A szavak hálójában: szabadszavas mélyháló-kereső program

TIKK DOMONKOS, KARDKOVÁCS ZSOLT, MAGYAR GÁBOR

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
{tikk,kardkovacs,magyar}@tmit.bme.hu

Kulcsszavak: mélyháló, szabadszavas keresés, természetes nyelvi feldolgozás, kontextus-felismerés, SQL transzformáció

E cikk „A szavak hálójában" című projekt keretében készülő komplex internetes kereső/kérdező egyik modulját, a szabadszavas mélyháló keresőt ismerteti. A mélyháló, amely alatt az internetes adatbázisok tartalmát értjük, és amely a szokásos keresőmotorokkal nem elérhető, rendszerint pontosabb, frissebb és több információt tartalmaz, mint a statikus Internet-oldalak összessége, azaz a felszíni háló. Munkánkban a mélyhálón való keresés technológiai megoldására teszünk javaslatot, bemutatva egy olyan rendszert, amely a szabadszavas, azaz természetes magyar nyelvű kérdésekkel történő keresést is támogatja.

1. Bevezetés

A szavak hálójában című NKFP 0019/2002-es projektnek egy komplex internetes kereső/ kérdező eszköz létrehozása a célja, amely mind szöveges dokumentumok, mind képek közti keresések terén új technológiákat tartalmaz. A szövegeket a felhasználó az internetes adatbázisok tartalmában, a *mélyhálón* való szabadszavas, azaz magyar nyelvű, kerek egész kérdő mondatokkal *(természetes nyelvű kérdés)* keresheti.

Ez a keresési mód a felhasználó számára két jelentős előnnyel jár. Egyrészt lehetőséget ad a jó minőségű adatokat tartalmazó, hagyományos kereső-motorok segítségével nem elérhető tartalmak keresésére közös kiinduló pontból. Másrészt a szabadszavas kereséssel jelentősen egyszerűsödik az információigényt megfelelően reprezentáló keresőkifejezések megadása.

A képi keresés támogatására egy vizuális tezaurusz kerül kifejlesztésre, ami a képi tartalmak jellemzésére és indexelésére használható szöveges leírások mint tartalmi kategóriák rendszere, strukturált szótára. A vizuális tezauruszt képállományok jellemzésére javasoljuk standardként. Segítségével az adatgazda megfelelő és könnyen kereshető metainformációkkal láthatja el az általa közreadott képek tartalmát, segítve a képek tartalmában való hatékony keresést. Jelen tanulmányban az alkalmazás mélyháló-kereső részét mutatjuk be részletesen.

2. A mélyháló

2.1. A fogalom meghatározása

Évszázadokkal ezelőtt, ha valakinek olyan információra volt szüksége, amellyel közeli és távoli ismerősei nem rendelkeztek, felkeresett egy könyvtárat, hogy ott a megfelelő könyveket fellapozva megtudja, amire kíváncsi volt. Az idő múlásával, a tudományos haladással párhuzamosan az ismeretek megszerzése egyre nehezebbé vált, akár egy bizonyos témakört, akár az egyetemes tudást tekintve. Ezen könyvtár- vagy tágabban médiahálózatok létrehozásával segítettek.

Manapság az Internetet, s annak domináns alkalmazását, a Világhálót közhelyesen régi idők könyvtárához szokás hasonlítani: minden természetesen felmerülő kérdésre megadja a választ – ha tudjuk, hol keressük. Ugyanakkor, míg a könyvtárban bármikor tanácsért fordulhatunk a készséges alkalmazottakhoz, a Világhálón, ha léteznek is, nincsenek *helyben* a regionális szakértők sem, akik útbaigazítanának a gombamód szaporodó oldalak között.

A szakértők helyét a modern keresőrobotok, vagy a keresőmotorok vették át. A Világháló eredeti felépítése tette lehetővé azt, hogy dokumentumok egymáshoz kapcsolódó halmazaként alapvetően bárki számára – így egy gép számára is – bejárhatóvá váljon. A keresőmotorok hagyományosan ilyen kapcsokon, linkeken keresztül járják a Világhálót mind a mai napig. Az oldalakat jellemzően kézzel szerkesztették, ezért az ilyen hagyományos oldalakat statikus oldalaknak nevezzük a továbbiakban.

A fejlődés azonban nem állt meg a tartalomipar előretörésével. Szükségessé vált, hogy az oldalak kinézete, struktúrája jellemzően változatlan maradjon, egyes részei gyakrabban, mások lassabban frissítődjenek – gondoljuk csak a hírszolgáltatással foglalkozó oldalakra. Kialakultak a gépek által készített, illetve előállított portáloldalak rendszere – a teljesen dinamikusan előálló oldal. Ez viszont azt jelentette, hogy a Világhálónak egyre nagyobb számban keletkezett olyan része, mely kapcsokon (linkeken) keresztül nem elérhető, így a keresőmotorok azokat nem látják, és nem találják meg.

Az ezredforduló környékén végzett mérések szerint csupán a Világháló statikus része mintegy 2,5 milliárd dokumentumot számlál, s naponta 7,5 millióval gyarapodik mindenféle központi ellenőrzés, nyilvántartás nélkül. Ebből az is következik, hogy a korábbi könyvtári párhuzamot a keresés terén lehetetlen fenntartani: nem várhatunk el teljességre törekvő információszolgáltatást.

A mennyiségi expanzió mellett évek óta megfigyelhető az a tendencia is, hogy a dokumentumok egyre nagyobb hányada válik *dinamikussá*, vagyis a dokumentum lekérése nyomán áll elő, majd továbbítódik az igénylőhöz. A szolgáltató részéről ennek két oka van. Egyrészt így az adott igénynek megfelelően tudja előállítani a rendelkezésre álló, strukturáltan tárolt információkból az éppen szükséges adatokat. Ennek következtében nem kell hagyományosan szerkesztett dokumentumokon keresztül eljuttatni a felhasználók vélt igényeit kielégítő információt. Másrészt lehetőséget nyújt aktualizált dokumentumok előállítására, amelyben a *lekérés* pillanatában érvényes adatok szerepelnek. Ezáltal a dokumentumon belüli adatfrissítés is leegyszerűsödik.

A dinamikusan előálló oldalakat azonban a kereső robotok nem látják, sőt a Világháló expanziója és az oldalak megújulása miatt a statikus oldalaknak is egyre kisebb részét képesek felderíteni. A legnagyobb kereső 1998-ban még a Világháló 32%-át, 1999-ben már csak 16 %-át ismerte. Mivel szolgáltatók részéről egyre jellemzőbbé válik a tartalmak dinamikus generálása, ezért a keresőmotorok hatékonysága romlik.

A portálok oldalai „mögött" található, strukturált, jellemzően adatbázisokban tárolt, dinamikusan elérhető tartalmak összességét *mélyhálónak* nevezzük (deep web, DW). Az elnevezés a tartalom nehezebb elérhetőségére utal, szembeállítva azt a klasszikus, felszínen található tartalommal.

Fontos megjegyezni, hogy a mélyháló nem azonos a láthatatlan vagy fekete hálóval. Láthatatlan háló részét képezik azok az oldalak is, amelyek tűzfal mögött, intraweben, jelszóval védett vagy más, általánosan meg nem közelíthető módon érhetőek el. A mélyháló jellemzője, hogy elvben bárki hozzáférhet ezekhez az információkhoz, de szisztematikus, keresőmotorok általi bejárása nem volt lehetséges – legalábbis mostanáig.

A mélyhálóról készült tanulmány [1] szerint csak a legnagyobb 60 mély adatbázisban 40-szer annyi adat van, mint a felszínen. Az összes adatot figyelembe véve mintegy 500:1 arány adódik, mindez kb. 200 ezer szolgáltatót jelent. Az átlagos méretű mély szolgáltató 5,43 millió adatrekorddal rendelkezik, de a méret szerinti középső mindössze 4950-nel. A mély oldalakat havonta átlagosan fele annyian látogatják, mint a felszínieket, de a mediánt kétszer annyian. Ezek a számok óriási mértékű adatkoncentrációra utalnak. A mélyháló mérete a becslések szerint lényegesen gyorsabban növekszik a felszíninél.

A nyomtatott adatokhoz viszonyítva is elképesztő a növekedés: 1998-ban nagyjából megegyezett a kettő, majd 2000-ben a mély háló javára billent a mérleg hétszeres aránnyal, s várhatóan 2003-ban elérte a 60-szorost is. Keresés esetén a *mély oldalak* nagyjából *10%-kal több találatot jelentenek* és empirikus mérések alapján közülük nagyjából *háromszor annyi az értékes,* mint a felszíniek esetében. A felsorolt tényezők miatt indokolt a mély oldalak bevonása az internetes keresési térbe.

2.2. A mélyháló keresése

Az interneten elérhető adatbázisok, akárcsak más adatbázisok, nem csupán szintaktikusan, hanem szemantikusan is strukturáltak – azaz az információegységek egyértelműen azonosíthatóak bennük. Ez hatalmas előny a Világháló más (adatokat tartalmazó) dokumentumaihoz képest, ám egyúttal hatalmas kihívás is, mert egy átfogó mélyháló-keresőnek egységesítenie kell ezen adatbázisok által leírt világot, annak ellenére, hogy a szolgáltatók a való világ ugyanazon elemeit jellemzően különböző módon modellezik (pl. más nyelven).

A mélyhálós adatbázisok kereshetőségének feltétele, hogy a kereső(motor) megfelelő információkkal rendelkezzen a tárolt adatokról és az adatbázisok struktúrájáról. Ez csak a kereső és mélytartalom-szolgáltatók közötti együttműködéssel valósulhat meg. A szolgáltatónak tehát biztosítania kell a tárolt adatokra vonatkozó adatbázis-hozzáférésen kívül egy további, immár metaadatok (sémainformációk) kinyerését támogató adatbázis-csatlakozást, de legalább egy kapcsolódási pontot, interfészt is. Az adatgazdának így nem kell foglalkoznia a metaadatok olyan módú előállításával, mely az összeillesztést lehetővé teszi, hanem annak kinyerését a kereső alkalmazásaira bízhatja. Ez egyrészt számára kisebb fáradtsággal jár, kevesebb erőforrást igényel, másrészt az integrátor mélyháló-kereső motort is jóval kevesebb leírási mód értelmezésére kényszeríti. Jelenleg a szemantikus információ leírására ugyanis rengeteg, széleskörűen elterjedt, egymással többé-kevésbé kompatíbilis megegyezés létezik (gondoljunk csak a városok és nevek kódolásának sokszínűségére), viszont sémainformációt az adatbázisok körében annyiféleképp ábrázolnak, ahány adatmodell létezik – itt gyakorlatilag vagy a relációs, vagy az objektumorientált típus jöhet szóba.

Ezek közül a ma leginkább az SQL nyelven lekérdezhető relációs adatbázis-kezelők használatosak, vagy röviden az SQL adatbázisok. Ezek jól definiált sémainformációval rendelkeznek, ami a hatékony, értelmes (szemantikus) kereséshez nélkülözhetetlen. Ugyanakkor az előző bekezdésben felvázolt kooperáción alapuló mélyháló-keresési stratégia keretében elégséges is.

2.3. Kulcsszó alapú vagy szabadszavas keresés

A hagyományos keresőmotorok a Világháló feltérképezése során a dokumentumokat indexelve katalógusállományokat készítenek. Egy adott keresés során a katalógusállományok és a keresőkifejezés szavainak összevetésével határozzák meg az eredményt. Szemantikus alapú keresést nem tudnak megvalósítani, mivel a keresőkifejezés szavairól szemantikus információ nem áll rendelkezésre. Hasonlattal élve, a mai keresők olyanok mint a szóelemzők, amelyek a mondat értelmétől, mondattani szerepüktől függetlenül értelmezik – gyakran tévesen – a szavakat.

A sémainformációra alapuló mélyháló-keresőben ez a megoldás nem járható út. Ennek oka egyrészt az,

hogy szemantikus információ nélkül nem lehet eldöntetni, hogy mely adatbázis melyik sémájában kell a keresést végrehajtani. Másrészt, ha a keresés eredményt hoz – például minden séma minden mezejére illesztve a keresőkifejezés szavait –, akkor annak interpretálása is problémát jelent.

A kulcsszó alapú keresés azonban hagyományos keresőmotorok használatakor sem vezet gyakran eredményre. A felhasználónak ugyanis olyan keresőkifejezést kell megadni, amelynek elemei (szavai) vélhetően szerepelnek majd azon az oldalon – tehát már legalábbis részlegesen rendelkeznie információkkal a válaszoldalról – ahol az információigényt kielégítő tartalom is szerepel. Ehhez a felhasználónak ki kell találnia, hogy milyen szövegkörnyezetben szerepelhet a keresett tartalom, ennek hiányában ugyanis a keresése sikertelen lesz.

További gondot jelenthet, hogy túl általános kifejezéseket használva feldolgozhatatlanul nagy mennyiségű válaszoldalt ad vissza a kereső, míg pontosan specifikált keresőkifejezések szavai együttesen gyakran egyetlen dokumentumban sem fordulnak elő. A felhasználónak alkalmazkodnia kell a gépi keresés technológiájához, igazán eredményesen csak akkor tudja a keresőket használni, ha megérti azok működési elvét, és sajátjává teszi ezt a „gondolkodásmódot". Természetes nyelvi kérdések esetén a kérdés fókuszát a kérdőszó (ki, hol, mikor stb.) határozza meg, azonban kulcsszó alapú kérdezésnél hiba lenne elvárni, hogy a kérdőszó a válaszban szerepeljen.

E problémák feloldását a természetes nyelvű kérdésfeltevés megengedése, a szabadszavas keresés jelentheti. Nyilvánvaló, hogy a természetes nyelvű kérdések gépi „megértése" csak nyelv szintaktikai szabályainak, valamint szemantikai elemeinek bizonyos részét tartalmazó tudáskomponensek birtokában lehetséges. A mélyháló tartalmában való keresés esetén a sémainformációba kódolt szemantikus adatok már kiindulást jelenthetnek a megfelelő tudásbázis felépítésére. A következő fejezetben bemutatásra kerül a projekt által kidolgozott szabadszavas keresést támogató mélyháló-kereső alkalmazás felépítése és működése.

3. Szabadszavas keresést támogató mélyháló-kereső

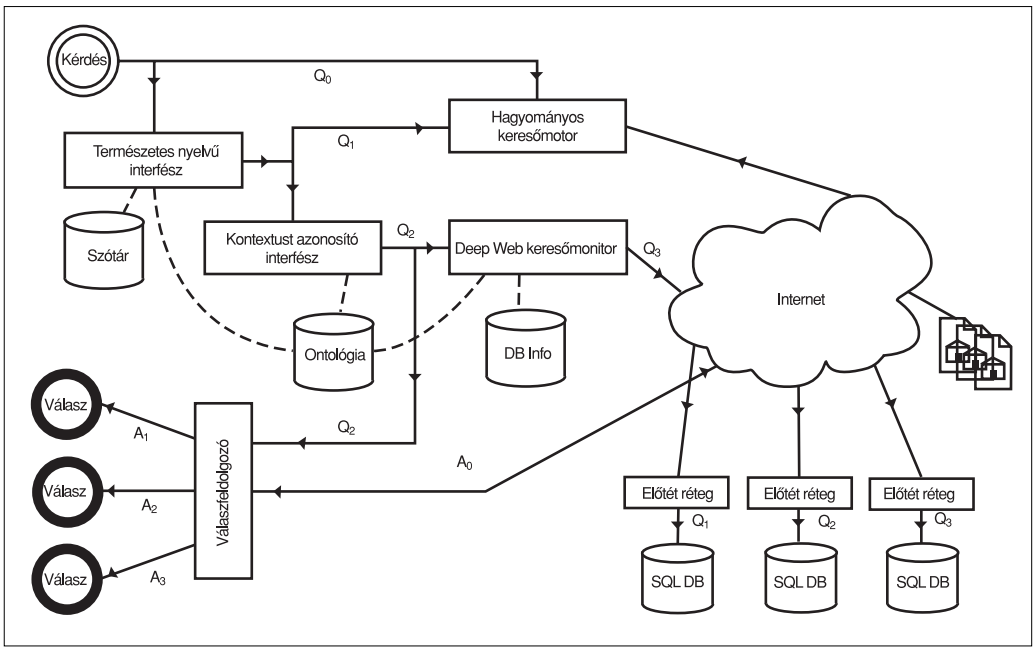
3.1 A rendszer áttekintése

A projekt keretében megvalósuló mélyháló-kereső prototípus-alkalmazás a mélyhálón jelenleg böngészővel el nem érhető, általában adatbázisban található tartalom egy részét kívánja elérhetővé tenni, amelyek a könyv, film, labdarúgás és étterem témakörébe esnek. Ennek érdekében a projekt felvette a kapcsolatot néhány, a fenti témakörökben érintett tartalomszolgáltatókkal (Országos Széchényi Könyvtár, Fókusz Online Könyváruház, port.hu, Axelero, eszemiszom.hu).

A mélyháló-kereső feladata a természetes nyelvű kérdés feldolgozása. Ez egyfelől a természetes nyelvű kérdések SQL lekérdező nyelvre való fordítását; másrészt a kifejezés-alternatívák továbbítását jelenti az SQL adatbázisok, és az onnan jövő válaszok kezelését, valamint az eredmények megjelenítését a felhasználó felé.

A mélyháló-kereső csak olyan jellegű kérdésekre képes válaszolni, amelyre a válasz megtalálható a mélytartalmat szolgáltató partnerek adatbázisaiban. Ez természetesen megszorításokat jelent a kérdés típusára, jellegére és témájára vonatkozóan.

1. A keresőmotor csak olyan egyszerű, azaz nem összetett, kérdőszóval kezdődő, a magyar nyelvtan és helyesírás szabályainak megfelelő kérdőmondatokat fogad el, melyek a mélyhálós partneradatbázisok által lefedett információtér-szegmens elemeire vonatkoznak. Néhány további, nem túl szigorú megszorítást alkalmazunk a kérdőszavakra, illetve bizonyos nyelvtani szerkezetekre vonatkozóan.
2. A keresőmotor tehát nem fogad el, illetve nem garantálja a jó választ eldöntendő, szubjektív, intencionális, kauzális, valamint az adatbázisban jellemzően nem tárolt információra vonatkozó kérdésekre.



1. ábra
A rendszer moduláris vázlata

A rendszer moduláris vázlata az *1. ábrán* látható. A felhasználó kérdése először a természetes nyelvi feldolgozó (angolból származóan a továbbiakban *NL*) modulhoz kerül, e modul bemenete egyúttal az egész mélyháló-kereső szoftverrendszer általános bemenete is. Hatékonyság vizsgálat esetén a mélyháló-kereső válaszait egy beépített, hagyományos keresőmotor találataival hasonlítjuk össze. Ilyen esetben tehát a kérdést a rendszer hagyományos keresőmotorhoz is továbbítja, de ezzel az ággal a továbbiakban nem foglalkozunk.

Az NL modul a kérdés nyelvi feldolgozását, releváns lexikai egységekre való bontását (tokenizáció), valamint morfológiai és szintaktikai elemzését végzi. Természetesen e feladatok elvégzéséhez különböző tudáskomponensekre (például (szó)tárakra, ontológiára) és segédeszközökre van szükség; ezeket a 3.2. szakaszban ismertetjük. Kimenete szintaktikailag elemzett, zárójelezett mondatalternatívák listája. A zárójelezett, szintaktikailag elemzett mondatokat a *kontextus felismerő modul* átalakítja az általunk definiált CL (Context Language) kifejezésekké, amely már a kontextusra vonatkozó információkat is tartalmazza. Ez képezi a mélyháló-kereső motorjának (angol megfelelőjéből a továbbiakban *DW motor*) a bemenetét.

A DW motor feladata többrétű. Egyrészt a kontextus-információk, az aktuális kérdés tárgya, a kérdésben szereplő tulajdonságok és a rendelkezésre álló adatbázis-leírók (Database Information, DB Info) segítségével a kérdés megválaszolására alkalmasnak tartott adatbázisok meghatározása. Másrészt a bemenetére érkező CL formalizált mondatokból a megfelelő adatbázisok felé küldendő, az adatbázisra jellemző, de szabványos sémákra illeszkedő speciális SQL lekérdezések (DWL nyelvű lekérdezések) előállítása és ezek továbbítása. Ez utóbbi feladat magában foglalja az adott adatbázisra vonatkozó DB Info alapján a történő átalakításokat.

A mélytartalom-szolgáltatók weboldalán működik a DWL nyelvű lekérdezéseket értelmező *előtét réteg,* mely végrehatja a helyi adatbázis-kezelőtől és adatmodell megvalósítástól függő DWL→SQL módosításokat, valamint ellátja és felügyeli a megfelelő jogosultsági és biztonsági feladatokat.

A tartalomszolgáltatótól kapott választ (amely több találatot is tartalmazhat) az előtét réteg továbbítja a *válaszfeldolgozó modulnak.* Ez összegyűjti az egyes adatbázisoktól beérkező eredményeket, és azokat különböző szempontok (például beérkezési idő, felhasználói profil, korábbi keresések felhasználói szokások alapján történő kiértékelése nyomán kialakult forráskontextus relevancia) szerint rangsorolja, és megjeleníti a felhasználó számára.

A továbbiakban részletesen bemutatjuk a mélyhálókereső rendszer legfontosabb komponenseit.

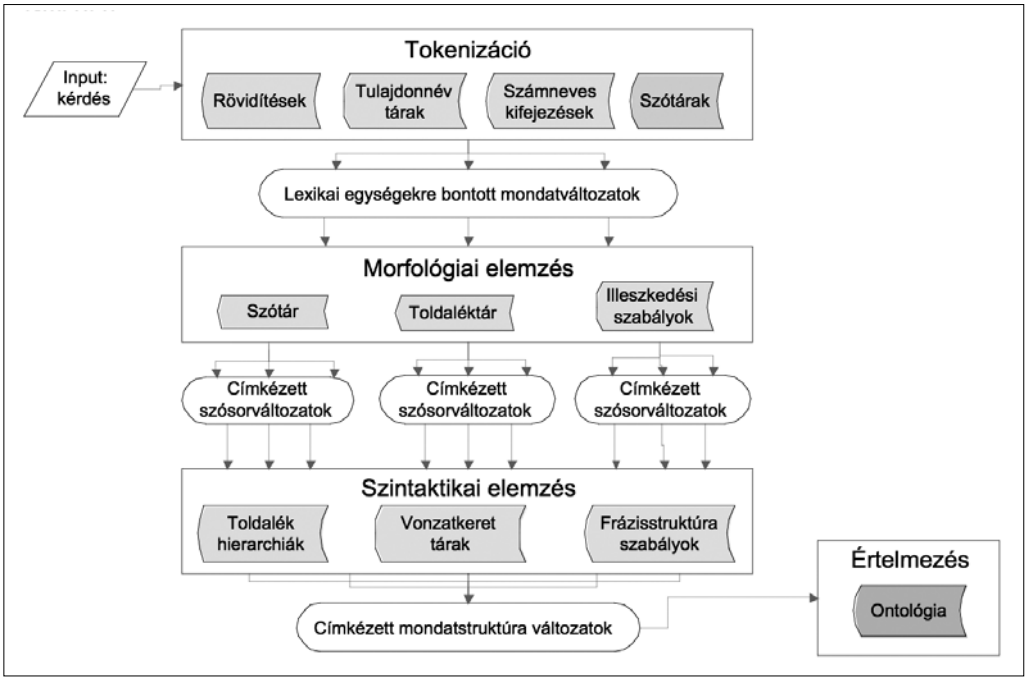
3.2. Az NL modul

Az NL modul a természetes nyelvi kérdést a számítógép által könnyen kezelhető formára alakítja át. A transzformáció lépéseit és a felhasznált tárakat és segédeszközöket a *2. ábra* ismerteti.

A kérdés feldolgozása két fő szakaszból áll. Az első szakaszban a mondatot releváns lexikai egységekre bontjuk, és ezen egységeknek elvégezzük a morfológiai elemzését. A második szakaszban meghatározzuk a mondat frázisait (szavaknál nagyobb mondatbeli egységeket) és azonosítjuk a lényeges nyelvtani szerkezeteket.

A későbbi üzemszerű működés gyorsítása céljából a gyakori kérdéstípusokra egy cache-tár felhasználásával kérdésséma-alapú mintaillesztést alkalmazunk a második szakasz előtt, amivel jelentős sebességnövekedés érhető el, ugyanis ekkor az ismert kérdésekre, illetve kérdéstípusokra kockázat nélkül kihagyható a bonyolult feldolgozó módszereket tartalmazó második szakasz. Ha az aktuális kérdésre nincs a cache-tárban megfelelő kérdésséma, akkor végrehajtódik a második szakasz.

Az első szakasz legfontosabb része, hogy a tulajdonnév jellegű, valamint a későbbiekben kiemelten kezelt entitásokat (például rövidítések, címek, dátumok,



2. ábra
A kérdéstranszformátor

tulajdonnevek, pénznemek számnévvel, e-mail és honlapcímek stb. – összefoglaló névvel: névelemek) tartalmazó különböző tárak segítségével felismerjük, a megfelelő névelem szerint címkézzük, és a továbbiakban egy egységként *(tokenként)* kezeljük, akkor is, ha több szóból állnak. A névelemként fel nem ismert szavakat a morfológiai elemzés során morfológia jegyeikkel (szófaj, toldelékok) címkézzük fel. Hasonlóan a névelemként címkézett tokenek toldalékait is morfológiai elemző segítségével határozzuk meg.

Fontos megjegyezni, hogy ha az elemzés bármely fázisában több lehetséges megoldás adódik (például morfológiai elemzésnél a homonimák esetén: *ég* [ige], *ég* [fn]), akkor azokat párhuzamosan, külön alternatívaként kezeljük, és minden ilyen elágazás új mondatváltozatokat generál.

A morfológiai jegyekkel felcímkézett mondatváltozatok összetartozó szintaktikai egységeit a zárójelező modul végzi, amely toldalékra és szófajra vonatkozó információk alapján egy szabályrendszer segítségével felismeri a legfontosabb szerkezeteket, például főnévi csoport, birtokos szerkezet, névutó, logikai operátorok, igei szerkezetek stb. Az eredményként keletkező zárójelezett és felcímkézett változatokat a kontextusfelismerő tudásbázisa alapján lehet értelmezni.

3.3. A kontextusfelismerő

A kontextusfelismerő feladata a szintaktikailag már elemzett, zárójelezett mondatváltozatokra meghatározni a megfelelő sémát vagy sémákat, amely(ek)ből vélhetően a kérdésre válasz adható. A kontextusfelismerő sémái és azok attribútumai a partneradatbázisokban elérhető elemek megfelelően absztrahált változatai.

Az eljárás részét képezi a szintaktikai elemzés szemantikai vizsgálata is, amelynek során a jelzős, logikai és más nyelvtani szerkezetekből egy értelmezés, interpretáció keletkezik. Az interpretáció maga egy logikai következtetés, amelynek a végén a releváns séma is előáll. Az eredményeket egy SQL-hez és XML-hez egyaránt közel álló belső nyelven, az úgynevezett CL nyelven állítja elő – ez kerül a DW motor bemenetére.

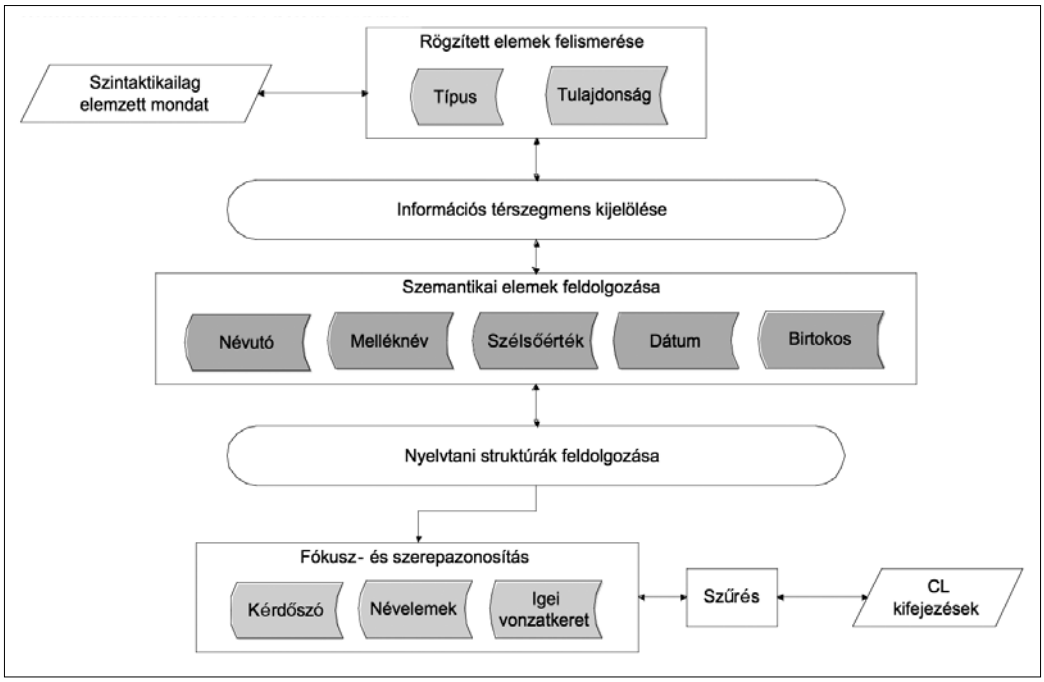
A modul a következőkben felsorolt erőforrásokat veszi igénybe:

• *Séma- és attribútumnévtár:* itt tároljuk azokat séma- és tulajdonságneveket (attribútumokat), amelyek a mélyháló-kereső aktuális témaköreinek leírásához szükségesek. A sémák attribútumait a sémában, vagy jellemző tartalmuknak nyelvtani struktúrákban betöltött szerepe alapján különböző különleges annotációkkal lehetnek ellátva.

• *Névelemek és kérdőszavak tára:* rendre a névelemekhez és kérdőszavakhoz tartozó sémák tárolására szolgál.

• *Igei vonzattár:* minden igéhez, amely a mélyháló kereső aktuális tematikájában szerepet játszhat, hozzárendeljük lehetséges vonzatainak halmazát. Egy igéhez több vonzathalmaz is tartozhat, amennyiben azok kizárják egymást, például az ige eltérő jelentéséből adódóan. Minden esetben a lehető legbővebb vonzathalmazt tároljuk. A vonzatok alapján következtetni lehet, hogy az adott szó milyen környezetben és értelemben, illetve milyen tulajdonság értékének feleltethető meg. Fontos látni, hogy a vonzattárra mindenképpen szükség van, hiszen ige és szavak gyökei, azok halmazai nem határozzák meg a szemantikai szerepeket; például „Mikor látogatta meg Bush Putyin elnököt?” és a „Mikor látogatta meg Putyin Bush elnököt?”. Hasonlóan: „Hol adnak virslit és zöldborsófőzeléket?”, „Hol adnak virslit zöldborsófőzelékkel?” és „Hol adna virslit zöldborsófőzelékért?” lényegesen különböző értelmű mondatok.

A kontextusfelismerő *(3. ábra)* először a kérdésben szereplő rögzített elemeket keresi (séma-, illetve attribútumnévtárban előforduló kifejezések), hiszen ezek előfordulása meghatározó lehet az információs térszegmens kijelölésében, vagyis a kérdés kontextusának kiválasztásában. Ezt követi a nyelvtani szerkezetek szemantikai feldolgozása, ahol az egyes elemek *értéke* alapján kényszerfeltétel-rendszert hozunk létre a lehetséges kontextusokra, illetve tulajdonságokra vonatkozóan.



3. ábra
A kontextusfelismerő működési vázlata

Ez a feltételrendszer tovább bővül a fókusz- és szerepazonosítás során, amikor kérdőszót, a névelemeket és az igei vonzatszerkezetet dolgozzuk fel. Kiemelt jelentőségűek köztük a névelemek, amelyek a legtöbb kérdésben előfordulnak, de az adatbázisok is jellemzően egyértelműen meghatározott entitásokra vonatkozó tényinformációkat tartalmaznak. Végül a kényszerfeltétel-rendszer megoldásaként a modul kiszűri az ellentmondásos kontextusokat és elkészíti a lehetséges kontextusokat és tulajdonságait leíró CL kifejezéseket.

A kérdésfeldolgozás lépéseit a *„Mikor játsszák a Mátrixot Budapesten”* példamondaton szemléltetjük.

1. NL modul; névelemek felismerése:
   Mátrix és *Budapest* entitásokat megtalálja a megfelelő névelemtárban.
2. NL modul; morfológiai elemzés:
   minden szónak megadja a morfológiai jegyeit, például *játsszák* = játszik [ige] + kijelentő mód jelen idő T/3 alak, vagy játszik [ige] + felszólító mód T/3 alak, *Mátrixot* = Mátrix [névelem] + tárgyrag, *Budapesten* = Budapest [névelem] + helyhatározó rag.
   Megjegyzés: a *játsszák* kétféle morfológiai elemzése miatt két a továbbiakban alternatívát kezelünk.
3. NL modul; zárójelezés eredménye:
   (Mikor) (játsszák) (a Mátrixot) (Budapesten).
4. Kontextusfelismerés;
   névelemek szerepének meghatározása:
   Mátrix – **film**, *Budapest* – **város** azonosítása.
5. Kontextusfelismerés; fókusz meghatározása:
   Mikor kérdőszó **dátum**ra vagy **időpont**ra vonatkozik.
6. Kontextusfelismerés;
   igei vonzatszerkezet feldolgozása:
   játszik+tárgy+helyhatározó vonzatséma illesztése, és a vonzatok szerepének meghatározása (tárgy = film, szerep, ...; helyhatározó = város, ...).
7. Kontextusfelismerés;
   szűrés és CL kifejezés előállítása:

```
Context = Esemény
Idôpont = ?
Esemény = ( Context  = Mûsor
            Cím     = Mátrix
            Hely    =  (    Context = Mozi
                            Város = Budapest ))
```

3.4. A mélyháló-kereső motorja és a tartalomszolgáltatókkal való kommunikáció

A *mélyháló-kereső* motorja az alábbi feladatokat látja el:

1. Relevanciafelismerés: CL kifejezés kontextusa, illetve a kitöltött tulajdonságmezők alapján kiválasztja azon adatforrásokat, amelyek elvileg képesek a formalizált CL kifejezésben kódolt kérdés megválaszolására. A kiválasztásnál figyelembe veszi, hogy minden hivatkozott fogalom (séma) létezzen az adott adatforrásnál, és a feltételként és kimenetként megszabott tulajdonságokat az adatforrás tárolja. Erőforrásként felhasználja a DB Info-ból kinyert úgynevezett Relevancia Táblát.
2. DWL-konvertálás: A CL kifejezést speciális DWL nyelvjárású SQL lekérdezéssé alakítja.
3. Szabványos egységek kezelése: a különböző, változó ábrázolású adatelemeket hozza egységes formára.
4. Hitelesítés, azonosítás: Vezérli a partneradatbázisokkal való kommunikációt, például hitelesítést, kérdés-válasz azonosítást és általában a biztonságos működést.

A mélyháló-kereső a mélytartalom-szolgáltató partnerek oldalán elhelyezett *előtét rétegen* keresztül kommunikál a webhely adatbázisaival, tehát a DWL lekérdezést is ezeken keresztül továbbítja az adatbázisok felé. Az előtét réteg feladata a webhely keresőszolgálathoz való csatlakoztatása, a jogosultságok ellenőrzése, a hitelesség, a lekérdezhetőség és a válaszküldés biztosítása. Hatóköre kizárólag a lekérdezhetőséggel összefüggő teendők ellátására szorítkozik, adatot nem tárol, folyamatokat nem indít be.

Az előtét réteg inicializálását az adatgazdák végzik, amikor csatlakoznak a mélyhálós keresőszolgáltatáshoz. Ekkor az adatgazda meghatározza, hogy a mélyháló-kereső által ismert témák közül melyekről tárol információt, és ezekből pontosan milyen adatokat kíván a mélyháló-keresőn keresztül elérhetővé tenni. Ezen adatokból készül a DB Info tár, amely ezeket az információkat a mélyhálós keresőszolgáltató oldalán tárolja; ez alapján választja ki a relevanciafelismerő, hogy a bejövő kérdéseket mely adatbázishoz kell elküldeni. Abban az esetben, ha az adatgazdánál az előtét réteg módosítását igénylő adatváltozás történik, akkor egy adminisztrációs felületen keresztül frissíthető a DB Info tartalma.

Az előtét réteghez továbbított információ három nagyobb, jól azonosítható részből áll; a kérdésazonosítóból, az SQL (DWL) lekérdezésből és az azonosításhoz szükséges elemekből. Az azonosítás ellenőrzése után az előtét réteg beállításai alapján az SQL lekérdezésben levő tábla- és mezőneveket az adott adatbázis elnevezési konvenciói alapján le kell cserélni. (Ezeket az információkat az adatgazda szolgáltatja az előtét réteg inicializálásakor.) A mezőnevek lecserélésekor, ha a helyi megvalósítás SQL lekérdezést támogat, akkor SQL transzformáción, minden más esetben csak az egyes nevek alkalmazásfüggő átnevezésen mennek keresztül.

A válaszadást hasonlóan kell megvalósítani. A válasz mindenképpen tartalmazza az eredeti kérdésazonosítót, a válaszok számát – ha az nem haladta meg a felsőkorlátot, de legalább egy elemet tartalmaz – továbbá a válaszok leírását, és a hitelességet garantáló mezőket digitális aláírás és nyilvános kulcsú titkosítás formájában. Ez utóbbi általában állandó magánhálózat, úgynevezett VPN (Virtual Private Network) keretein belül is megvalósíthatjuk – ami a projekt keretében meg is valósul.

Már a lekérdezés és a válaszadás során is előkerültek a biztonsági kérdések. A biztonság mindenekelőtt az azonosítható kérdezőt és válaszadót jelenti, de éppen úgy az egyes résztvevő felek szuverenitását és jogvédelmi kérdéseket is magában foglalja – a szemantikus web kezdeményezéssel összhangban.

A működés során két nyilvános kulcsú titkosítást, vagy kódolt VPN hálózatot alkalmazunk biztonságtechnikai protokollként. A titkosítás a motor és az előtét réteg adatcseréjére terjed ki.

A *válaszfeldolgozó modul* feladata, az egyes adatbázisoktól beérkező eredmények összegyűjtése és rendezése. Mivel az eddigi felméréseink szerint a különböző adatforrásoktól kapott adatok típusa heterogén (azaz hol rekordok, hol rekordok halmaza, vagy csak egy URL), ezért *válaszként a felhasználó az adatforrás választ tartalmazó oldalára mutató linket kapja keresőszolgáltatótól*. Ezzel megvalósul a szolgáltatás biztonsága is, hiszen információszivárgásra, illetve elszívásra nem kerülhet sor.

A válaszok helyességének elemzése rendkívül időigényes feladat lenne, ezért a válaszokat az alábbiak szerint csoportosíthatjuk, illetve rangsorolhatjuk:

– Amennyiben az eredeti kérdés értelmezése nem egyértelmű, akkor az abból generált különböző kérdésreprezentációk szerint;
– adatforrás és azon belül séma szerint;
– a válaszok beérkezési sorrendje szerint;
– felhasználói profil szerint;
– és végül a korábbi keresések felhasználói szokások alapján történő kiértékelése nyomán kialakult forrás-kontextus relevancia alapján.

Az érvényes rendezési módszert a felhasználó választhatja ki, amit egy cookie segítségével tárol a mélyháló kereső.

4. Nemzetközi összehasonlítás

A projekt által tervezett szabadszavas mélyháló-kereső alkalmazás nemzetközi viszonylatban is élenjáró technológiákat tartalmaz. Egyedülálló módon széleskörű természetes nyelvű feldolgozást valósít meg, aminek segítségével képes magyar nyelvű kérdőmondatokat SQL lekérdezésekké átalakítani, továbbá egy olyan komplex internetes keresőszolgáltatást javasol, amely három keresési technológiát integrál (felszíni, mélyhálós, illetve az itt nem részletezett vizuális tezaurusszal indexelt képi).

A projekt által integráltan kezelt feladatokra különállóan már léteznek technológiák. Az Interneten több mélyháló-kereső is található, amely angol nyelvű adatbázisokkal van kapcsolatban. Ilyen például a BrightPlanet CompletePlanet [2] nevű keresője, amely a mélyhálós oldalak több mint felét indexeli, de ez csak kulcsszó alapú keresést támogat. A keresések eredménye ezért többnyire a mélyhálós tartalomszolgáltatók főoldalára mutat, ahol a felhasználónak kell megtalálnia a keresett információt. Hasonlóan kulcsszó alapú keresőszolgáltatást nyújt még a ProFusion [3] és a Copernic [4] is.

Az integrált, képi keresést is támogató szolgáltatások egyre elterjedtebbé válnak, hiszen újabban már a Google és Yahoo! is támogatja ezeket, míg korábban csak olyan kisebb kereső oldalak nyújtották, mint például az iBoogie [5]. A szolgáltatások a képeket fájlnevek, illetve az esetleges egyéb képhez csatolt információk alapján indexelik, a képi tartalomban való keresést, annak bonyolultsága miatt egyik sem teszi lehetővé.

Nyelvtechnológiai projektek keretében főleg angol nyelvű szabadszavas kérdező-válaszoló rendszerek ismertek, melyek közül például az MIT fejlesztett START [6] projekt az Internetről összegyűjtött információk alapján válaszol. Hasonló módon dolgozik az Answerbus [7] és az AskJeeves [8] kereső is.

E tanulmány zárásaként a néhány legismertebb szabadszavas kérdező rendszer működését mutatjuk be egy példán keresztül. A „*When does the Siam Cuisine Restaurant open?*” kérdésre az alábbi válaszok születtek:

1. START:
   Unfortunately, I wasn't told when Siam Cuisine Restaurant opens.
2. Answerbus:
   Siam Orchids Authentic Thai Cuisine Restaurant was opened on February 5, 2003.
3. AskJeeves:
   This Center City location is open for lunch and dinner seven days a week.

5. Összefoglalás

Cikkünkben ismertettük „A szavak hálójában” projekt keretében megvalósítandó komplex internetes kereső eszköz mélyháló-kereső moduljának architektúráját. Bemutattuk az egyes részfeladatokat ellátó egységek működését.

A kézirat leadásakor az itt bemutatott rendszer megvalósítása a befejezéséhez közelít. A már elkészült részegységek a tesztelésre használt kérdésgyűjteményen hatékonyan működnek. A közeljövőben tervezzük a szoftver összekapcsolását az első mélyhálós adatbázisokkal, majd annak sikeressége esetén kibővítjük a lefedett témák körét, majd a keresőrendszert nyilvánosan elérhetővé tesszük az Interneten.

Irodalom

[1] M. K. Bergman, Deep Content, 2001, http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp
[2] http://www.completeplanet.com
[3] http://www.profusion.com
[4] http://www.copernic.com
[5] http://www.iboogie.com
[6] http://www.ai.mit.edu/projects/infolab/
[7] http://www.answerbus.com/index.shtml
[8] http://www.ask.com/