TMT 52. évf. 2005. 5. sz.

Varga Katalin

Intelligens információkereső rendszerek

Automatizálási lehetőségek és projektek a szövegelemzésben

A tanulmány a szerző nemrégiben megvédett PhD disszertációjának része. A disszertáció a tartalmi elemzés és feltárás egyik legégetőbb problémakörével, a szövegek elemzésének legújabb, legkorszerűbb módszereivel foglalkozik, ezen belül is hangsúlyosan az automatizálás lehetőségeivel. A számítógépes nyelvészeti kutatások igen előrehaladottak ezen a területen, az érdekes és gondolatébresztő kísérletek azonban még mindig csak viszonylag szűk térben működőképesek. Ezekből a kutatásokból ad a tanulmány egy kis ízelítőt, azzal a nem titkolt céllal, hogy a könyvtárak megértsék, most kell megtalálniuk a helyüket az új igények piacán, mielőtt tényleg mások veszik a kezükbe a minőségi információszolgáltatás kulcsát.

A növekvő információmennyiség, a minőségi információs szolgáltatások iránt fokozódó igények és a technika rohamos fejlődése az információtudományi kutatás-fejlesztés számára az automatizálás kérdéskörét állítja fókuszba. Az elektronikus dokumentumok terjedésével együtt nő a probléma, hogyan igazodjunk el az információk között. Mivel a tartalmi feltárás az egyik legidőigényesebb és legdrágább munkafolyamat, mind több kutatás irányul az automatikus megoldások keresésére. A szövegek jelenléte és gépi kezelhetősége kézenfekvővé teszi a tartalmi feltáró eszközök automatikus meghatározási módszereinek alkalmazását. A kutatások rendkívül figyelemreméltóak, az emberi intelligenciát azonban még nem sikerült mesterséges intelligenciával felváltani. A mai napig nincs olyan működő projekt, amely teljes egészében automatikusan tudja elvégezni a tartalmi feltárás feladatait.

A legintenzívebb kutatások az információkereső rendszerek területén folynak, itt csapódnak le az elvárások, és itt a legerősebb a verseny is. A kutatási irányok a szövegelemzés irányába mutatnak. A cél, hogy a keresőrendszerek lássák el az információfeldolgozás feladatát is, vagyis ne legyen szükség a szövegeket képviselő szurrogátumokra. Ezek a rendszerek arra épülnek, hogy a teljes szövegek képesek legjobban képviselni önmagukat, az intellektuális energiákat pedig a keresési oldalon kell befektetni.

A mai elvárások szerint korszerűnek minősíthető információkereső rendszer tartalmi alapú hozzáférést biztosít, interaktív, integrálni tudja a különböző médiatípusokat, nyelvtől független, és azonnal tud reagálni a változó felhasználói igényekre. Ezeknek a tényezőknek együttesen kell befolyásolniuk a tervezést. A természetes nyelven alapuló szövegelemző és -kereső rendszerek erősen függnek a nyelvi feldolgozás mélységétől és pontosságától. Többek között az alábbi magasabb szintű elvárásoknak kell eleget tenniük:

- A válogatás támogatása tartalmi kivonatok segítségével.
- Rugalmas, többszintű nyelvi elemzés.
- Többnyelvű keresési lehetőség.
- Különböző navigációs eszközök.
- Az igényeknek megfelelő tudásbázisok integrálása.
- Különböző információs források egy platformon történő kereshetősége (pl. bibliográfiai adatok és webforrások).

Az információkeresés modern rendszerei nem állhatnak meg a természetes nyelvi szövegeknél, éppúgy meg kell találni a hangzó, video-, multimédia szövegek kereshetőségét is. A kutatások a tartalom alapján történő keresésre koncentrálnak. A keresőrendszerek számára olyan felületeket kell tervezni, amelyek segítségével a felhasználó természetes nyelven tud kommunikálni a rendszerrel, és keresni a szövegek között. Ezek a kérdés-felelet rendszerek szintén tudásbázisokon, illetve a mesterséges intelligencia és a szakértői rendszerek alkalmazásán alapulnak. Az információkereső rendszereknek elemezniük kell a szolgáltatás tárgyát jelentő szövegeket és a kérdéseket egyaránt. Ezenkívül biztosítaniuk kell a két szöveg, illetve azok reprezentációjának összehasonlíthatóságát.

Teljes szövegű információkeresés

Az adatbázispiacon a teljes szövegre épülő keresőrendszerek a legelterjedtebbek és a legkedveltebbek. A digitális technológia olcsó szövegtárolási lehetőségeket kínál, és egyben igen gyors keresést is a tárolt teljes szövegekben. A felhasználó számára kényelmes, hogy nagy dokumentumtárakban kereshet mindössze egy-egy szó megadásával. Az eljárás azért is olcsó, mert nem igényel emberi indexelő munkát. A teljes szövegű keresőrendszerek közelebb állnak a tényleges felhasználói igényekhez, amelyek gyakran nem úgy jelentkeznek, ahogy azt az indexelő gondolta. A használók jobban kedvelik, ha maguk állíthatják össze a természetes nyelvű keresőprofilt, és azt nem kötik az indexelési elvek, illetve szabályok. A teljes szövegre épülő keresőrendszerek a teljesség tekintetében sokkal jobb eredményeket mutatnak, mint a szabályozott szótárakra építő indexelő szolgáltatások.

A másik oldalon azonban, a pontosságot illetően nem jók az eredmények. A teljes szövegű keresőrendszerek nagyon deficitesek, sok fölösleges találatot is adnak, és nem kínálnak semmiféle megoldást a minőségi válogatásra. A felhasználónak tehát nagy mennyiségű szövegből kell válogatnia, és mivel ideje általában nincs, ezért sajnos egyre inkább az az eljárás, hogy az első 10-20 találatnál megáll. A minőség és a relevancia helyett a sorrend lett a meghatározó, és ez egyáltalán nem kívánatos tendencia.

Az elmúlt évtizedben a teljes szövegű információkeresésre irányuló kutatások felerősödtek, különösen mióta az Egyesült Államok Nemzeti Szabványügyi és Technológiai Intézete elindította a *TREC* programot (*T*ext *RE*trieval *C*onference), amely a szövegfeltárást és -keresést támogatja *(http://trec.nist.gov)*. Éves konferenciáin fórumot ad a legfrissebb kutatási eredmények bemutatására. A TREC mára szinte fogalommá vált. A konferenciák igazolják, hogy sokkal kifinomultabb szövegfeltáró rendszerekre van igény.

A TREC kutatási program tematikus szekciókban zajlik. A kutatásokban központi szerepe van az értékelésnek, ami nagyban segíti, hogy valóban használható, a felhasználók számára is hasznos fejlesztések történjenek, és ne csak presztízskutatások. Mindig van egy fő kutatási irány, és emellett számtalan kisebb témacsoport is. Jelenleg a főirány azt vizsgálja, hogyan lehet újfajta kérdésfeltevéssel keresni statikus dokumentumgyűjteményekben, vagyis amikor a gyűjtemény ismert, a várható kérdések azonban nem. A témacsoportokban olyan kutatási témák szerepelnek, mint a többnyelvű információkeresés (cross-language retrieval), 100 Gbájtnál nagyobb dokumentumgyűjteményekben való keresés, interaktív információkeresés, hangzó, video- és multimédia dokumentumokban történő tematikus keresés.

Relevancia-visszacsatolás

Az információkeresés fontos, de nehéz problématerülete, hogy hogyan fogalmazza meg úgy a keresőkérdést, hogy az csak a releváns kognitívumokat hozza ki találatként. Ideális kérdésfeltevés csak akkor képzelhető el, ha pontosan ismerjük a dokumentumgyűjtemény összetételét, ezért a keresést ismétlődő lépésekben, mintegy fokozatosan puhatolózva kell végrehajtani. Minden egyes keresés után értékelni kell a kapott találatok pontosságát és teljességét, és az értékelés alapján kell a kérdést továbbfejleszteni. Ez a keresési módszer tehát a *relevancia* értékelésén alapul.

A relevanciára épülő információkeresés mögött az az elv áll, hogy az egyazon kérdésre megfelelő választ adó dokumentumok hasonlítanak egymásra. Ha találunk egy releváns dokumentumot, akkor a keresőkérdést ehhez kell közelíteni, így remélhetően további releváns tételekre lelünk. Vagyis a kérdést a találatok segítségével lehet finomítani. *G. Salton* két alapmódszert ajánl ehhez [1]:

- A releváns találathoz tartozó tárgyszavak, deszkriptorok beépítése a keresőkérdésbe.
- Az eredeti kérdés keresőelemei súlyának megváltoztatása a releváns tétel alapján.

A kísérletek azt igazolják, hogy érdemes a keresőprofilt mindaddig finomítani, míg a felhasználó maximálisan nem elégedett a találatokkal. A módszer az interaktív információkeresésben és a találatok szűrésében egyaránt használható.

Szövegelemzést és információkeresést támogató kutatások

Az alábbiakban néhány konkrét kutatási projekt bemutatásával érzékeltetem, mennyire sokszínű ez a tudományterület, és milyen irányok jellemzőek a legújabb kérdésfeltevésekben. A kutatások rendkívül szerteágazóak, a teljes spektrumból lehetetlen válogatni. A példák valóban csak példák, nem jelentenek minőségi preferenciát.

Fókuszált információkeresés [2]

A hirerarchikusan szervezett webes dokumentumok körében a hatékony kereséshez a tartalom és a struktúra (a hiperlinkek rendszere) viszonyát is kutatni kell. Minél jobban ismerjük a dokumentumok természetét, annál könnyebben tudjuk megtalálni az optimális szövegeket, vagyis azokat, amelyek releváns információt tartalmaznak, és amelyek segítségével, a bennük található kapcsolatok (linkek) mentén haladva a felhasználó további releváns szövegekhez is eljuthat. Ezt a keresési típust nevezik fejlesztői *fókuszált keresés*nek.

A felvetett probléma a hiperszövegek természetéhez kötődik, ahhoz a jelenséghez, hogy két szöveg közötti utalásos kapcsolat maga is tartalmi információ. Tételezzük fel például, hogy egy adott kérdésre *A* és *B* szöveg egyaránt találatot jelent, *A* szövegben pedig van egy link *B* szöveghez. A hagyományos keresőrendszerekben ez az információ (hogy a két szöveg utal egymásra) nem derül ki, csak akkor, amikor a szöveget kezdi el olvasni valaki. A rangsorolást alkalmazó keresőrendszereknél az is megeshet, hogy a rangsorban a két kapcsolódó dokumentum távol kerül egymástól.

A hagyományos tartalmi alapú információkeresés és a hipertext szolgáltatásait kihasználó böngésző keresés csak együtt alkalmazva jelenthetnek hatékony módszert a nagy mennyiségű elektronikus szövegtengerben. A fókuszált keresés találatként adja azt a dokumentumot, amelynek valamennyi „gyermeke" (amelyekre utal) szintén tartalmaz releváns információkat, de csak a gyermekeket hozza ki akkor, ha csak ezekben van releváns válasz.

A fókuszált keresés a *Dempster-Shafer* bizonyítási elméleten alapszik. Egy dokumentum tartalmi reprezentációja az alapszöveg és a hozzá kapcsolódó „gyermek" dokumentumok halmazaként van definiálva a Dempster-féle kombinációs szabály segítségével. A fókuszált keresési modellt az alábbi elemek határozzák meg: a webtér logikai struktúrája, a dokumentumok reprezentációi, a tartalmi és a strukturális tudást figyelembe vevő reprezentációk halmaza, a keresési funkció és fókuszált keresés.

A modell hierarchikusan szerkesztett szövegeket tud kezelni, amelyek fastruktúrában ábrázolhatók, és ahol a „szülő" dokumentum általánosabb szinten tárgyalja a témát, mint a hozzá kapcsolódó „gyermek" dokumentumok. A fókuszált keresés egy adott keresőkérdés esetében azt jelenti, hogy a talált dokumentum is és a nála alacsonyabb hierarchiaszinten elhelyezkedő kapcsolódó dokumentumok is relevánsak. A halmazba tartozó dokumentumokat indexelő kifejezések, kulcsszavak, tárgyszavak csoportja reprezentálja. A reprezentációban tükröződnie kell a kifejezések súlyának, vagyis annak, mennyire erősen jellemzik a szöveg tartalmát. A fölérendelt szövegek reprezentációja tartalmazza az alárendelt szövegek reprezentációit is. Az automatizálás számára persze mindezek a lépések bonyolult matematikai képletekkel modellálhatók.

A fókuszált keresési modellt az Ermitázs múzeum hierarchikusan szervezett weboldalán tesztelték 15 különféle témára irányuló keresőkérdéssel. A teszteléshez használt dokumentumgyűjtemény nem túl nagy, a kérdések is válogatottak voltak. A módszer ezek között a körülmények között hatékonynak bizonyult, és mindenképpen figyelemre méltó ötleteket adhat a tartalmi reprezentációk és a hipertext struktúrára együttesen építő információkeresés további kutatásához.

Bekezdés szintű információkeresés [3]

Az információs igények gyakran nem teljes dokumentumokra, csupán ezeken belüli releváns szövegrészekre irányulnak. A felhasználó szempontjából a kognitívum nem mindig egyezik meg a feltárási egységgel. A kutatások speciális köre irányul arra, hogyan lehet a keresőt rögtön a releváns szövegrészekhez vezetni, anélkül, hogy ehhez hosszabb szövegeket kelljen végigolvasni és elemezni.

Az Ausztráliában kifejlesztett *Taylor* nevű keresőprogram a keresőkérdésre egy virtuális dokumentumot ad válaszként, amely a dokumentumok releváns *bekezdés*eit, illetve ezekre mutató linkeket tartalmaz. Az eljárás két nagy lépésből áll: először ki kell válogatni azokat a bekezdéseket, amelyek relevánsak lehetnek a kérdésre, majd ezekből össze kell állítani a válaszként megjelenő virtuális szöveget. A bekezdések közötti sorrend nem kötött, de relevanciaértékük szerint lehet őket rangsorba állítani.

A Taylor hatékony működéséhez pontosan megfogalmazott, lehetőleg specifikus keresőkérdések szükségesek. Az is fontos, hogy a rendszer fel tudja térképezni a dokumentumok szövegszerkezetét, ehhez speciális elemzőrendszert is kifejlesztettek. A fejezetcímek sokat segíthetnek, különö-

sen ha összehasonlíthatók a keresőkérdéssel. A Taylor először is elemzi az adott dokumentumgyűjteményt, és felépít egy indexfájlt a dokumentumok szerkezetéről és tartalmáról. A bejövő keresőkérdéseket ezzel a fájllal hasonlítja össze, és az összehasonlítás eredménye a megfelelő bekezdések rangsorolt listája.

TREVI (Text Retrieval and Enrichment for Vital Information) [4]

A *TREVI* projekt egy megosztott objetumorientált Java alapú rendszer, amely a statikus/dinamikus specifikációk szisztematikus feldolgozásán és a nyelvi műveletek ellenőrzésén alapul. A TREVI-t a tematikus szövegelemző rendszerek közé kell sorolni, amely az alábbi szolgáltatásokat nyújtja:

- Természetes nyelvű szövegek elemzése különböző nyelvészeti modulok együttes alkalmazásával.
- Tartalom szerinti kategorizálás.
- A szövegek kiegészítése hasonló forrásokra mutató linkekkel.
- Szövegek publikálása a weben, megfelelő böngésző eszközök támogatásával.

A TREVI konzorcium kifejlesztette azokat az integrált szoftvereket, amelyekkel szűrni és osztályozni lehet a bejövő adatokat a használói igények függvényében, és ugyanakkor további kapcsolódó háttér-információkkal is ki tudják egészíteni őket. A szoftvercsomagot hírek elemzésére használják. Az eszközkészlet a következő részekből áll:

- A bejövő szövegeket kezelő, az adatokat standardizáló modul.
- Nyelvi feldolgozó modul. Az elemző támogatja nagy tömegű szöveg elemzését, a fogalmak szemantikai meghatározását, személynevek, eseménynevek felismerését stb.
- Független lexikon- és tezauruszkezelő modul, amellyel főként angol és spanyol terminológia kezelhető.
- Felhasználói profilokat kezelő modul.
- Szövegek kategorizálását végző modul, amely a felhasználói profilokhoz igazodva osztályozza a szövegeket.
- A szövegek linkekkel történő kiegészítését végző modul, amely a tartalom alapján összekapcsolja a szövegeket már feldolgozott hasonló témájú szövegekkel vagy adatokkal.
- Publikációs modul, amely a feldolgozott és kiegészített szövegeket hozzáférhetővé teszi a weben.
- Speciális, az egész folyamatot vezérlő modul.

A TREVI szoftvercsomag újdonsága, hogy kombinálni tudja a szisztematikus megközelítést a fejlett és adaptív nyelvi elemzéssel, illetve a szövegek tartalmi alapú kategorizálásával. A program mind az osztályozás pontossága, mind a használói vélemények szerint jó eredményekkel kecsegtet.

Televízió- és rádióműsorok tartalmi alapú keresése [5]

Az AT&T cambridge-i laboratóriuma DART (Digital Asset Retrieval Technology) projektjének célja, hogy lehetővé tegye a digitális média – amely szöveget, hiperszövegeket, képeket, audio- és videoanyagokat egyaránt tartalmaz – indexelését, annotálását és visszakeresését. Egy különleges szövegtípusról van tehát szó, amely azonban mindinkább kihívást jelent az információs rendszerek számára. Egyszerre kell megoldani az írott, a hangzó és a videoszöveg feldolgozását.

Az angol televíziócsatornák műsorait a normál sugározható jelek mellett teletext formában is tárolják. Ez tartalmazza a program vázlatát, címét, időpontját és egyéb információkat. A *rádió- és televízióprogramok* tartalmát strukturált, hierarchikus rendszerben ábrázolják. A hierarchia csúcsán a program neve található, kiegészítve metaadatokkal és a műsoridő hosszával. A programokat szegmentálják, vagyis kisebb részekre darabolják, ezek jelentik a feltárás és a keresés egységeit, vagyis a kognitívumokat. Egy ilyen kognitívum önálló témával rendelkezik (pl. egy hír vagy riport). A szegmensek közötti határt akusztikai jelek vagy videoszünetjelek jelölik. Léteznek olyan algoritmusok, amelyek fel tudják ismerni a beszélő személyének megváltozását, a mikrofonváltást, vagy a zene kezdetét, illetve végét. A videorészleteknél is meg tudják állapítani, hol vannak vágások, illetve hol változik a kamera mozgása. A televízió-műsorok szegmentálásánál általában az audio- és a videoegységek együttes figyelembevételével dolgoznak; ahol a váltások egymáshoz közel vannak, ott nagy valószínűséggel témaváltás is van. A rádióprogramoknál természetesen csak az audioeszközök használhatók.

Az audio/video eszközökkel történő szegmentálást megerősítik egy nyelvi elemzéssel is, amely ellenőrzi, hogy a kijelölt egységek lexikai tartalma homogén-e, vagyis a benne szereplő szavak egy témára utalnak-e. Az így kvantált műsorok visszakereséséről egy többféle keresőeszközt is alkalmazó rendszer gondoskodik, amely az alábbi keresési típusokat kínálja fel:

- Képrészletek, imidzsek keresése keretek segítségével. A szegmentálás során meghatározzák azokat a kereteket, amelyek az egyes jelenetek határait jelentik. Ezek a keretek ahhoz is segítséget nyújtanak, hogy az ismétlődő jeleneteket könnyebben lehessen felismerni. A képek indexeléséhez a hisztogram technológiát használják. Az imidzsek alapján történő keresés azonban sokkal lassabb és nehézkesebb, mint az egyes jelenetek szöveges leírásában történő hagyományos keresés.
- Az akusztikai keresések a hasonlóságon alapulnak. Az ilyen kereséseknek főként akkor van hasznuk, ha például egy bizonyos beszélőt keresünk.
- Kombinált akusztikai és kulcsszavas keresés. A kulcsszavas keresések további szűrésére használható az akusztikai hasonlóságon alapuló rangsor. A vizsgálatok nem igazolták ennek a keresési módszernek a hatékonyságnövelő hatását.
- A lexikai ellenőrzés során minden szegmenst a lexikai egységek halmazával, illetve az ezt ábrázoló vektorral jellemeznek. A vektorok összehasonlításával mérhető az egymás melletti szegmensek tartalmi hasonlósága. Meghatározott küszöbérték felett ezeket a szegmenseket egy egységgé vonják össze.
- A televízió-műsoroknál gyakran előfordul, hogy a riportok mellett feliratokkal is tudatják a nézővel a beszélő kilétét vagy a témát. Az elemző rendszer ezeket a feliratokat is felhasználja a tartalom reprezentálásához.

A felhasználó először egy útmutató segítségével tájékozódhat a televízió-műsorokról, amely megadja a programokra vonatkozó alapvető információkat (cím, rövid leírás stb.). A kiválasztott programokon belül lehetőség van a szegmensek közötti böngészésre. A képernyőn fel vannak sorolva az egyes szegmenseket jellemző képkockák és a hozzájuk tartozó audiorészletek, ezek és egy szöveges keresőablak segítségével lehet keresni. Így aztán ha valaki egy hosszabb magazinműsorból csak egy bizonyos témával foglalkozó részre kíváncsi, a rendszer segítségével megkeresheti, és azonnal meg is nézheti. Mindezek felett a rögzített műsorokat egy egyszerű osztályozási rendszerbe is besorolják, amely újabb könnyítést ad a válogatáshoz (pl. beszélgető műsorok, hírek, filmek).

Ez a keresőrendszer tehát tulajdonképpen egy hagyományos szöveges kereső, amelyet kiegészítettek video- és audioeszközökkel. A felhasználók szövegesen keresnek, kulcsszavak alapján. A háttérben segédprogramok működnek, amelyek felajánlják a keresőnek az általa megadott kulcsszavakkal jelölt fogalmakhoz kapcsolódó további kulcsszavakat, így próbálván megoldani a szabályozatlanság problémáját.

Az információkeresés új generációja [6]

A 21. század információkereső rendszereitől elvárjuk, hogy legyenek képesek konkrét kérdésekre konkrét válaszokat adni, javaslatot tenni, a választ adott esetben önálló szövegben megfogalmazni, vagyis újfajta kérdésfeltevésekhez is alkalmazkodni. Az is elvárás, hogy az információ azonnal érthető és használható módon jelenjen meg a kérdező számára. A jelenleg működő keresőrendszerek által szolgáltatott rangsorolt találati listák nem felelnek meg ennek a követelménynek. Lehet, hogy a válasz érthető (bár gyakran elég könnyen félreérthető is), de ritkán hasznosítható. Az ideális információs szolgáltatás képes arra, hogy a felhasználó által szövegesen megfogalmazott kérdésre egy célzottan összeállított szöveges választ adjon.

A General Electrics kutatócsoportja egy ilyen fejlesztésen dolgozik, az *információkeresés új generációján (Next Generation Information Retrieval = NGIR)*. A kutatás kiindulópontja, hogy a keresés eredményessége, vagyis a teljesség és a pontosság összefüggésben van a keresőkérdés hosszával, illetve kidolgozottságával. Minél jobban, bővebben van megfogalmazva a kérdés, annál könnyebben hajtható végre eredményes keresés. A felhasználók által megfogalmazott kérdések azonban többnyire szűkszavúak. Ezért az információkeresés hatékonyabbá tétele érdekében a keresőkérdések megfogalmazásánál is alkalmazni kell a nyelvi feldolgozó technológiákat.

A módszert kiterjesztett tematikus keresésnek nevezik, lényege, hogy a felhasználói kérdéseket kiegészítik néhány dokumentum releváns bekezdéseivel, szövegrészleteivel. Ezáltal a téma többféle megvilágításban, megfelelőbb kontextusban fogalmazható meg a kérdés számára. A konkrét keresést már ezzel a kibővített keresőkérdéssel végzik el. A módszer sokkal jobb eredményeket mutat, mint a hagyományos statisztikai alapú keresések, ezért ígéretesnek tűnik egy új generációs információkereső rendszer megalapozásához.

A kiterjesztett keresőkérdés tulajdonképpen egy metadokumentum, amely minden olyan információs elemet tartalmaz, amelyre a felhasználó kíváncsi. Ez a metadokumentum azután folyamatosan

alakítható, változtatható további releváns szövegek részleteivel, és végezetül előáll egy olyan szöveg, amely maga a válasz a kérdésre.

A folyamatot próbálják teljesen automatizálni. Az egyszerűbb nyelvi feldolgozó technikák alkalmazása – emberi beavatkozás nélkül – nem adott sokkal jobb eredményeket, de a fejlettebb technológiák reménnyel kecsegtetnek. Az egyik kedvelt módszer a relevancia-visszacsatolás, amikor a felhasználó értékeli az első találatok relevanciáját, és a kérdést ennek az érékelésnek megfelelően finomítják, módosítják. A relevancia-visszacsatolás módszerével könnyen eljutunk az ismert releváns dokumentumokhoz, újakat viszont nehezebb találni. A jobb kérdések megfogalmazásához tehát más módszerek is szükségesek.

A relevancia-visszacsatolás során általában újabb fogalmakkal egészítik ki a kiinduló kérdést. Az új módszerek nemcsak fogalmakat, hanem mondatokat, illetve egész szövegrészeket is beépítenek ebbe a folyamatba, remélvén, hogy az így szövegesen is kiegészített keresőkérdés hatékonyabb. Az eredeti kérdésre kapott találatok közül a relevancia szerinti rangsor első 10-30 dokumentumát használják a kiegészítéshez. Ezekben megkeresik azokat a szövegrészeket, amelyekben előfordulnak az eredeti kérdésben szereplő fogalmak, és ezeket a szakaszokat építik be az újabb keresőkérdésbe.

A módszer problémája, hogy a relevancia megítéléséhez a felhasználónak sok szöveget kell elolvasnia, ez pedig időigényes, és rontja a hatékonyságot. Ezért a módszert tovább finomították, beépítettek egy előzetes automatikus szövegtömörítési fázist. A relevanciát ezután nem teljes, hanem tömörített szövegek alapján kell megítélni, és ezekből lehet a kiegészítéshez szükséges részeket átemelni a keresőkérdésbe.

A kutatási projektnek része tehát egy automatikus szövegelemző és -tömörítő modul is, amely a *DoX* névre hallgat. Ez a modul önmagában is érdekes és hasznosítható tapasztalatokat nyújt. A DoX program kétféle tömörítést végez. A *tematikus tömörítés* csak arra a témára koncentrál, amelyet a felhasználó a keresőkérdésben megfogalmazott. Ha egy szöveg nem szól a témáról, akkor nem készül róla tömörítés. Az *általános tömörítés* a szöveg főtémáját keresi és fogalmazza meg, függetlenül attól, hogy mi ez a téma. A kétféle megközelítés szerint ugyanarról a szövegről kétféle tömörítés is készíthető. A DoX program indikatív és informatív referátumot egyaránt tud készíteni. Az indikatív referátum az eredeti szöveget kb. 5-10%-ára tömöríti, ez éppen arra elég, hogy a leglényegesebb tartalmi elemekre utaljunk. Az informatív referátum az eredeti szöveg 20-30%-a, az eredeti minden fontos állítását tartalmazza. Az automatikus tömörítési folyamat a következő lépésekből épül fel:

- A szöveget először szakaszokra bontják. Ez történhet a bekezdések mentén, a szöveg tipográfiai elrendezése (behúzások, SGML tagek, üres sorok stb.) nyújt segítséget. Ha a szöveg nem oszlik bekezdésekre, akkor többé-kevésbé egyforma részekre osztja a program.
- Második lépésben a program kiválogatja a legjellemzőbb bekezdéseket, illetve szövegszakaszokat a kulcsszavak, szövegszavak, illetve a felhasználó által megadott szempontok szerint.
- Ezután fel kell térképezni az egymás melletti bekezdések kapcsolatát. Ha egy kiválasztott bekezdés egyértelmű előre- vagy hátrautalással kapcsolódik a mellette állóhoz, akkor ez utóbbi is a kiválasztottak közé kerül.
- A következő lépés a bekezdések súlyozása. Minden szakasz pontértéke attól függ, hogy a keresőkérdés hány elemét tartalmazza.
- A bekezdések súlyát a bekezdés hosszához viszonyítva normalizálják, figyelembe véve a kitűzött célt, hogy milyen hosszúságú tömörítést akarunk. Ezt a célt minél jobban meg kell közelíteni.
- Azokat a bekezdéseket, amelyeknek hossza több mint másfélszerese a megengedettnek, kiiktatják. Ezáltal csökken a tömörítésnél figyelembe veendő szakaszok száma, így nő a hatékonyság. Ha van olyan veszély, hogy minden bekezdés hosszabb a megengedettnél, akkor be lehet állítani úgy a program működését, hogy az első bekezdést mindenképpen tartsa meg.
- Ezután a megmaradt bekezdéseket tartalmuk, szerkezetük és hosszuk alapján kettesével-hármasával csoportosítják. Bármely bekezdések kerülhetnek egy csoportba, nem kell, hogy egymás mellett legyenek a szövegben. Az eredeti egymásra utaló kapcsolatokat azonban figyelembe veszi a rendszer.
- Az újonnan keletkezett csoportokat újból súlyozzák, és a másfélszeresnél hosszabbak ismét kiesnek.
- A megmaradó csoportokat súlyuk alapján rangsorba állítják. A rangsor élén álló bekezdésekből a kitűzött célnak megfelelően készül el a tömörítés.

A tömörítés célja, hogy a felhasználó el tudja dönteni a szöveg relevanciáját, és ki tudja választani azokat a szövegrészeket, amelyekkel a keresőkérdés kiegészíthető. A keresés tehát a következő:

- A természetes nyelven megfogalmazott kérdést a rendszer lefordítja a keresőnyelvre, és lefuttatja az adatbázisban.
- Eredményül egy maximum 30 referátumból álló listát ad vissza, amelyek a fenti tömörítési módszerrel készültek.
- A felhasználó átnézi a referátumokat (egynek az átnézése kb. 5-15 másodpercet igényel), és kiválogatja azokat, amelyek nem relevánsak.
- A relevánsnak ítélt tömörítések bekerülnek a keresőkérdésbe.
- Az így kiegészített témát alávetik a szokásos természetes nyelvi indexelési eljárásoknak, majd az így kialakuló keresőkérdéssel elvégzik a végső keresést.

Ez a programcsomag tulajdonképpen ötvözi mindazokat az eljárásokat, amelyeket a természetes nyelvek feldolgozásából az információkeresés fejlett technikái hasznosítani tudnak. A kutatás természetesen nem annyira kiérlelt még, hogy nagy tömegű szövegen, különféle kérdéstípusokkal tesztelték volna. A gondolatmenet azonban ígéretesen illusztrálja, hogyan hasznosíthatók az automatizálási lehetőségek a kereséssel egybekötött szövegelemzésben.

Irodalom

[1] SALTON, Gerard: Automatic text processing. The tranformation, analysis, and retrieval of information by computer. Reading, MA., Addison-Wesley, 1989. p. 307.

[2] LALMAS, Mounia–MOUTOGIANNI, Ekaterini: A Demster-Shafer indexing for the focussed retrieval of a hierarchically structured docuent space. Implementation and experiments on a web museum collection. = RIAO'2000: Content-based multimedia information access. Conference proceedings. Paris, College de France, 2000. p. 442–456.

[3] PARADIS, Francois: Information extraction and gathering for search engines. The Taylor approach. = RIAO'2000: Content-based multimedia information access. Conference proceedings. Paris, College de France, 2000. p. 78–85.

[4] BASILI, Roberto–PAZIENZA, M. T.: An adaptive and distributed framework for advanced IR. = RIAO'2000: Content-based multimedia information access. Conference proceedings. Paris, College de France, 2000. p. 908–922.

[5] MILLS, Timothy J. (et al.): AT&TV: Broadcast television and radio retrieval. = RIAO'2000: Content-based multimedia information access. Conference proceedings. Paris, College de France, 2000. p. 1135–1144.

[6] STRZALKOWSKI, Tomek (et al.): Towards the Next Generation Information Retrieval. = RIAO'2000: Content-based multimedia information access. Conference proceedings. Paris, College de France, 2000. p. 1196–1207.

[7] KUPIEC, J.–PEDERSEN, J.–CHEN, F.: A trainable document summarizer. = Proceedings of the Eighteenth SIGIR Conference. New York, ACM, 1995. p. 68–73.

[8] LUHN, H. P.: The automatic creation of literature abstracts. = IBM Journal of Research and Development, 2. sz. 1958. p. 159–165.

[9] MOENS, Marie-Francine: Automatic indexing and abstracting of document texts. Boston, Kluwer, 2000.

[10] PRÓSZÉKY Gábor: Számítógépes nyelvészet. Bp., Számítástechnika-alkalmazási Vállalat, 1989.

[11] RUGE, Gerda–SCHWARZ, Cristoph–WARNER, Amy J.: Effectiveness and efficiency in natural language processing for large amounts of text. = JASIS, 1991. július, p. 450–456.

Beérkezett: 2005. I. 19-én.

Varga Katalin
az Országos Pedagógiai Könyvtár és Múzeum könyvtárának vezetője, főosztályvezető.
A Pécsi Tudományegyetem Könyvtártudományi Tanszékének egyetemi adjunktusa.
E-mail: *kvarga@hu.inter.net*