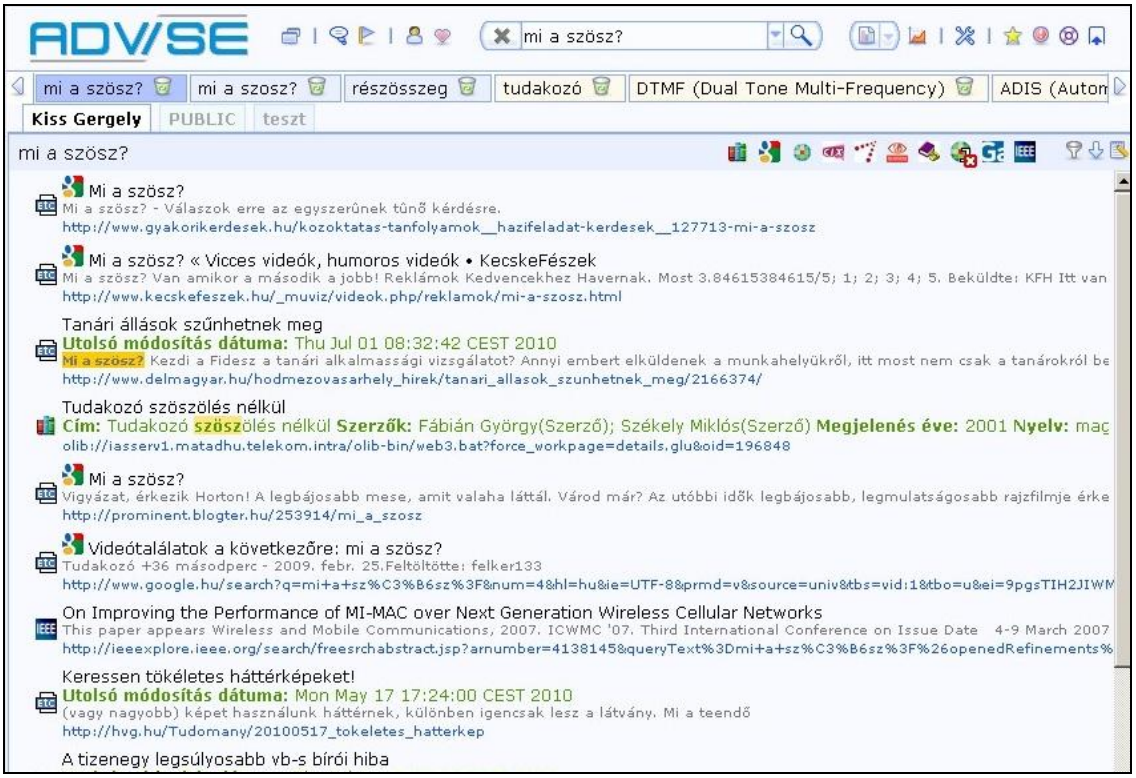
lehetővé teszi további rendszerek bevonását is, ezért alkalmazható könyvtári környezetben a különböző adatbázisok, adatforrások integrálása nyomán a könyvtári információkeresésre, tudásmenedzsment-feladatokhoz és integrált portálkeresőként.

ADVISE – adaptív, tanuló, automatikus keresőrendszer

Az ADVISE fogalomalapú kereső automatikus tárgyszavazási folyamatot végez az adatforrások indexálása során. Természetesen ez a gépi algoritmus önállóan nem tudja azt a pontosságot elérni, mint amire egy ember képes. A felhasználók keresési szokásait azonban adaptálja a rendszer, ezáltal a fontos tárgyszavak köre behatárolható és azoknak a kapcsolati hálója felépíthető. Ez lehetővé teszi, hogy egy riport definiálása előtt a rendszer már a felhasználó szokásai alapján a riportparaméterek döntő részét automatikusan meghatározza. Egy keresési folyamat gyakorlatilag az ad hoc riport fogalmához közelít, és meg is valósítja azt, amikor a felhasználó kéri az eredmények rendszerezését és formázását. Ezt az információt felhasználva a dokumentumokat újraindexálva jelentős pontosság érhető el az automatikus tárgyszavazás módszerében. További fontos tény, hogy számos ismeretterületen már rendelkezésre állnak tárgyszavazott vagy csupán tematikai besorolással rendelkező dokumentumok, melyek szintén felhasználhatók a kapcsolati háló felépítésére és finomítására. Ilyenkor a rendszer pontosan úgy viselkedik, mintha a tárgyszavazást végző felhasználó „annak idején” ezt az ADVISE tanulófelületén keresztül tette volna meg.

Hasonlóan a szemantikus webhez, képes témahelyek közötti meghatározott kapcsolatokra épülő automatikus akciók generálására. Szerkesztése a taxonómiák felső szintje, illetve azok kapcsolatai és előfordulásai szerint történik. Kapcsolódhat osztályozási rendszerekhez, emellett felhasznál(hat)ja a tezaurusz szemantikai ugrópontjait és keresési módszereit is.

A riportdefiniáláshoz, ha információt szeretnénk kinyerni, az ADVISE a fogalmi háló vizualizációs képességével is támogatást nyújt. Az asszociációs kapcsolatok megjelenítése és szerkesztése könnyedén, intuitív módszerekkel megtehető. Kísérleti megvalósításunk alapján a megoldás egyszerűsége egy közönséges wiki vagy más web2-es technológiához hasonlítható.



11. ábra **ADVISE keresőfelület különböző tartalmakból (projektkönyvtár, jogosultságellenőrzés, névadatok stb.) Az egyes adattípusok kiemelten jelennek meg.**

Analitikai funkciók a lekérdezésekhez

Az ADVISE analitikai funkciói közül az alábbiakban néhányat felsorolunk:

Kereséshez kapcsolódó kimutatások

- Melyek a leggyakoribb keresések?
- Melyek az adott témakör legrelevánsabb fogalmai?
- Melyek a legkeresettebb témakörök?
- Melyek a hiánytémakörök?

Dokumentumok

- Melyek a leggyakrabban letöltött dokumentumok?
- Egy adott témakörhöz melyek a legrelevánsabb dokumentumok?
- Melyek a felhasználók szerint leghasznosabb dokumentumok?
- Melyek a felhasználók szerint haszontalan dokumentumok?
- Melyek azok a dokumentumok, amelyeket még senki nem használt?
- Milyen tipikus szűrőfeltételekkel található meg egy adott dokumentum?
- Mekkora a dokumentum eszmei értéke? (Hányan használják és milyen értékeléssel rendelkezik?)

Forrásrendszerek, adatbázisok

- Milyen az egyes rendszerek sebessége, rendelkezésre állása?
- Mekkora az egyes rendszerek, illetve előfizetések kihasználtsága?
- Mely felhasználók vagy csoportok használnak egy adott adatforrást a legintenzívebben?
- Mely adatforrások adatják a legrelevánsabb találatokat?

Felhasználók

- Mennyire aktívak a felhasználók (keresés, letöltés)?
- Mennyire elégedettek a felhasználók a szolgáltatással, a találatok minőségével?
- Milyen célcsoportok, kompetenciák vannak?
- Mely felhasználók, illetve csoportok kompetenciái azonosak?
- Milyen a csoportok közötti kollaboráció?
- Kik vannak feliratkozva egy-egy riportra?
- Ki milyen riportokra van feliratkozva?
- Mely szervezethez tartoznak a feliratkozottak?

Asszociációs jelleg – hogyan lehet riportot definiálni és elkészíteni?

A riportkészítés első lépése mindig annak a meghatározása, hogy mire vagyunk kíváncsiak, azaz milyen információra van szükségünk. Ehhez sok esetben nem egyszer futunk neki a kérdésnek, és fokozatosan, iteratív módszerrel próbáljuk behatárolni, hogy mi lenne számunkra az igazán hasznos információ. Ezt a folyamatot tekintjük a riport definiálásának. Ennek a lépésnek a fontossága a strukturálatlan adatok esetében sem kisebb, mint a jelenlegi megoldásoknál. Kulcskérdés, hogy hogyan lehet megfogalmazni azt, hogy „milyen információra van szükség”? Ebben segít az asszociációs gondolkozás, mely az ADVISE-kereső motorjának és adatelemző rétegének kulcsfontosságú eleme. A fogalomalapú keresés során nem szövegrészleteket bocsátunk a rendszer rendelkezésére, hanem fogalmakat, amelyek között egy asszociációs háló írja le a kapcsolatokat, az agy alapvető működéséhez hasonlóan – azt természetesen lényegesen leegyszerűsítve. Az asszociációk mentén a rendszer az adott fogalomkörhöz legrelevánsabb információhalmazt tudja a forrásokból meghatározni, akkor is, ha a definiálás során a felhasználó az adatforrásokban található terminológiától eltérően fogalmazott.

A fentiek alapján már látható, hogy a strukturálatlan adatokra épülő riport nem más, mint egy jól meghatározott kritériumrendszer mentén végrehajtott adatvagyon-feltérképezés és -keresés majd -rendszerezés és formázott összegzés.

A végső jóváhagyást természetesen mindig az ember adhatja meg, az ADVISE fogalomvizualizációs felületén lehetőség van az automatikusan elkészített tárgyszóhalmaz megtekintésére és felülbírálására. Utóbbi esetben implicit módon ismét tanítottuk a rendszert – amely információ a következő indexálás során ismét felhasználható.

ADVISE – az automatikus fogalmi vizualizáció újszerű megoldása

A fogalmi vizualizáció automatikus előállításával az adott tárgykör fogalmi struktúrája érzékletesen mutatható be. Az ADVISE intuitív vizuális környezete egyszerűen módosíthatóvá teszi a taxonómiát, a későbbi terveink szerint „drag-and-drop” technikával is. A jelenlegi verzióháló export-importot tesz lehetővé, és a vizualizációt a yED eszköz szolgáltatja. A hálók és csomópontok korlátozás nélkül fejleszthetők.

A vizualizáció a lexikai egységek közötti relációkat áttekinthetően mutatja. Rendelkezik intuitíve használható interfésszel, hogy ösztönözze a felfedezést. Érdekessége, hogy olyan elemeket rendel

egymáshoz, amelyeket nem lehet számszerűsíteni, vagyis a kifejezések jelentéseit és kapcsolatait.

A vizuális háló megfigyelései során rugalmasan változó lehet a háló, megfigyelhető és ábrázolható a használók egy adott oldalhoz kapcsolódó internet-használati magatartása is. A vizualizációval elemezhetők akár a rejtett szerkezetek, például az üzleti struktúrákban. Ha például kompetenciatémában készül vizualizáció, látható a hiány vagy a telítettség, amely döntési, stratégiai információ az irányítás kezében. A kompetenciainformációkhoz tartozó és személyekre vonatkozó háló szemléletesen mutatja egy-egy munkatárs tevékenységi struktúráját, kapcsolati rendszerét, vagy tudásának, tevékenységének irányultságát, színvonalát.

Portál és ADVISE-integráció automatikus kategorizáló, lekérdező és tartalomszolgáltató feladatokhoz

Az érdeklődőkkel folytatott konzultációk alkalmával a szakterületen jártas kollégák számára a szemantikus keresésnek és az automatikus taxonómia építésének előnyei könyvtári környezetben pillanatok alatt nyilvánvalóvá váltak, ezért előzetes egyeztetéseket végeztünk mind üzleti, mind architekturális témában az IQPortál és az ADVISE integrációjáról.

Az ADVISE mint kifejezetten heterogén adatforrásokra tervezett szemantikus kereső, az automatikus fogalomépítés mellett biztosított fogalmi hálóbetáplálás képességével a portál heterogén információforrásaiból egy keresési folyamatban képes az információkat kinyerni – ha szükséges, jogosultságokhoz kötve. Az integráció révén a portál nemcsak információs portálszerepet tud betölteni, hanem tudásportál-funkciókat is. Az ADVISE alkalmas az explicit információk sokféle formájából a kompetenciák felderítésére és a kompetenciákhoz tartozó tartalmak kinyerésére, anélkül, hogy a terület szakértőjének vagy művelőjének közreműködését kellene kérni. (Ezt a szabadságot korlátozhatja a jogosultság limitálása.)

Az integrált motor támogatni fogja az OAI-PMH protokollt, a keresőfelület pedig az IQPortál felületébe illeszkedő módon fog megjelenni. A taxonómiaépítés támogatásához a felületen lehetőséget adunk a fogalmi háló megjelenítésére és szerkesztésére is. A könyvtári szakma számára az ADVISE kereső automatikus fogalmi hálóépítési képessége és vizualizációja, a rendszer adaptív, asszociációs képessége és rugalmas módosítási lehetősége jelenti a fő vonzerőt.

Az integráció a fentiekben megfogalmazott célok felhasználói felületét képezik az alábbi megoldásokkal:

a. Felület funkcionalitásának differenciálása felhasználói tapasztalat szerint.
b. Kompetenciamenedzsment-felület.
c. Exchange Server illesztő.
d. SAP BW illesztő (főként vállalati környezetben fontos).
e. Lokális keresés támogatása.
f. SSO támogatás (Single-Sign-On – egyszeri bejelentkezés).
g. OpenSearch illesztés.
h. Windows tálcakomponens.
i. Dokumentumkezelő illesztés.

Technológia

Az ADVISE alapja egy elosztott architektúrára tervezett skálázható keresőmotor. A motor biztosítja a rendszer adaptív funkcióinak integrálását a klasszikus indexálási feladatokon keresztül az ún. okos indexáló bővítményekhez, amelyek metaadatokból, adatbázis/tábla adatokból, szöveges adatokból, dokumentációból, illetve minden olyan tevékenységből származnak, amelyet a felhasználók a keresőrendszerrel végeznek. A keresőmotor a legkorszerűbb lineáris hálózati analízisen alapul, számos specialitással kiegészítve. Ez ad lehetőséget arra, hogy a gyakorlatban fokozatosan módosuló adatbázis változását rentábilisan le lehessen követni algoritmusokkal. A fenti apparátus az igen elterjedt *Hibernate* eszközön keresztül kapcsolódik a JDBC-kompatibilis adatbázisokhoz a legrobusztusabb elosztott gyorstározási és kapcsolatkezelési megoldások támogatása mellett. A rendszer használata nem igényel fejlesztői beavatkozást, üzemeltetése minimális IT-erőforrást köt le. Lehetőség van a moduláris bevezetésre, a többszálú, többgépes működésre, amely a terheléselosztással javítja a teljesítményt. Rugalmas konfigurációt nyújt JVM-en belül / JVM-ek (gépek) között, és lehetővé teszi a runtime (üzemidő alatti) újrakonfigurálást. Folyamatosan megoldott a teljesítménymérés, a memória-nyomkövetés és a távoli hibaelhárítás-funkció.

Az ADVISE automatikus fogalmi hálóépítési mechanizmusa támogatja fogalmi háló importálását, illetve exportálását. A rendszer alapja a platform- és adatbázis-független Java technológia, mely lehetőséget biztosít a megannyi forrásrendszerhez

és adatbázishoz való illesztéshez, valamint kiválóan támogatja a vállalati webes alkalmazások fejlesztését. A rendszer fejlesztése során törekszünk az adatbázis-függetlenségre, hogy a teljesen szabad forráskódú és ingyenes szoftverkomponensektől (pl. MySQL adatbázis, Apache webszerver) a nagyvállalati méretekig (Oracle adatbázis és OC4J webszerver) minden fontosabb és szabványos adatbáziseszközt lefedjünk a közbülső lépcsőkről sem megfeledkezve (pl. Microsoft SQL Server).

Az architektúra tervezésénél kiemelkedő fontosságot tulajdonítunk a skálázhatóságnak. A rendszer moduláris, valamint klaszterezett felépítésű. Bővítésre lehetőség van már telepített rendszer esetében is számottevő költség nélkül.

A webes felületek AJAX technológiát alkalmaznak, amely letisztult, interaktív tájékozódást tesz lehetővé a keresések és a riportok elkészítése, valamint böngészése során. A rendszer többi részéhez hasonlóan itt is törekszünk a platformfüggetlenségre, ezért gyakorlatilag minden kurrens böngészőt támogatunk (Explorer, Firefox, Chrome, Opera, Safari).

A rendszer architektúráját az alábbi *komponensek* alkot(hat)ják, de ettől részben eltérő is lehet:
- *Adatforrás-illesztő.*
- *E-mail-illesztő.*
- *Adatbázis*: a rendszer fogalmi hálóját valamint a felhasználói és az analitikai adatokat tároljuk itt.
- *Riportalkalmazás*.
- *Analitikai funkciók.*
- *Keresőmotor*.
- *Fogalomtár-funkciók.*
- *Webes felület (analitikai felület a rendszer analitikai funkciói és jelentései eléréséhez, keresőfelület, fogalomtár-felület, igénylőfelület).*

Az ADVISE mint termék

Az információmenedzsment területén folyamatosan fennálló költségelvonás és -hiány miatt különösen fontos az integrációban rejlő szinergiák kihasználása, és a fenti fejlesztés termékszerű megjelenítése az egyes szakterületek számára (könyvtár, levéltár, múzeum, üzleti vállalkozás, MOKKA stb.) – mindazon informatikai lehetőségekkel, amelyet az ADVISE automatikus szemantikus kereső és az IQPortál integrációja kínál. Ezért tervezzük különböző intézménytípusok számára a rendszer termékként való értékesítését, amely a meghatározott funkciócsoportokhoz standard megoldásokat kínál – főként az alábbi modulokra bontva:
- Automatikus egyidejű keresés elektronikus dokumentumokban, fogalmi rendszerezés a szemantikus keresés pontosítása érdekében (automatikus osztályozás és keresés, taxonómiák és ontológiák, tématérképek, fogalmi háló stb.).
- Integráció az IQPortál termékkel és a preferált rendszerek illesztésével (adatbázisok, fájlrendszerek, digitális könyvtárak, távoli adatbázisok, megvásárolt adatbázisok, internet, kompetenciamenedzsment-felület, SSO támogatás stb.).
- Használat során megfogalmazott igények (OpenSearch és Windows tálca bővítmény, dokumentumkezelő illesztés, saját webfelülettervezés stb.).

oooOOOooo

Az ADVISE és az IQPortál együttesen teszi megoldhatóvá az elektronikus tartalmak egyidejű keresését a költséges keresőnyelvek előállításának kötelezettsége nélkül. Ugyanakkor a keresések során a rendszer „tanulja” és tárolja az adott szervezet által használt fogalmakat és kapcsolataikat, ezáltal a keresőnyelv mégis automatikusan létrejön a használat során, és kis közreműködéssel további előnyös automatikus rendezési lehetőségeket tesz lehetővé (fogalmi vizualizáció, tématérkép, ontológia támogatása stb.).

A könyvtári szakma számára az ADVISE kereső automatikus fogalmi hálóépítési képessége és vizualizációja jelenti a fő vonzerőt, e képességek beépítése az IQPortál termékbe a legfontosabb eddig megismert igény. Meglátásunk szerint a létrejövő MOKKA rendszer összetett keresési igényeihez is jelentős támogatást nyújthatna az ADVISE, tekintettel arra, hogy bármely szabványos alapokon álló rendszerrel képes integrációra.

Jegyzetek és hivatkozások

[1] Az ADVISE mozaikszó az *Adaptive DataWarehouse Search Engine* összetételből ered, a márkanév „ADVISE” formában használatos.

[2] DRESNER, Howard: The performance management revolution; Business results through insight and action. New Jersey, Wiley, 2008. 231 p.

[3] CSIK Tibor – VARGA Katalin: A tudás és az információfeldolgozás. = http://tmt.omikk.bme.hu/show_news.html?id=4007&issue_id=464

[4] Autonomy, vö.: BÁNHEGYI Zsolt: Vállalati-üzleti információszerzés: a szoftveripar újdonságai. = TMT, 55. köt. 5. sz. 2008. http://tmt.omikk.bme.hu/show_news.html?id=4894&issue_id=493

[5.] Adaptivitás: Alkalmazkodóképesség, tanuló rendszerek jellemzője. Kiemelt fontosságú tulajdonság olyan környezetekben, ahol a megoldandó feladat algoritmikusan nem kódolható előre.

[6] Crawling: *http://en.wikipedia.org/wiki/Web_crawler* ill. *http://en.wikipedia.org/wiki/Distributed_web_crawling*; Carlos CASTILLO: Effective Web Crawling *http://www.webir.org/resources/phd/Castillo_2004.pdf*

[7] PageRank: *http://en.wikipedia.org/wiki/PageRank*

[8] Levensthein-távolság: *http://en.wikipedia.org/wiki/Levenshtein_distance*; GILLEAND, Michale: Levenshtein Distance, in Three Flavors. *http://www.merriampark.com/ld.htm*

[9] *Hash*technika: Értékek hasítása, olyan hatékonyan (gyorsan) végrehajtható függvénnyel, amely során az értelmezési tartomány elemei egyenletesen szórnak a *hash*értékkészlet tartományában. Alkalmazásával gyorsan kikereshetők értékek, ez adatbázisoknál és a keresés területén egyaránt fontos technológiai kitétel.

[10] Elosztott hash-tábla: http://en.wikipedia.org/wiki/Distributed_hash_table; http://en.wikipedia.org/wiki/Hash_function; *http://www.prototypejs.org/api/hash*

[11] Taxonómiákra nem térünk itt ki, mert több típusról kellene szólni. Ugyancsak nem térünk ki a taxonómia és a tezaurusz viszonyának taglalására. A vállalati taxonómiák a gépi tartalomrendszerezés céljára készülnek, vagy internetes tartalom rendezésére. L. bővebben: HORVÁTH Zoltánné: Taxonómia – az egyezményes nyelvek szerepe és rokonságai – útközben a szemantikus webhez. = http://tmt.omikk.bme.hu/issue.html?issue_id=472

[12] Szemantikus web: Értelmező/intelligens web, olyan világháló, amelyen az információk számítógépes értelmezésre is felkészített formában állnak rendelkezésre, ezzel elősegítve a gépi keresés és egyéb intelligens szolgáltatások készítését.

[13] Az RDF és az OWL nyelvről: [OWL] Web Ontology Language = http://en.wikipedia.org/wiki/Web_Ontology_Language; *[RDF]* Resource Description Framework = http://en.wikipedia.org/wiki/Resource_Description_Framework

[14.] *yEd*: szabad forráskódú Java alapú, hálózati vizualizációs (diagramkészítő) alkalmazásról: http://www.yworks.com/en/products_yed_about.html

Irodalom

BARÁTNÉ HAJDÚ Ágnes: A percepció és megjelenítés jelentősége az információkereső nyelvekben. 2007. = *http://tmt.omikk.bme.hu/show_news.html?id=4785&issue_id=487*

BOGNÁR Katalin: Tudásalapú rendszerek és technológiák. 2006. = http://www.inf.unideb.hu/~bognar/mestint4/mestint_konyv.pdf

BRODER, Andrei: A taxonomy of web search. = *http://www.sigir.org/forum/F2002/broder.pdf*

FAJSZI Bulcsú – CSER László – FEHÉR Tamás: Üzleti haszon az adatok mélyén. Az adatbányászat mindennapjai. Budapest, Alinea, 2010. 414 p. Tartalomjegyzék. = http://www.alinea.hu/pages/uzletihaszon/adatbanyaszat_tartalom.pdf

Information retrieval on the WWW and active logic. A. A. Barfourosh et al. = http://www.lib.umd.edu/drum/bitstream/1903/1153/1/CS-TR-4291.pdf

KISS Gergely: Skálázható, intelligens megoldások fejlesztése Java technológiával. = IQSYMPOSIUM, 2009. október 7. Összefoglaló. = *http://www.iqsys.hu/web/guest/iqsymposium-operativ-informaciotechnologia-2009*

KISS Gergely: Adaptív adattárház újdonságok. ADVISE 1.2. = IQSYMPOSIUM, 2010. április 14. (pdf). Összefoglaló. http://www.iqsys.hu/c/document_library/get-file?44i

KOVÁCS László – MICSIK András: Szemantikus webszolgáltatások tervezése és megvalósítása. = *http://www.hiradastechnika.hu/data/upload/file/2006/2006_1/HT_0601-4.pdf*

LEHNMANN Miklós: Vizualitás. A képek szerepe a tudományban. = http://www.tofk.elte.hu/tarstud/filmuvtort_2001/lehmann.htm

Integration and verification of semantic constraints in adaptive process management systems. DADAM, Peter et. al. = http://www.informatik.uniulm.de/dbis/01/dbis/downloads/LRD07.pdf

REEVE, Larry: Information retrieval on the semantic web using ontology-based visualisation. = http://www.pages.drexel.edu/~lhr24/courses/Info780-06Paper.pdf

Semantic data integration for the Enterprise. Oracle White papers. = http://www.oracle.com/technology/tech/semantic_technologies/pdf/semantic11g_dataint_twp.pdf

Taxonomies. Frameworks for Corporate Knowledge. Second ed. by Jan WYLLIE, Trand Monitor… David SKYRME., ed. Simon LELIC, Ark Group. - London, ARK Group, 2005. 80 p. ISBN 0-9549674-1-0

UNGVÁRY Rudolf: Tezaurusz és ontológia, avagy a fogalmi ismertetőjegyek generikus öröklődésének formalizálása. = http://tmt.omikk.bme.hu/show_news.html?id=3615&issue_id=450

Beérkezett: 2010. VII. 11-én.

Horváth György
az IQSYS Zrt. üzleti intelligencia szakértője.

Horváth Zoltánné
az IQSYS Zrt. könyvtári szakértője, termékmenedzser.
E-mail: horvath.zoltanne@iqpp.hu

Papp Attila
az IQSYS Zrt. üzleti intelligencia vezető szakértője.
E-mail: papp.attila@iqpp.hu

Jelentkezési felhívás segédkönyvtáros tanfolyamra

A Budapesti Műszaki és Gazdaságtudományi Egyetem Országos Műszaki Információs Központ és Könyvtár (BME OMIKK) emelt szintű OKJ-s segédkönyvtáros tanfolyamot hirdet.

A végzett hallgató munkaköre:
segédkönyvtáros.
Az oktatás elsősorban gyakorlati jellegű, amely a vizsgakövetelményekben is érvényesül.

A tanfolyam **2011. januárban**, keresztféléves képzési formában indul.
A képzés időtartama két félév.

A foglalkozásokat heti egy alkalommal csütörtökönként, valamint minden hónap utolsó hetében szerdán és csütörtökön 8-tól 17 óráig tartjuk.

Részvételi díj a két félévre
150 000,– Ft, a vizsgák költsége **50 000,– Ft.**

Felvételi vizsga nincs, a beiratkozás feltétele az érettségi bizonyítvány bemutatása.

A tanfolyam jegyzeteit, segédkönyveit kölcsönzés formájában biztosítja a szervező intézmény.

A képzésre azoknak a jelentkezését várjuk, akik a könyvtári munka gyakorlatát rövid idő alatt kívánják elsajátítani, és a számítógép használatában négy ECDL modul megismerésével jártasságot akarnak szerezni.

Jelentkezni az alábbi címre eljuttatott (kitöltött, kinyomtatott) jelentkezési űrlappal lehet:

BME OMIKK
segédkönyvtáros képzés
1111 Budapest, Budafoki út 4-6.

A jelentkezési űrlap a BME OMIKK honlapjáról letölthető
http://www.omikk.bme.hu/main.php?folderID=1159&articleID=1816&ctag=articlelist&iid=1
Jelentkezési határidő: **2010. december 15-ig**

További felvilágosítás **463-3534**-es telefonszámon és a **gylengyel@omikk.bme.hu** e-mail címen Lengyel Gyöngyitől kérhető.