HADMÉRNÖK

VII. Évfolyam 2. szám - 2012. június

Vadász Pál
vadasz.pal@montana.hu

EGY NYÍLT FORRÁSOKRA ÁLLÍTOTT SZEMANTIKUS KERESŐRENDSZER BEMUTATÁSA

Absztrakt

Az exponenciálisan növekvő strukturálatlan adatmennyiség a weben szükségessé tette a kulcsszavas keresésen túlmutató szemantikus keresés széleskörű elterjedését. Minősített adatok kezelésére bemutatok egy biztonságos architektúrát egy, a web felé nyitott és egy, a külvilágtól elzárt kereső-rendszerrel.

The exponentially growing amount of unstructured data in the web has caused the widespread use of semantic search as opposed to the character based search. A secure architecture is represented for handling classified data using the combination of an open source and a cut-off search system.

Kulcsszavak: *OSINT, szemantikus keresés, karaktersor alapú keresés ~ OSINT, semantic search, character string based search*

BEVEZETÉS

Az internet megjelenésével a múlt század 90-es éveitől exponenciálisan nő az elérhető adatok mennyisége. Mike Lynch, az Autonomy alapítója becslése [1; 545. o.] szerint ezek 80%-a adatbázisba nem szervezett, strukturálatlan adattömeg, amelyet a klasszikus technológiákkal nem lehet feltárni. A kifejlesztett webkeresők (AltaVista, Yahoo, később Google) kulcsszavakra keresnek, ezért a találati arány csekély, emberi mértékkel átláthatatlan mennyiségű eredményt hoz ki. Számottevően jobb eredményt kapunk a jelentéstartalmú (szemantikus) keresők alkalmazásával, ahol a rokon értelmű szavak hierarchikusan egymás alá rendelt fogalmak értelmezésével, szövegkörnyezet elemzésével a felesleges, pontatlan találatokat kiszűrhetjük, és csak a valódi eredményeket kapjuk meg.

A weben, azaz a nyílt forrásokban való keresés igénye mind a gazdasági életben, mind pedig a katonai területen azonnal felmerült. Feltételezhetően itt is érvényesül a Pareto-szabály, miszerint az eredmény 80%-a 20% erőfeszítéssel nyílt forrásból is elérhető, ami viszonylagosan olcsó, névtelen, és főleg nem kockáztat emberi életet, diplomáciai bonyodalmakat. A NATO már 2002-ben kiadott több nyilvános kézikönyvet is a témában [2; 3; 4]. A lehetséges felhasználási területek bőven kimerítenék egy külön cikk kereteit is. A nyílt forrásokból való információ-szerzés egyetlen szakmai vélemény szerint sem pótolja az egyéb felderítési lehetőségeket, hanem kiegészíti azokat.

Egy átlagos biztonsági igényű szervezet rosszindulatú külső hatások elleni védelmére egy egy- vagy kétszintes tűzfal rendszer elegendő. Azonban egy kényes vagy minősített tartalmakat kezelő környezetet semmilyen technológia nem véd meg teljes biztonsággal. Ezért ezeket fizikailag nem kötik össze a külvilággal. Mivel egy nyílt forrásokra dolgozó kereső a dolog természetéből fakadóan állandó kapcsolatban kell, hogy legyen a webbel, meg kell oldani a külső keresőrendszer által generált adattartalmak egyirányú bejuttatását a belsőbe. Ennek több technikai megoldása van. Ezek ismertetése meghaladja ennek a cikknek a kereteit.

Magyarországon tudomásom szerint csak igen kevés ilyen struktúra működik. A téma aktualitását a nyílt forrású keresők iránti robbanásszerű érdeklődés adja. Elég csak a bűnüldözésben felhasználható közösségi oldalakból származó információkra gondolni. Az alább ismertetett keresőrendszer egy létező architektúra leírására épül. Célja kettős. Egyrészt a döntéshozók részére tájékozódást adjon, másrészt az informatikai szakemberek számára irányt mutasson.

1. A KERESŐRENDSZER ARCHITEKTÚRÁJA, KAPCSOLÓDÓ SZEREPKÖRÖK

Egy kényes vagy minősített anyagokat kezelő rendszer két részre osztható (külön hálózat éri el az internetet és a belső hálózat ettől teljesen szeparált). A kialakítandó kereső is két fő elemből kell, hogy álljon. A külső hálózaton található szerver indexeli a megadott paraméterek alapján az internetes információkat. Mindemellett előre elkészített szabályok alapján az indexelt adatok közül kiválogatja a felhasználó különböző egységei és munkatársai számára releváns dokumentumokat, és ezeket le is tölti a külső rendszerbe. A letöltés nemcsak szűrés alapján kell, hogy működjön, hanem lehetnek bizonyos tartalmak, amelyeket szűrés nélkül kell letölteni. A külső rendszerben rendelkezésre fog állni az a keresőfelület, melynek segítségével az indexelt információk között lehet majd keresni, értesítéseket beállítani, előre megadott taxonómia[1] alapján kategorizálni[2] stb.

[1] Dokumentumok csoportosítására szolgáló hierarchikus kategória rendszer. Két része van, maga a kategória struktúra és az egyes kategóriákat meghatározó szabályok.

[2] Lekérdezés, melynek eleget tevő dokumentumok bekerülnek a hozzá tartozó taxonómia kategóriába.

A letöltött anyagok áttöltésre kerülnek egy belső rendszerbe, majd azokat le kell indexelni a belső rendszerben működő szerverrel. Ennek eredményeképpen a letöltött anyagok között a belső rendszert használó felhasználók keresni tudnak (taxonómiák és egyéb tudásháttérek segítségével is). Ennek során:

- a külső rendszerben letöltött dokumentumok átvitelre kerülnek a belső hálózatba, és ott importálással kereshetővé válnak (és alapjait képezhetik értesítő szolgáltatásoknak) a belső rendszert használó felhasználók számára;
- miután a letöltött dokumentumok vagy az importálandó index fájlok a belső hálózatba kerültek, a rendszer indexeli őket és kategorizálásra kerülnek a belső hálózat taxonómiái alapján;
- az indexeléskor az index alapon áttöltött dokumentumoknak minden metaadata megtalálható lesz a belső hálózatban külön konfiguráció nélkül (a mezőket a szerveren létre kell hozni);
- azokat a dokumentumokat, melyeknek eredeti formája töltődik át, a külső hálózatban meglévő tartalomszűrési és adatkinyerési konfigurációkhoz hasonló módon kell indexelni.

1.1 A rendszerrel kapcsolatos szerepkörök

A rendszerrel kapcsolatos szerepkörök azonosítása a működés sikeréhez elengedhetetlen. A szerepkörök meghatározásához nemcsak a rendszer sajátosságait, hanem az azt használni készülők munkafolyamatait és információs igényeit is figyelembe kell venni. Ezek a szerepkörök a munkafolyamat során jól elkülöníthető, más és más képzésszintet igénylő feladatcsoportokhoz kapcsolódnak.

A tapasztalat szerint egy komplex keresőrendszer technikai megvalósítása néhány hét vagy hónap alatt megtörténhet, míg a szervezetben való tényleges abszorbeálása lényegesen hosszabb időt vesz igénybe. Több példa is ismeretes hatalmas beruházások elhalására a folyamatok szabályozása és a képzés elmaradása miatt.

Alkalmi felhasználónak nevezzük azt a munkatársat, aki a rendszer által biztosított keresőfelületet használja bizonyos rendszerességgel, de automatizált információszolgáltatást nem vesz igénybe.

A rendszer *folyamatos felhasználója* az, aki a keresőfelület mellett igénybe veszi a különböző értesítési szolgáltatásokat. Olyan felülettel rendelkezik, amelyen a frissülő értesítési információk mindig testre szabottan megjelennek.

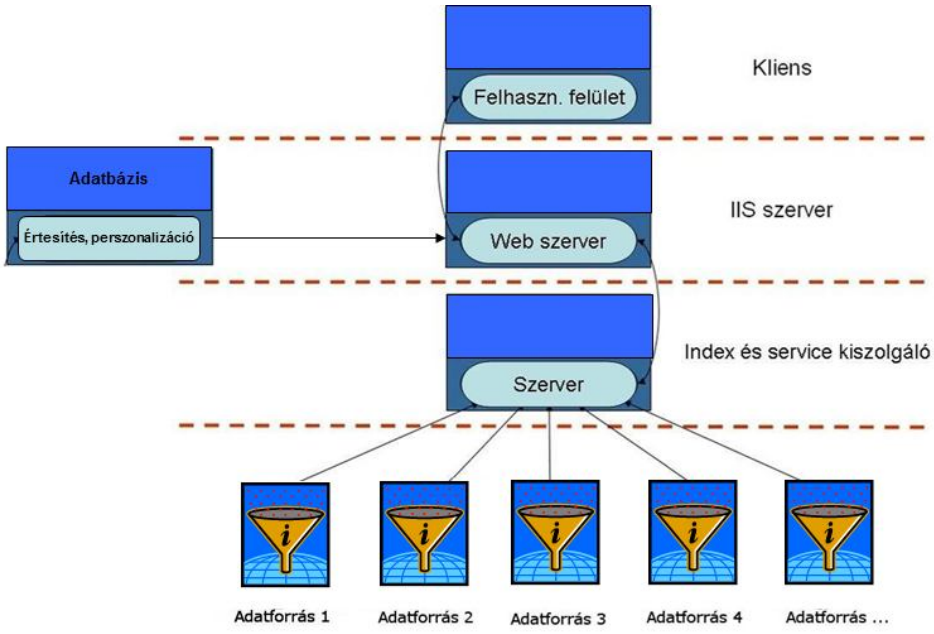
A *kulcsfelhasználó* amellett hogy a rendszer folyamatos használója, kapcsolatban van a tudásmenedzserrel, akinek javaslatot ad a rendszer tartalmi oldalával kapcsolatosan, illetve támogatja szakmai oldalról.

A *tudásmenedzser* felelős a rendszer tartalmi karbantartásáért, bővítéséért, teszteléséért. A feladatok ellátásához elsősorban a kulcsfelhasználókkal történő egyeztetés a célszerű, de folyamatosan kapcsolatban kell lenni minden felhasználóval. Mivel a rendszer nagy mennyiségű és folyamatosan változó információtömeg szintén változó szempontok alapján történő feldolgozását végzi, a jó működéshez elengedhetetlenül fontos a tudásmenedzser megléte.

Az *üzemeltetői szerepkör* jelenti a rendszer informatikai szempontú kezelését. Természetesen a rendszer jellegéből adódóan több feladat is van, amelyet a tudásmenedzseri szerepkörrel közösen, konzultálva kell elvégezni.

1.2 Egy lehetséges szoftver architektúra

A felhasználói felület és az értesítő szolgáltatások a háttérben a szerver segítségével végzik a kereséseket. A szerver indexeit az indexelő behívások (*fetch*[3]-ek) frissítik. Az indexek naprakészsége függ az indexelési ütemezésektől, melyeket adatforrásonként lehet beállítani, és a beállításoktól függően akár kézzel, akár automatizáltan lehet futtatni.



1. ábra. A keresőrendszer szoftver architektúrája (saját)

2. INDEXELÉS, KATEGORIZÁLÁS, LETÖLTÉS ÉS SZŰRÉS

2.1 Az indexelés folyamata, feladatai

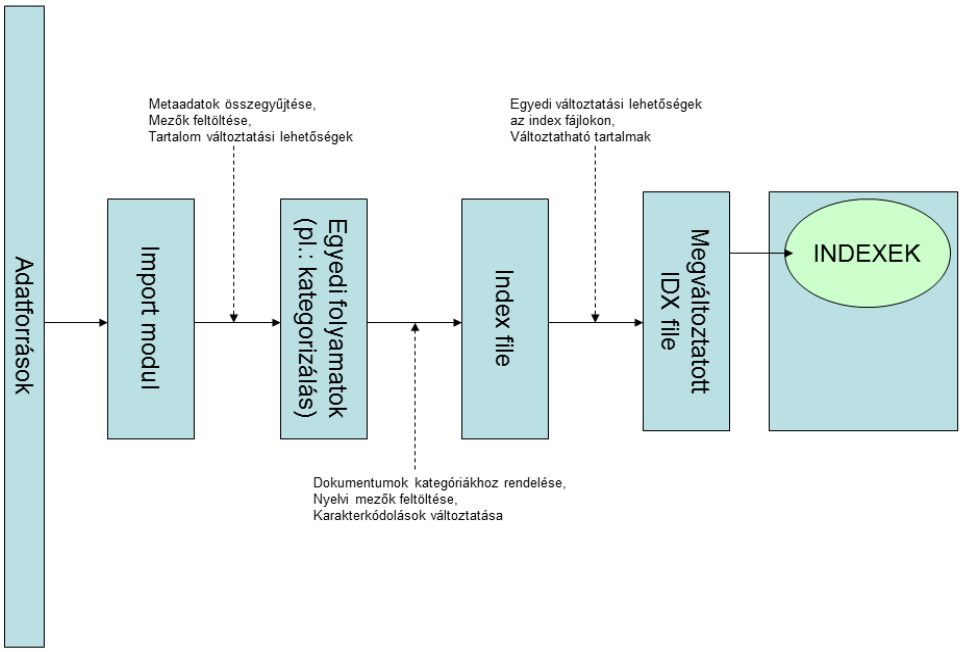
Az indexelési feladatok ellátására a rendszer a különböző adattípusokhoz tartozó behívásokat használhat. Minden behívás külön konfigurálható, és mindegyiken belül önálló index munkafolyamatok (*job*-ok) hozhatók létre külön időzítéssel. Weboldalak indexeléséhez a http behívás, míg a hálózaton tárolt tartalmakhoz a fájlrendszer (vagy a http) behívás használható.

Az egyes behívásokat *windows* szolgáltatásokon (*service*) keresztül lehet elindítani. Működésük közben index fájlok keletkeznek, melyek tartalmazzák az indexelendő információkat, és indexelés után megtarthatók vagy törölhetők.

A meghatározott kritériumoknak megfelelő dokumentumokat a rendszer nemcsak indexeli, hanem le is tölti a külső rendszerbe abból a célból, hogy a megfelelő kategóriába tartozó vagy a felhasználók által letölteni kívánt információk áttöltésre kerüljenek a belső hálózat rendszerébe. A külső rendszerben az internetes keresésre jogosult felhasználók egy grafikus felületen keresztül tudnak keresni az indexelt információk között.

Az indexelés weboldalanként (vagy részenként) külön történik, külön oldalra jellemző beállításokkal, azonban már az index konfigurációk létrehozása előtt el kell dönteni az adott oldalról, hogy milyen indexelési alaptípussal érdemes indexelni.

[3] A rendszer indexelő mechanizmusa. Külön *fetch* kapcsolódik a különböző típusú (fájlrendszer, http, stb.) információforrásokhoz.



2. ábra. Az indexelés folyamata (saját)

2.2 A dokumentumok kategorizálásának folyamata, feladatai

A kategorizálási folyamatra azért van szükség, hogy a dokumentumokat több szempontrendszer szerint is csoportosítani és válogatni lehessen automatikus módon. Ez teszi lehetővé a taxonómia alapú navigációt a felhasználói felületen.

A kategorizálás megvalósítására két lehetőség kínálkozik:

1.) A kategorizálási folyamat minden dokumentumot megfeleltet a meglévő kategóriáknak, és amelyikbe beleillik, azzal megcímkézi (*taggeli*) az indexelendő dokumentumot, valamint lehetővé válik a kategória alapú dokumentumlista keresése a rendszerben. A rendszer kategorizálási folyamata megadott időközönként fut le, és az azóta érkezett dokumentumokat, valamint az új és módosított kategóriák miatti változásokat átvezeti a rendszeren anélkül, hogy újra kellene indexelni a dokumentumokat. A módszer hátránya, hogy a kategóriák változtatása nem érvényesül, csak egy komolyabb folyamat lefuttatásával, valamint az indexelési időben jelentősen terheli az indexelőt. Előnye, hogy az indexelt kategóriákból parametrikus indexek is építhetőek (persze ezek is nem túl nagy paraméter mennyiség esetén csak).

2.) Az indexelés során nem kategorizálja a rendszer a dokumentumokat, hanem minden kategóriaválogatás (pl. taxonómia keresés) lefutásával röptében válogatja le a kategória szempontjából releváns dokumentumokat. Ennek a módszernek az előnye, hogy nem lassítja az indexelést (ami nagy mennyiségű dokumentum esetén fontos szempont, és a kategóriák változtatása esetén is mindig naprakész. Hátránya, hogy nem lehetséges kategóriák alapján parametrikus indexeket képezni.

A kategóriák között lennie kell ún. letöltési kategóriáknak is, amelyekbe a beletartozó dokumentumok kerülnek áttöltésre a belső rendszerbe. A kategóriákat a tudásmenedzser hozza létre.

A rendszernek kezelnie kell a tudásmenedzser által létrehozott taxonómiákat úgy, hogy azokat gyorsan publikálni lehessen, ezáltal megjelenhessenek a felhasználói felületen, segítve a felhasználói navigációt.

A taxonómiák hierarchikus felépítésű kategória rendszerek, melyek minden eleméhez (ezek a kategóriák) kapcsolódik lekérdezés, ami a dokumentumok kategóriákhoz való tartozását meghatározza. Ezen kívül kategória szabályokat létre lehet hozni minta dokumentumok alapján is a rendszerben.

2.3 Dokumentumok letöltése

Mivel a belső és a külső hálózat teljesen szeparált, szükség van az információk letöltésére, hogy azokat át lehessen másolni a belső hálózatot használóknak. A letöltésnek szűrtnek kell lennie, azaz nem minden indexelt dokumentumot kell letölteni. A dokumentumok automatikus letöltése kategorizálás alapján történik. Azokat a kategóriákat, amelyek letöltendő dokumentumokhoz kapcsolódnak, külön taxonómiá(k)ban kell rögzíteni. Amikor a külső rendszer kategorizál, hozzárendeli az indexelt dokumentumokat a kategóriákhoz.

2.4 Tartalomszűrés

Ez a funkció weboldalak indexelésekor rendkívül hasznos, mivel a következőket teszi lehetővé:

- weboldal információs szempontból felesleges részeinek eldobása, indexelésből kihagyása;
- adott weboldal információinak egyedi mezőkben történő tárolása;
- adott weboldalra jellemző adatforrás-specifikus keresések létrehozása;
- web információk konfigurálható megjelenítése.

Így csak az érdemi információban történik keresés, nincsenek olyan találatok, melyeket csak egy link vagy reklám hoz be. Ez a módszer szükség esetén megváltoztatható úgy, hogy a rendszer az oldalak teljes szövegében keressen, de ez visszamenőleg nem végezhető el, mivel a teljes adatbázis újraindexelése szükséges.

A tartalomszűrés alapfeltétele, hogy az indexelt weboldalon olyan meta-címke (*metatag*) struktúra legyen (meta-címke alatt ebben az esetben bármilyen karaktersorozatot érthetünk), amely egyértelműen azonosítja a kiszűrendő vagy kiemelendő tartalmat. Minden más esetben csak a hagyományos teljes indexelés lehetséges.

A fentiek szerint a tartalomszűréses indexelés használatához minden egyes indexelendő weboldal (vagy weboldal rész) esetén fel kell mérni a konfigurációhoz szükséges információkat, és azokat oldalanként beépíteni az indexelési konfigurációba.

3. KERESÉS ÉS JELENTÉSKÉSZÍTÉS

3.1 Keresési lehetőségek, módszerek

Ez a téma lényegében a szabadszavas, logikai operátoros és a beépített (minden adatforráshoz tartozó) metaadat[4] keresést jelenti, amely a felhasználói felületről elérhető. Ez a keresési mód lehetővé teszi, hogy ad hoc módon kutasson a felhasználó az általa látható adatforrásokban, kihasználva a felület által nyújtott kényelmi funkciókat. Ennek a keresési módnak az alapjai szabadszavas kulcsszavak, melyek a kiválasztott logikai operátorokkal kereshetők le.

Az *adatforrás specifikus keresési lehetőség* lényege, hogy a felhasználók a keresőfelületükön egyébként is adott keresőmezőkön túl keresni tudnak csak adott adatforrásokra jellemző speciális metaadatokra is. Ezt a funkciót az adatforrások felmérésekor kialakítandó metaadat-

[4] Leíró, jellemző adat, amely egy dokumentumhoz kapcsolódik, de nem feltétlenül része a tartalmának (pl. szerző, dátum, stb.).

struktúra teszi lehetővé, mivel az itt meghatározott adatokra teszünk lehetővé speciális kereséseket.

A *parametrikus keresés* lehetővé teszi az adatforrás adott paraméterek mentén történő szűkítését, a szűkítő paraméterek dinamikus változásával együtt. Ez azt jelenti, hogy ha valamely paramétert kiválasztjuk, akkor az összes többi választható dolog az első szűkítésnek megfelelően fog változni. Ennek a dinamikus szűkítési módnak a feltétele, hogy az adatforrás paraméterként használni kívánt metaadatai véges tartalmúak legyenek, és legyenek jelen az adatforrás minden dokumentumában.

A *taxonómia alapú keresés* lényege, hogy az előre meghatározott kategóriarendszerek és az egyes kategóriákhoz tartozó szabályok (vagyis a taxonómiák) alapján egy olyan keresést tegyen lehetővé a felhasználói felületen, amely a fa struktúra grafikus böngészésével képes leválogatni az adott kategóriába tartozó dokumentumokat. Mindemellett a taxonómia fa egyes elemeihez tartozó szabályok önmagukban is alkalmasak különböző automatikus szolgáltatások (pl. értesítés) alapjául szolgálni.

Az *összevont keresés* (*federated search*) lényege, hogy amikor a felhasználó egy keresést végrehajt a rendszerben, nemcsak a rendszer keresőjét használhatja, hanem a keresést „elküldheti" különböző publikus keresőknek, amelyek a találataikat a rendszerben előálló csoportosított találati listában jelenítik meg.

Ahhoz, hogy a találati lista a felhasználói felületen megjelenhessen (mint egy külön találati lista fül), a megjelenítést konfigurációval az adott keresőre kell szabni. Ez azt jelenti, hogy új kereső felvételekor (fejlesztést igényel) mindig testreszabott új konfigurációt kell készíteni. A konfigurációs fájlban be kell állítani, hogy az adott kereső hogyan jeleníti meg a címet, összefoglalót, url-t, stb., valamint, hogy hány találatot adjon vissza.

3.2 Duplikátum szűrés

Indexeléskor a rendszer a beépített duplikátumszűrő mechanizmusát képes használni, melynek lényege a referencia alapú ellenőrzés. A dokumentumok referenciája az url-ből képződik. Ez a módszer alkalmazható fájlrendszer és webes tartalmak esetén is.

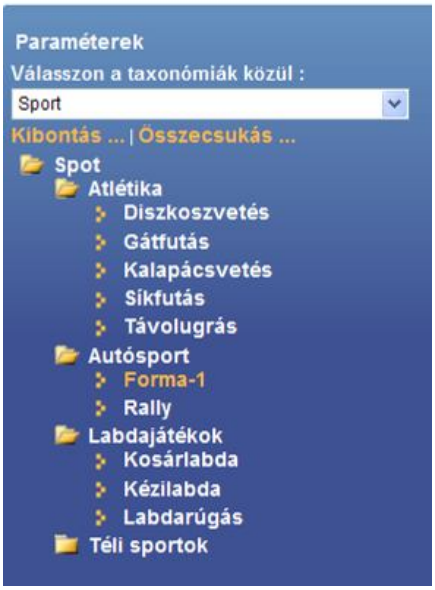
A duplikátumszűrés fájlszinten ellenőrző összeg (*checksum)* alapján is történhet, és lehetővé teszi, hogy az egyforma dokumentumok csak egyszer jelenjenek meg a rendszerben. Ez a módszer kifejezetten ugyanolyan fájlokra működik, és nem célszerű alkalmazni jogosultságkezeléssel használt adatforrások esetén.

3.3 Eredmények megjelenítése

A rendszer az *összefoglalókat* nem indexeléskor, hanem kereséskor dinamikusan állítja elő, így figyelembe vehető a kereső-kifejezések elhelyezkedése a dokumentumban az összefoglaló készítésénél. Az összefoglaló típusa változtatható, ennek beállításaiban a mindenkori kereső motor lehetőségei az irányadóak. Alapkiépítésben a rendszer a kereső-kifejezésekhez mérten legjobb néhány mondatot (összefüggő szöveget) teszi bele az összefoglalóba.

A rendszer képes a *fontosnak ítélt kifejezéseket* a szövegből kiválogatni, és minden dokumentum mellé metaadatként hozzáfűzni, ezzel is javítva a dokumentum kereshetőségét. A működés lényege, hogy a találati listában minden dokumentumnál elérhető egy olyan megjelenítési lehetőség, melyre kattintva a felhasználó a fontosnak ítélt kulcsszavakat is megjeleníti a dokumentum szövege mellett.

Megjeleníthető a *taxonómia* is, kiválasztása egy listadobozból történhet. A lista azokat a taxonómiákat tartalmazza, amelyek a konfigurációs fájlban be lettek állítva.



3. ábra. Taxonómia kiválasztásának lehetősége (saját)

A taxonómia kereséskor is használható a szabad szöveges kereső mező, így egyszerre több módszerrel is szűkíthető a találati lista.

A taxonómiák csoportba foglalását és az általuk használt adatforrásokat konfigurációs fájlból lehet beállítani. Egy csoportba több fa is tartozhat. A konfigurációban meg lehet adni tájékoztató információkat az egyes taxonómia csoportokhoz, melyek segítik a felhasználót a tájékozódásban (például, hogy mely adatforrásokat használja az adott taxonómia).

ÖSSZEGZÉS

A bemutatott rendszer és módszer lényege, hogy könnyebben és gyorsabban feldolgozhatóvá tegye a publikus forrásokból (főleg internet) származó információkat. Egyfelől közös felületről teszi elérhetővé az információforrásokat, amelyeket egyszerre lehet keresni, másfelől mindezt olyan előre létrehozható tudásházterek segítségével, melyek adott témákat sokkal mélyebben és pontosabban tudnak meghatározni és keresni, mint a hétköznapi keresési módok. Ezeket a tudásházterekat a rendszer egyszerre tudja használni minden adatforrásán, ami azt jelenti, hogy képes az információt annak forrásától függetlenül bármilyen szempontrendszer szerint kategorizálni.

Egy ilyen rendszer képes értesítési szolgáltatások megvalósítására is, ami azt jelenti, hogy az egyes felhasználók létrehozhatnak saját előfizetéseket, melyek számukra érdekes témákhoz kapcsolódnak. A rendszer folyamatosan figyeli a kijelölt forrásokat, és ha új, friss információ érkezik az adott témában, akkor a felhasználó saját felületén az automatizáltan megjelenik.

Mindezek a szolgáltatások egyszerre teszik lehetővé, hogy mélyebben feldolgozhatóvá, mindemellett gyorsabbá tegyük az elemző munkával járó információgyűjtési tevékenységet.

FELHASZNÁLT IRODALOM

[1] TIDD, J.-BESSANT, J. R.-PAWITT, K.: Managing Innovation: Integrating Technological, Market and Organizational Change. – John Wiley & Sons, 2005.

[2] Intelligence Exploitation of the Internet. – SACLANT, 2002 okt.

[3] NATO Open Source Intelligence Reader. – SACLANT, 2002 okt.

[4] NATO Open Source Intelligence Handbook. – SACLANT, 2001 nov.