MUNK SÁNDOR ny. ezredes

INFORMÁCIÓKERESŐ RENDSZEREK ALAPJAI, ÖSSZETEVŐI

BASICS AND COMPONENTS OF INFORMATION SEARCH/RETRIEVAL SYSTEMS

Korunk robbanásszerűen bővülő információtömege már új kezelési módszereket igényel. Az információkeresés alapvető eszközét napjainkban az informatikai eszközökkel támogatott információkereső rendszerek képezik. Az információkeresés eredményességének növeléséhez az információkereső rendszereknek egyre inkább jelentés-alapú megoldásokat kell alkalmazniuk. A szemantikus keresés módszerei a hagyományos információkereső rendszerek keretei között jelennek meg, így ezek vizsgálata alapját képezi a szemantikus keresés kutatásának is. Ennek érdekében jelen publikáció az információkereső rendszerek átfogó bemutatását tűzte ki céljául. Ezen belül: összegzi az információkereső rendszerek alapjait, általános jellemzőit; meghatározza az információkereső rendszerek egyes összetevőit és ezek rendeltetését, rendszerezi alkalmazott megoldásaikat. Kulcsszavak: információkeresés, információkereső rendszerek, információkereső rendszerek összetevői.

The information explosion of our days requires new management methods. Nowadays the essential tools of information retrieval are IT assisted information retrieval systems. To increase the effectiveness of information retrieval, information retrieval systems have to apply meaning-based solutions. Semantic search methods appear in the framework of traditional information retrieval systems, so their examination forms the basis of semantic search research too. For this reason this publication aims to present a general view of information retrieval systems. In particular: summarizes the basics, and general characteristics of these systems; specifies the components of information retrieval systems, and their purpose, reviews their solutions which are applied. Keywords: information search/retrieval, information search/retrieval systems, components of information search/retrieval systems

Bevezetés

Egy — az információkeresés alapjaival foglalkozó — korábbi publikációban [1] már hivatkoztunk az információk napjainkban szinte mindenki által magasra értékelt szerepére, jelentőségére. Az úgynevezett in-

formációs társadalomban, az információ- vagy tudásgazdaságban az információ önálló értékké, a társadalmi, gazdasági, kulturális és hétköznapi tevékenységek egyre jelentősebb szerepű tárgyává és erőforrásává vált. A robbanásszerűen bővülő mennyiségű információtömeg új kezelési módszereket igényel. A szervezeti folyamatok, vagy a személyes célok megvalósításához szükséges információk megszerzésének egyik, növekvő jelentőségű lehetősége a már valahol meglévő, rögzített információk közötti keresés.

Az információkeresés, vagy más kifejezéssel információ visszakeresés értelmezésünk szerint olyan tevékenység, amely információreprezentációk (adatok) meghatározott köréből meghatározott információigény kielégítését segítő információreprezentáció(k), adat(ok) megtalálására, kiválasztására irányul. Ez a meghatározás független a keresés tárgyát képező adatok strukturáltságától, így magában foglalja a táblázatos formában rendezett (relációs) adatok, a szöveges, álló és mozgókép-, hang-dokumentumok, valamint az XML (vagy más formatizált, félig strukturált) dokumentumok közötti keresést. [1, 253. o.]

A rendelkezésre álló információk napjainkban már szinte minden témakörben túlnyomórészt informatikai[1] rendszerekben kerülnek tárolásra, sőt egyre növekvő mértékben azokban is keletkeznek, így az információkeresés is lényegében csak informatikai eszközökkel, módszerekkel támogatva valósítható meg. Emellett a hagyományos formában (könyvek, sajtótermékek, iratok, stb.) tárolt információk keresését is informatikai megoldások (pld. elektronikus katalógusok[2]) támogatják. Az informatikai eszközökkel támogatott információkeresést különböző információkereső rendszerek biztosítják, amelyek nélkül a folyamatosan keletkező, illetve a tervszerűen összegyűjtött információk rejtve maradnak, nem jutnak el lehetséges felhasználási helyeikre.

A szervezetek nem jutnak hozzá a működésükhöz szükséges külső információkhoz, rendelkezésre álló információik nem hasznosulnak szervezeti szinten. A személyek nem tudják megszerezni a számukra fontos információkat. Az információkeresés jelentőségét, hétköznapivá

[1] Jelen publikációban az 'informatikai' jelzőt tág értelemben, 'információs tevékenységeket támogató, megvalósító technikai [megoldás]' tartalmú értelmezésben használjuk.

[2] OPAC = Online Public Access Catalog.

válását szemlélteti, hogy a Google-lal történő keresés, a guglizás mára már közszóvá vált.[3]

A különböző formátumú információk (táblázatos adatok, szöveges dokumentumok, multimédia anyagok, stb.) között kereső rendszerek az évek során újabb és újabb megoldásokat, módszereket alakítottak ki, hogy növeljék a keresés eredményességét és hatékonyságát. A keresés minőségének alapfogalma a relevancia, a keresési igénynek történő megfelelés. Erre épül a teljesség (a releváns információk mekkora része kerül megtalálásra) és a pontosság (a megtalált információk mekkora része releváns).[4] Az információkeresés alapvető nehézsége abban áll, hogy a kereső rendszernek a relevancia megítéléséhez a reprezentáció alapján elő kell állítania, meg kell közelítenie, vagy 'el kell találnia' a hordozott jelentést, információt.

Az információkereső rendszerek esetében — elsőként az internetes kereséshez és a 2000-es évek elején megszületető szemantikus web kezdeményezéshez kapcsolódóan — merült fel az információk tartalmi alapú keresésének erősítése és jelent meg a szemantikus keresés. A jelentés fogalmához (szemantika = jelentéstan) kapcsolódó kifejezés tartalma, értelmezése megjelenését követően folyamatosan változott, bővült és pontosodott, napjainkra az információkereső rendszerek talán legfontosabb elméleti és gyakorlati kérdésévé vált.

A szemantikus keresés a védelmi, közigazgatási szférában is egyre növekvő jelentőségű szerepet játszik. Csak két nagy területet említve megjelenik a nyílt forrású hírszerzés katonai, nemzetbiztonsági és rendőri célú felhasználásában (a nemzeti biztonságot fenyegető terrorizmus, a szervezett bűnözés, a szélsőséges csoportok, és a korrupció elleni harcban), vagy a jogszabályokban, bírói ítéletekben, közigazgatási döntésekben, eljárási iratokban történő keresés során.

Mivel a szemantikus keresési módszerek megjelenése eddig alapvetően nem, vagy nem sokat változtatott az információkeresés alapelvein, átfogó feladatstruktúráján, a hagyományos információkereső rendszerek képezik alapját a szemantikus keresés megvalósításának. Mindezek alapján jelen publikáció célja az informatikai eszközökkel támogatott

[3] A 'to google' kifejezés 2006 nyarán bekerült az Oxford English Dictionary-be és a Merriam-Webster Collegiate Dictionary-be.

[4] Recall, precision.

információkereső rendszerek átfogó bemutatása és ezzel közvetve a szemantikus kereséssel kapcsolatos kutatások megalapozása. Ennek érdekében:

— összegzi az információkereső rendszerek alapjait, általános jellemzőit, felépítését;
— meghatározza az információkereső rendszerek egyes összetevőinek rendeltetését, rendszerezi alkalmazott megoldásaikat.

Információkereső rendszerek alapjai

Az informatikai rendszerekben tárolt információk közötti keresés technikai feltételét megfelelő funkciókat megvalósító informatikai rendszerek, alkalmazások képezik. Ezek két nagy típusát napjainkban — egy korábbi publikációban [1] megfogalmazottaknak megfelelően — az adatbázis-kezelő rendszerek és az információ visszakereső rendszerek képezik.[5] A keresés funkciói alapjukat tekintve mindkét rendszerben szinte azonosak: meghatározott formátumú felhasználói lekérdezésekre az annak megfelelő információk kiválasztása és rendelkezésre bocsátása.

Az azonosság mellett számos jelentős különbség is kimutatható, amelyek közül meghatározó a keresési tartomány jellege, amely az adatbázis-kezelő rendszerek esetében alapvetően elemi (egyedek jellemzőire, kapcsolataira vonatkozó) információk együttese, míg az információkereső rendszerek esetében ilyen elemi információk sokaságát közvetve, strukturálatlanul tartalmazó információforrások együttese. A szemantikus keresés megoldásai mindkét esetben értelmezhetőek, alkalmazhatóak, azonban szerepük, jelentőségük eltér.

Jelen publikációban részletesebben csak a strukturálatlan információk közötti kereséssel, az ún. információ visszakereső rendszerekkel és ezek közül is csak az informatikai megvalósításokkal foglalkozunk. A következőkben:

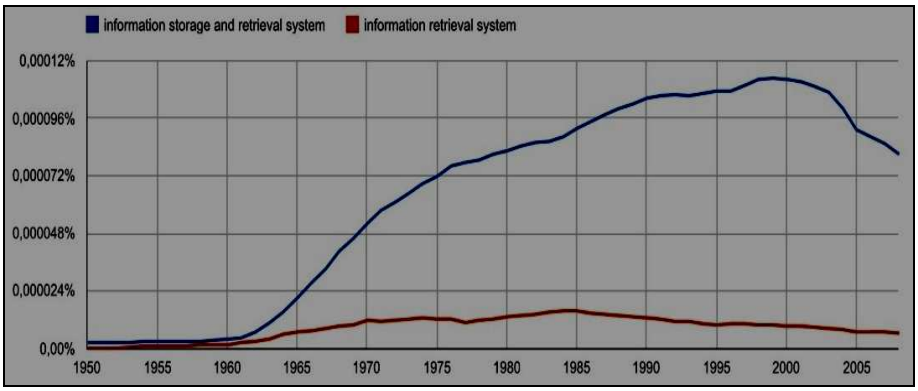
— értelmezzük az információkereső rendszerek határait, kapcsolatrendszerét;
— röviden bemutatjuk az informatikai eszközökkel támogatott információkeresés előzményeit és általános folyamatmodelljét;

[5] Database Management System (DBMS), Information Retrieval System (IRS).

— meghatározzuk az információkereső rendszerek alapvető architektúráját;
— végül áttekintjük a dokumentumok reprezentációjának alapjait.

Az információkereső rendszerek határai, kapcsolatrendszere

Elsőként rendszerszemléletű megközelítésben vizsgáljuk meg, hogyan határozható meg az információkereső rendszerek környezete és határai. A rendeltetésből következően a rendszer környezetének alapvető eleme(i) a keresési igénnyel rendelkező felhasználó(k). A következő lényeges összetevő a keresési tartományt képező információhordozók, dokumentumok összessége, amelyek elvileg tartozhatnak a környezethez, de alkothatják a rendszer részét is. Az előbbi esetben információtároló és visszakereső rendszerről, az utóbbi esetében 'tisztán' információkereső rendszerről beszélünk. Egy több millió szakkönyv tartalmát feldolgozó rendszer szerint az információtároló és visszakereső rendszer kifejezés jelentősen többször szerepel a szakirodalomban.



1. ábra: Rendszer kifejezések előfordulása 1950-2008 között[6]

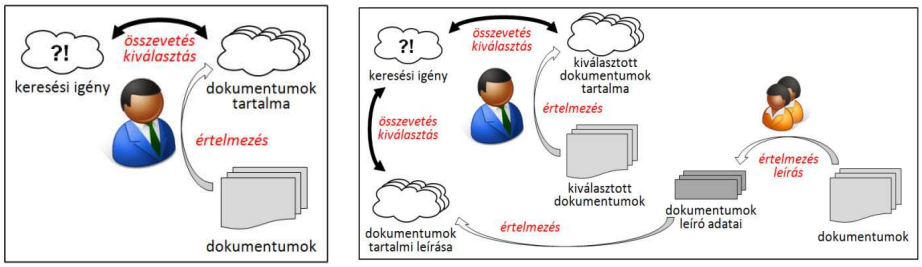
Az információkeresés tehát megvalósítható az információtárolástól lényegében független rendszer segítségével (pld. weblapok keresése), vagy integrált megoldás keretében (pld. jogszabályok keresése). Napjainkban

[6] Forrás: Google Ngram Viewer.

az információkeresés — az ábrán látható adatokkal ellentétben — jellemzően önállóan, a tárolástól függetlenül kerül megvalósításra, de legalábbis egy információtároló és visszakereső rendszer önálló alrendszerét képezi. Ennek megfelelően a továbbiakban e szűkebb értelmezésre építünk. A szűkebb értelemben vett információkereső rendszerek inputját tehát a keresési tartományt képező információhordozók, dokumentumok, valamint a felhasználói információigények képezik, outputjai pedig az információigény szempontjából hasznosnak minősített dokumentumok listái. Figyelembe véve, hogy a valós információkereső rendszerek nagyszámú információforrásban történő keresést biztosítanak, szinte minden megoldásban a rendszer önálló alrendszereit képezi a dokumentumok előzetes feldolgozása és a lekérdezések megválaszolása során az előfeldolgozás eredményeként létrehozott adatok közötti keresés.

Informatikai eszközökkel támogatott információkeresés modellje

A „hagyományos” információkeresés legegyszerűbb változata esetében a kereső személyesen, közvetlenül nézte át az elérhető információhordozókat (könyveket, iratokat, rajzokat, képeket, hangfelvételeket, filmeket, stb.) és választotta ki azokat, amelyeket információigényének megfelelőnek talált, illetve szerezte meg azokból a keresett információkat. Ez a XX. század második felében bekövetkezett információrobbanás, majd az informatika folyamatos fejlődése következtében gyakorlatilag kivitelezhetetlenné vált. A közvetett információkeresés első — hagyományos — megoldása a könyvtári dokumentumok esetében jelent meg, amely a keresett dokumentumok előzetes feldolgozására, különböző leíró adatok összegyűjtésére és meghatározására, majd ezek különböző katalógusokba rendezésére épült. A keresés ezen — az eredeti dokumentumot reprezentáló — leíró adatok felhasználásával indult, lényegében kétlépcsőssé vált. A kereső személy ebben az esetben először nem az eredeti dokumentumok, hanem azok (bibliográfiai) leírásai között keresett, majd az ezek segítségével kiválasztott dokumentumokat vizsgálta át részletesen. A keresés eredményessége, minősége így már jelentős mértékben a dokumentumok leírásának a kereső fél által nem befolyásolható minőségétől függött.



2. ábra: Hagyományos és közvetett információkeresés folyamatmodellje[7]

Az informatikai eszközökkel támogatott keresés a XX. század utolsó évtizedeiben vált egyre jelentősebbé, amikortól a tárolt, különböző célokra felhasználható információk köre folyamatosan bővülő mértékben állt rendelkezésre, sőt keletkezett informatikai eszközökkel feldolgozható formában (mára ami nem érhető el ebben a formában, az szinte 'nem is létezik').

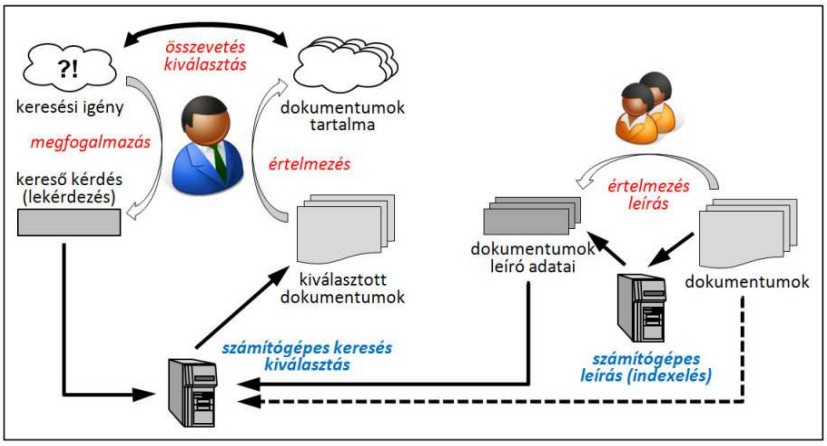
Ebben az esetben újabb feladat jelent meg: a keresési igényt meg kellett fogalmazni az alkalmazott informatikai megoldásnak megfelelő formában.

Innentől a dokumentumtartalmak között a keresési igény alapján történő keresés helyébe a dokumentumtartalmak leírása, reprezentációi között a keresési igény reprezentációja alapján történő keresés lépett.

Az informatikai megoldás alapvető sajátossága volt, hogy az információkat hordozó dokumentumok, reprezentációk korábban emberi kezelése (értelmezése, leírása, illetve szelektálása) helyébe részben, vagy teljesen informatikai eszközök által végzett műveletek léptek.

A feladat szépsége és nehézsége mindmáig abban áll (és ez vezetett el aztán szükségszerűen a szemantikus keresés témaköréhez), hogy a reprezentációk (adatok) szintjén működő 'technikai' megoldások hogyan tudják 'kezelni' a reprezentációk által hordozott információkat, jelentést. A korszerű információkeresés általános folyamatmodelljét a következő ábra tartalmazza.

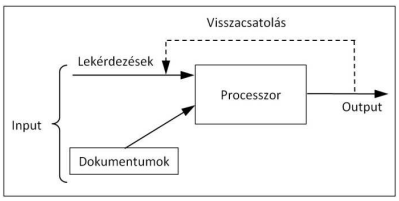
[7] Valamennyi forrás megjelölés nélküli ábrát a szerző maga készítette.

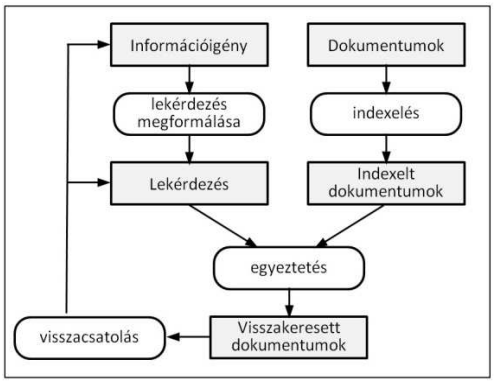


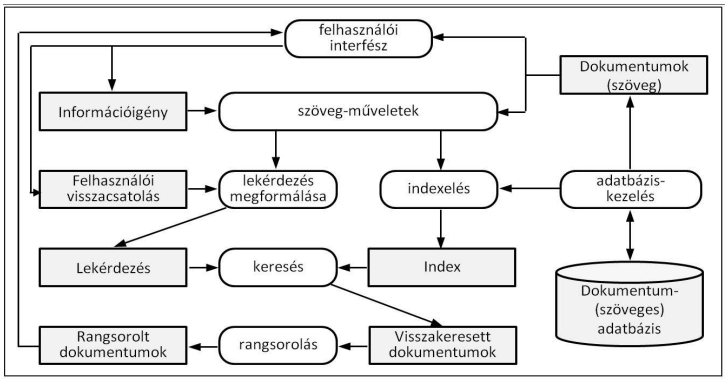
3. ábra:
Informatikai eszközökkel támogatott információkeresés folyamatmodellje

Információkereső rendszerek felépítése, összetevői

Az információkereső rendszerek az egyik első átfogó munkában [2] fekete dobozként jelentek meg. Ezt követően általános rendszermodelljük a szakirodalomban (pld. [3], [4]) különböző részletezettséggel és többnyire szöveges információ visszakeresésre specializálva jelenik meg került bemutatásra.

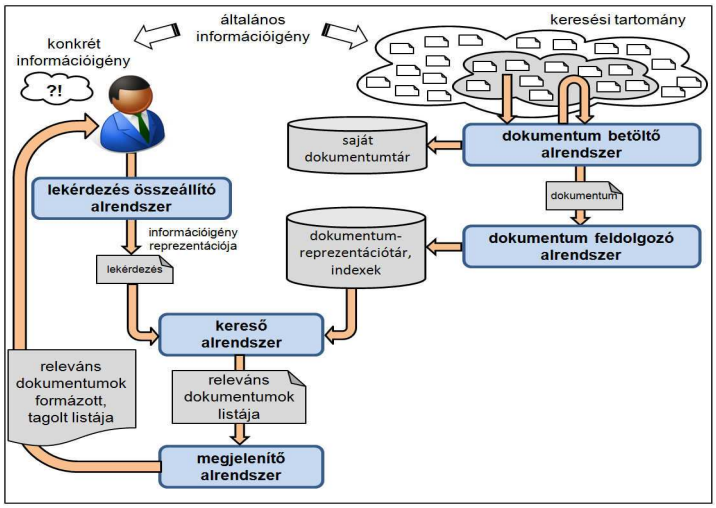






4. ábra: Információkereső rendszer-architektúrák
[2, 4. o.; 3, 2. o.; 4, 10. o. alapján]

Valamennyi architektúrális leírásban megjelenik a felhasználói információigény és az azt reprezentáló keresési igény (lekérdezés), valamint a keresési tartomány (az információkat hordozó dokumentumok köre). A leírásokban jól elkülönülnek a kereséseket előkészítő feladatok és összetevők, valamint a konkrét keresés során végbemenő tevékenységek. Mindezek és a korábbiakban elmondottak alapján az információkereső rendszerek átfogó felépítését, alapvető összetevőit és azok kapcsolatrendszerét a következő ábra szemlélteti.



5. ábra: Információkereső rendszer átfogó felépítése

Az ábrán látható megközelítés alapján az információkereső rendszerek öt alapvető összetevőre, alrendszerre tagolhatóak, amelyek részletesebb vizsgálatára, bemutatására a következő pontban kerül sor.

Az információkeresés a legtöbb megoldás (a későbbiekben bemutatott modellek) esetében megvalósíthatóak a párhuzamos feldolgozás módszereivel is, ami különösen a nagy keresési tartományok, pld. a web esetében szinte megkerülhetetlen. Alkalmazásuk jelentős mértékben növeli a keresés sebességét. Technikailag párhuzamos keresésnek tekinthető, de lényegét tekintve alapvetően eltérő megoldás az úgynevezett összevont keresés, vagy metakeresés,[8] amely különböző — sokszor különböző jellegű dokumentumokat tartalmazó — keresési tartományok egyidejű keresését biztosítja. Az összevont információkereső rendszerekben nem történik dokumentum feldolgozás, nincsenek index-adatbázisok, a keresési igényt megfelelő formába alakítva adják át más kereső rendszereknek, majd azok eredményét hozzák egységes formára, rendezik össze és továbbítják a megjelenítő alrendszernek. Az összevont keresés nehézségei elsősorban az ismétlődések kiszűrésében és az egységes rangsorolás, relevancia-mérték meghatározásában rejlenek.

Az összevont keresés sajátos változata az ún. szervezeti keresés[9], amely egy szervezet különböző információforrásaiban (adatbázisaiban, önálló számítógépes állományaiban, dokumentumkezelő rendszereiben, elektronikus leveleiben, stb.) történő integrált keresést biztosítja. A szervezeti információkereső rendszerek a webes összevont kereső rendszerekkel szemben általában nagyobb összehangoltságot (pld. egységes leíró, kulcsszó/címke rendszert), szerepkörökre szabott hozzáférést és hozzáférési kontrollt biztosítanak

Dokumentum reprezentációk információkereső rendszerekben

Az információkereső rendszerekben a keresési tartományt képező dokumentumok közötti keresés — néhány egyedi megoldástól eltekintve — a dokumentumokat helyettesítő, tartalmukat leíró reprezentációk segítségé-

[8] Federated search, metasearch.
[9] Enterprise search.

vel történik. Ennek oka elsősorban az eredeti dokumentumok méretében, hozzáférhetőségében, valamint a hatékony keresésre való 'alkalmasságában' rejlik. A kereső rendszerek ennek megfelelően az egyes dokumentumokból egy keresést támogató reprezentációt állítanak elő. A számos különböző megközelítésre (modellre) épülő információkereső rendszerek túlnyomó többségének közös megoldása a dokumentumok leíró jellemzők, ún. index kifejezések listájával történő reprezentációja.

Az index kifejezések[10] tágabb értelemben a dokumentum tartalmát, vagy más — a keresés szempontjából fontos — jellemzőjét megjelenítő elemi adatok (szavak, kifejezések, numerikus vagy logikai értékek). Az index kifejezések jellegüket tekintve metaadatok, vagyis olyan adatok, amelyek más adatokat (esetünkben információkat hordozó dokumentumokat) határoznak meg, vagy írnak le. [5, 10. o.] A metaadat strukturált adat, amely leír, érthetővé tesz, feltár vagy más módon megkönnyíti egy információs erőforrás visszakeresését, használatát, vagy kezelését. [6, 1. o.]

A metaadatok egy értelmezés szerint három típusba sorolhatóak. A leíró metaadatok azonosítás és feltárás, megtalálás céljából írják le az információs erőforrást. Ezek közé tartozhat pld. a szerző, a cím, a tartalmi összefoglaló, vagy a kulcsszavak. A strukturális metaadatok az összetett erőforrások felépítését, összetevőit írják le. Végül az adminisztratív metaadatok közé az erőforrások kezelését, megőrzését segítő, köztük technikai jellemzők tartoznak. [6, 1. o.] Az információkeresés során használt index kifejezések között a legjelentősebb szerepet a leíró metaadatok játsszák, azonban sok esetben felhasználásra kerülnek adminisztratív metaadatok (pld. formátum, méret, elérhetőség helye, stb.) is.

A dokumentumot a keresés során reprezentáló leíró jellemzők egy része (azonosító adatok, technikai jellemzők, stb.) a dokumentum alapján egyértelműen, szükség esetén automatizált módon meghatározható. Másik — a keresés szempontjából talán fontosabb — részük, a dokumentum tartalmát leíró jellemzők meghatározása, kiválasztása a szöveges és média dokumentumok esetében viszont sokkal bonyolultabb feladat. A dokumentumok tartalmának leírása, jellemzése a manuális feldolgozásra épülő könyvtári információkereső rendszerekben először kategóriákba (később hierarchikus rendszerbe) történő besorolással[11], ké-

[10] Term, index term, subject term, descriptor.
[11] Pld. Egyetemes Tizedes Osztályozás (ETO).

sőbb a tartalmat tömören jellemző szavakkal, kifejezésekkel (tárgyszavak, kulcsszavak) történt. Ezek a megoldások — a dokumentumok létrehozói révén — részben még ma is működőképesek, azonban a tartalom-előállításban érintettek és a keresési tartományt képező dokumentumok körének óriási mértékű kibővülésével és ezzel az automatizált tartalom leírás, jellemzés előtérbe kerülésével szerepük nagyon háttérbe szorult.

A szöveges dokumentumok tartalmát reprezentáló kifejezések származhatnak a dokumentum szövegéből és lehetnek a szövegben szereplő szavaktól formailag független, a tartalmat jól leíró kifejezések. Előbbiek meghatározásának alapját a természetes nyelv feldolgozás[12] módszerei, eszközei, informatikai megoldásai képezik. Az utóbbiak meghatározásához már szemantikus technológiák és a szövegek mellett formalizált tudásreprezentáció (ezen belül fogalomrendszer reprezentáció[13]) is szükséges. [7, 276-277. o.]

A dokumentumot reprezentáló kifejezések száma több szempontból is befolyásolja az információkeresés eredményességét. Kevés, erősen jellemző kifejezés csökkenti a teljességet (a megtalált releváns dokumentumok arányát az összeshez képest), de növeli a pontosságot (a releváns dokumentumok arányát a megtaláltak között), míg sok kifejezés növeli a teljességet, viszont csökkenti a pontosságot.

Az ún. teljes szöveges keresés esetében lényegében a dokumentum valamennyi 'érdemi' szava[14] részét képezi a dokumentum reprezentációjának. Ebben az esetben különböző technikákkal, megfelelő reprezentációval biztosítható, hogy a több szóból álló kifejezések, szövegrészek alapján történő keresés is lehetséges legyen.

A (multi)média dokumentumok tartalmát reprezentáló kifejezések meghatározhatóak a dokumentumok esetleges szöveges összetevői (felirat, kiegészítő leírások, szöveggé alakított beszéd, stb.) alapján, azonban a tartalom érdemi leírása — amennyiben nem emberi közreműködéssel történik — speciális megoldásokat igényel. Napjainkban a multimédia dokumentumok keresése még jellemzően médiatípus specifiku-

[12] Natural Language Processing (NLP).

[13] Pld. ontológiák, taxonómiák.

[14] Általában a néhány betűből álló vagy nagyon gyakori, önálló jelentéssel nem bíró – ún. tiltólistás, 'stop word' – szavak (pld. névelők, kötőszavak, igekötők) kivételével, amelyeknek a keresések szempontjából semmi jelentőségük nincs.

san történik és ehhez igazodóan a keresési szempontok és így a média tartalmak reprezentációja, a tartalmat leíró jellemzők is típus specifikusak.[15]

Információkereső rendszerek összetevői

Az információkereső rendszerek előző pontban meghatározott összetevői, alrendszerei egymástól többé-kevésbé függetlenül, időben is általában két időszakra szétválva működnek. A konkrét keresést megvalósító, központi szerepet játszó kereső alrendszer csak megfelelő előkészítés alapján, más alrendszerek eredményeit felhasználva működőképes és eredményeit egy további alrendszer bocsátja a felhasználó számára legmegfelelőbb formában rendelkezésre. A különböző alrendszerek természetesen eltérő rendeltetéssel bírnak, ebből következően eltérő funkciókat, feladatokat valósítanak meg és ehhez illeszkedően eltérő módszereket, eszközöket használnak. A következőkben egyenként áttekintjük az információkereső rendszerek öt alapvető összetevőjét, meghatározzuk rendeltetésüket, bemenetüket és kimenetüket, működésük elveit, valamint főbb alkalmazott megoldásaikat.

Kereső alrendszer

A keresést végrehajtó alrendszer rendeltetése az információkereső rendszer előírásainak megfelelő formában megfogalmazott konkrét információigény és a már feldolgozott dokumentumokról összegyűjtött, összeállított információk alapján az igénynek megfelelő dokumentumok listájának meghatározása. Az alrendszer bemenetét a lekérdezés összeállító alrendszer által létrehozott információigény reprezentáció, illetve a dokumentum feldolgozó alrendszer által létrehozott dokumentum reprezentációk és a teljes dokumentumkörre vonatkozó adatok képezik. Az alrendszer eredménye a kiválasztott dokumentumok listája, amelyet a megjelenítő alrendszer bocsát a felhasználó rendelkezésére.

[15] Pld. fényképeken, videókon szereplő objektumok (személyek, tárgyak), helyszínek, események/tevékenységek; hangfelvételeken hallható szereplők (előadók), jellegzetes hangok, dallamok.

Az információkeresés modelljei, az információkereső rendszerekben alkalmazott megoldások csoportosítására, taxonómiájára a szakirodalomban számos javaslat született, amelyek a több mint tíz modellt különböző szempontok alapján osztályozzák (ezek jó áttekintését adja [3] és [10]). Jelen publikációban a keresés matematikai alapjai szerinti osztályozásra építünk, amely halmazelméleti, algebrai és valószínűségelméleti alapú megoldásokat különböztet meg. Mindhárom csoportban létezik egy ún. klasszikus modell, amelyek alapján az idők során újabb modellek születtek. [4] A továbbfejlesztések mindhárom csoportban elsősorban a keresés során felhasznált leíró jellemzők (index kifejezések) közötti függőségek figyelembevételéhez kapcsolódtak.

A halmazelméleti alapú információkereső modellek klasszikus változata — egyben az információkeresés legelső megoldása — az ún. Boole modell. Ebben a lekérdezés, a kereső kérdés leíró jellemzők[16] logikai kifejezése. Minden egyes leíró jellemzőhöz tartozik egy dokumentumhalmaz, amelyekben a jellemző szerepel, a keresés eredménye pedig ezekből a lekérdezésben szereplő logikai kifejezésnek megfelelően halmaz-algebrai úton meghatározott dokumentumhalmaz. A módszer sajátossága, egyben hiányossága, hogy a lekérdezésnek megfelelő dokumentumokat nem rangsorolja, nem enged részleges megfelelést és nincs mód a leíró kifejezések súlyozására. Mindezekből következően a Boole modellre épülő információkeresés gyakran túl kicsi, vagy túl nagy eredményhalmazt szolgáltat.

A kiterjesztett Boole modell a későbbiekben bemutatandó vektortér modellből vesz át megoldásokat (leíró jellemzők súlyozása, hasonlóság mértékek), ezzel biztosít rangsorolást és szolgáltat részlegesen megfelelő megoldásokat is. Az eredeti modell kiterjesztésének tekinthető az életlen (fuzzy) halmazokra épülő életlen (fuzzy) kereső modell. Ebben egy leíró jellemző halmazához egy dokumentum [0,1] közötti értékekkel jellemzetten — kevésbé és erősebben — tartozik, a keresés eredménye pedig ezen életlen halmazokon elvégzett, egyes változatokban a súlyozást is figyelembe vevő 'puha Boole' műveletek alapján kerül meghatározásra.

Az algebrai alapú információkereső modellek klasszikus változata a vektortér modell. A modell lényege, hogy a dokumentumokat egy sok-

[16] A leíró jellemző lehet kulcsszó (tárgyszó), vagy egy 'jellemző=érték' páros.

dimenziós tér vektoraként írja le, ahol a tér dimenzióit a leíró jellemzők alkotják, a koordináták pedig az adott jellemzőnek a dokumentumra vonatkozó súlyát, fontosságát reprezentálják (0, ha nem szerepel, nem jellemző)[17]. A lekérdezés szintén leírható ilyen vektorként és a keresés, rangsorolás alapját (a relevancia mértékét) a vektorok közötti hasonlóság-mérték[18] képezi.

A leíró jellemzők dokumentumra vonatkozó súlyának meghatározására különböző megoldások születtek. A leggyakrabban használt változat figyelembe veszi a jellemző dokumentumon belüli gyakoriságát és a jellemző gyakoriságát (ritkaságát) a dokumentum-gyűjteményen belül (inverz dokumentum gyakoriság): W(T,D,C) = TF(T,D) * IDF(T,C).[19]

Az eredeti vektortér modell hiányossága, hogy nem veszi figyelembe a leíró jellemzők közötti összefüggéseket, függőségeket. Ennek feloldására születtek továbbfejlesztett modellek, mint az általánosított vektortér modell, amely a hasonlóság-mértékben figyelembe vesz a leíró jellemzők közötti korrelációt leíró értékeket, amelyeket más módszerekkel (a dokumentum-gyűjtemény feldolgozásával, vagy szemantikus eszközökkel) kell meghatározni. A rejtett szemantikus indexelés lényege a leíró jellemzők terének algebrai módszerekkel történő átalakítása dokumentum tulajdonságok – dokumentumokat jellemző fogalmak — terére. Ezt követően a leíró jellemzők és ezek alapján a dokumentumok és a lekérdezések is e 'rejtett' — sokszor szavakkal értelmesen le sem írható — fogalmak vektoraiként kerülnek meghatározásra, a keresés pedig ebben a térben kerül végrehajtásra. A módszer egyik nagy előnye a többjelentésű és az azonos jelentésű szavakból fakadó problémák kiküszöbölése.

A valószínűségelméleti alapú információkereső modellek klasszikus változata[20] a dokumentumokat a vektortér modellhez hasonlóan leíró jellemzők bináris vektoraként írja le és feltételezi, hogy a leíró jellemzők egymástól függetlenül oszlanak meg a releváns, illetve nem releváns dokumentumok körében. A modell a dokumentumokat annak valószínűsége alapján rangsorolja, választja ki, hogy az adott kereső kérdés-

[17] A koordináta értékkészlete lehet diszkrét (0 vagy 1), vagy folytonos [0,∞].
[18] Jellemzően a koszinusz-távolság.
[19] T = leíró jellemző (term), D = dokumentum, C = dokumentum-együttes (corpus), W = súly, TF = leíró jellemző gyakoriság a dokumentumban, IDF = inverz dokumentum gyakoriság.
[20] Binary Independence (Retrieval) Model = bináris függetlenség modell.

re relevánsak-e. A kiválasztás kritériuma, hogy a relevancia valószínűsége nagyobb legyen, mint a nem relevanciáé[21]. A modell a dokumentum relevanciájának (adott lekérdezésre vonatkozó) valószínűségét visszavezeti az egyes leíró jellemzőknek a releváns és nem releváns dokumentumokban történő előfordulásának valószínűségére. Ezek az értékek a releváns és nem releváns dokumentumok ismeretében meghatározhatóak, részleges ismeretükben, vagy annak hiányában különböző értékekkel közelíthetőek, becsülhetőek, illetve a felhasználói visszacsatolás révén iteratív módon javíthatóak.

A valószínűségelmélet eszköztárára alapozott további modellek közé tartozik a Bayes hálózatokra épülő következtetési hálózat modell. Ez a modell a dokumentumok, az őket leíró jellemzők, az információigény és az azt leíró jellemzők közötti feltételes valószínűségek hálózata alapján határozza meg, hogy az egyes dokumentumokból milyen valószínűséggel 'következik' egy adott információigény kielégítése. Az első két csomópont típus közötti hálózatrész (dokumentum hálózat) előzetesen elkészíthető, a másik két típust magában foglaló lekérdezés hálózatot esetenként kell kialakítani. A valószínűségi nyelvi modell az egyes dokumentumok nyelv statisztikai modelljére (szavai, szó-sorozatai valószínűségeloszlására) épül. Ezek felhasználásával lehet meghatározni, hogy egy adott lekérdezés milyen valószínűséggel származtatható a dokumentum nyelvi modelljéből.

Dokumentum betöltő alrendszer

A dokumentumokat betöltő alrendszer rendeltetése az érintett, a keresési tartományt képező dokumentumok átvétele, vagy megkeresése és ezt követően rendelkezésre bocsátása a dokumentum feldolgozó alrendszernek. Az alrendszer alapvető feladata, hogy fenntartsa az általános információigények kielégítéséhez szükséges, a keresés során figyelembe vehető/veendő dokumentumok körének lehetséges teljességét.

Feladata lehet továbbá az is, hogy a heterogén formában beérkező, elért dokumentumokat az információkereső rendszer által preferált, egységes formátumra alakítsa.

[21] Vagyis esélye, esélyértéke (odds) nagyobb legyen mint 1.

A dokumentumok átvétele a betöltés passzív változata, amikor az új, vagy megváltozott tartalmú dokumentumok külső kezdeményezésre kerülnek be az információkereső rendszerbe (pld. könyvtári dokumentumok feldolgozása; új jogszabályok, módosítások beépítése; kép- és hangfelvételek tárolása; hírfolyamok, elektronikus levél folyamok feldolgozása stb.) Erre a megoldásra jellemzően integrált tároló és visszakereső rendszerek, vagy dokumentum folyamok esetében van lehetőség és ez jelent egyedüli megoldást hagyományos dokumentumok (pld. könyvek, folyóiratcikkek, stb.) keresésének előkészítése esetében. A dokumentumok megkeresése a betöltés aktív változata, amikor a kereső rendszer a dokumentumokat tároló rendszer aktív közreműködése nélkül gyűjti össze, vonja be az információkereső rendszerbe az új, vagy megváltozott dokumentumokat. Az elérhető dokumentumok nagy száma miatt ez jellemzően automatikusan történik, az ezt megvalósító rendszer-összetevők megnevezése: kereső robot.[22]

A kereső robotok először a World Wide Weben elérhető információk keresésének támogatására jelentek meg, de alkalmazhatóak bármely, informatikai rendszerben tárolt és elérhető dokumentumkör esetében. Az automatikus keresés alapvető kérdése, egyben szép szakmai feladata, hogy milyen módon és rendben lehet és célszerű bejárni a feldolgozandó dokumentumokat, hogyan lehet eljutni hozzájuk, mekkora részüket sikerül elérni.

A hagyományos kereső rendszerek robotjai által elért, feldolgozott weblapok egyes felmérések szerint a World Wide Webnek csak mintegy egykétezred részét képezik és az információigények szempontjából fontos dokumentumok éppen a nagyobb részt alkotó mély, rejtett, vagy láthatatlan részben[23] találhatóak. Emiatt a kereső robotok megoldásai, technológiái a kutatások ígéretes részét képezik.

Az információkeresés speciális esetének tekinthető a dokumentumok folyamában történő keresés, a megadott feltételeknek megfelelő dokumentumok kiszűrése. Erre elsősorban a hírszerzés, információgyűjtés különböző változataiban kerülhet sor, amikor a 'keresési tartományt' alkotó dokumentumok (pld. elektronikus levelek, telefonbeszélgetések, műsorok, rádió és tévéműsorok, hírszolgáltatások) tárolt változatai nem érhetők el, de

[22] Search (ro)bot, web crawler (bejáró, 'csúszómászó'), web spider ('pók').
[23] Deep web, hidden web, invisible web.

áramlásuk során hozzáférhetőek. Ebben az esetben a betöltés, a feldolgozás és a keresés együtt valósul meg.

Dokumentum feldolgozó alrendszer

A dokumentumokat feldolgozó alrendszer alaprendeltetése, hogy az egyes dokumentumokhoz a keresések céljára alkalmas – az adott keresőrendszer jellegétől függő típusú – reprezentációt készítsen és a keresést támogató, a teljes dokumentumkörre vonatkozó, további adatokat alakítson ki, tartson naprakészen. Az egyes dokumentumok keresés során közvetlenül felhasznált reprezentációi lehetnek az eredetivel tartalmilag lényegében azonosak, de formailag eltérőek, vagy tartalmazhatnak (általában jóval) kevesebb információt. Az alrendszer végső eredményét tehát a konkrét keresés során felhasználható egyedi dokumentum reprezentációk, illetve a teljes dokumentumkörre vonatkozó adatok alkotják.

A dokumentumok feldolgozása, a keresés során felhasználható reprezentáció létrehozása bonyolultság szempontjából két részre osztható. Mint azt korábban már említettük, a reprezentációt alkotó leíró jellemzők egy része (azonosító adatok, technikai jellemzők, stb.) a dokumentum alapján könnyen, egyértelműen meghatározható. A tartalmat leíró — a keresés során talán legfontosabb — jellemzők nagy része viszont általában nem szerepel a dokumentumban (média dokumentumok esetében pld. csak rendkívül ritkán és korlátozott mértékben). A továbbiakban részletesebben csak ez utóbbiakkal foglalkozunk.

A dokumentumok feldolgozása (tartalmi jellemzése) a keresési sebesség érdekében jellemzően a keresést időben megelőzően történik, de természetesen végrehajtható a keresés során is. Előbbi esetben a feldolgozás csak az általános keresési igényekre és a korábbi keresések tapasztalataira épülhet, míg a második esetben szorosabban igazodhat a konkrét kereséshez. Az előzetesen végrehajtott feldolgozás esetén is szükség lehet újrafeldolgozásra, például az általános keresési igények módosulása, illetve a keresési tapasztalatok változása esetén.

A dokumentumban nem szereplő leíró jellemzők azért szükségesek, mert egy adott (szöveges, rajz, kép, hang, stb.) dokumentum elemi szinten — néhány leíró jellemzőtől (pld. cím, szerző, stb.) eltekintve — csak az alkalmazott reprezentáció jellegének megfelelő összetevőket

(szavak, írásjelek, szövegformázás; vonalak; képpontok; hangelemek) tartalmaz, nem pedig az ezek által önállóan, vagy együttesen hordozott információkat. Az információkeresés pedig nem a reprezentációs elemek (szavak, képpont-mintázatok, stb.), hanem az általuk hordozott információk keresésére irányul.

A dokumentum tartalmát képező (az általa hordozott) információkat a felhasználás (olvasás, megtekintés, meghallgatás) során az emberi agy hozza létre. Ezt a folyamatot kell az általános keresési igényekhez igazodóan, legalábbis részeiben megvalósítani, helyettesíteni és az eredeti – nem strukturált, vagy félig-strukturált – reprezentáció mellett további, a hordozott tartalmat leíró (esetleg magára a dokumentumra és elkészültére vonatkozó) strukturált információkat előállítani.

A dokumentum reprezentációt alkotó leíró jellemzők (index kifejezések) két alapvető – részben eltérő – típusa a tárgyszó és a kulcsszó. A tárgyszó[24] a könyvtártudomány szakkifejezéseként, a könyvtári dokumentumok manuális feldolgozásához kapcsolódóan jelent meg. Az alkalmazott tárgyszó rendszer lehet szabad, de a gyakorlatban jellemzően kötött (előre meghatározott körből választható).

Ez utóbbiak közé tartoznak a hierarchikus osztályozási rendszerek, valamint a további, összetettebb fogalmi kapcsolatokat tartalmazó tezauruszok is. A dokumentumok, információegységek tartalmi leírásának és ezzel a keresés elősegítésének eszköze a kulcsszó, címke[25] is. Ezeket jellemzően a szerző határozza meg szabadon, például a tudományos publikációk esetében és különösen az Internet világában (hírek, blogbejegyzések, képek, videók, elektronikus levelek, stb.).

Az indexelés — az információkeresés szempontjából — a dokumentum leírása, osztályozása, olyan leíró jellemzőkkel (meta adatokkal) történő ellátása, amelyek jól jellemzik a dokumentum tartalmát, segítik keresését, megtalálhatóságát. Meghatározásuk történhet manuálisan és automatizált módon.

Az információkeresés történetében elsőként alakult ki és mindmáig legfejlettebb a szöveges dokumentumok feldolgozásának, indexelésének módszertana, eszköztára, azonban az 1990-es években megjelent a média dokumentumok keresési célú feldolgozása. Az egyes dokumentu-

[24] Subject term, index term, descriptor.
[25] Keyword, tag.

mokat reprezentáló index kifejezés listák általában nem önállóan, hanem a keresés céljait segítő adatstruktúrákban, ún. indexekben kerülnek tárolásra.

A szöveges dokumentumok tartalmi feldolgozása emberi közbeavatkozás nélkül a szöveget alkotó szavakra és a szöveg struktúrájára épül, az automatizált indexelésre különböző szövegfeldolgozó megoldások alakultak ki. Egyes módszerek a legjellemzőbbnek ítélt szavakat (kifejezéseket) határozzák meg, a teljes szöveges megoldások pedig gyakorlatilag minden 'érdemi' szót felvesznek az indexbe[26]. Az index kifejezések jellemzően a dokumentum szavainak nyelvészetileg egyszerűsített (pld. szótőre szűkített) változatai. Az információkeresés egyes modelljei esetében az index kifejezésekhez jelentőségüket, az adott dokumentumot jellemző erejüket, relevanciájukat leíró értékeket (súlyokat, valószínűségeket) is kell rendelni.

A média dokumentumok tartalmi feldolgozása emberi beavatkozás nélkül a kép- és hangfeldolgozás, alakfelismerés technikáira épül. A tartalom-alapú média dokumentum keresés alapját a dokumentumok jellemzői és ezek meghatározása képezik.[27] [részletesebben lásd pld. 8, 9] A keresés során média típusoktól függően különböző szintű és típusú jellemzők használhatóak fel. Az alacsonyabb szintű jellemzők közé tartozhatnak képek esetében textúra, szín- és alakjellemzők, hangfelvételek esetében hangerő, hangmagasság, hangszín jellemzők. Magasabb szintű jellemzők lehetnek képek esetében tárgyak, személyek (arcok), események, hangfelvételek esetében hangszerek, előadók, hangot kiadó dolgok, jelenségek. A média dokumentumok keresésének sajátossága, hogy sok esetben nem megadott leíró jellemzők, hanem egy konkrét dokumentummal fennálló hasonlóság alapján történik.

Lekérdezést összeállító alrendszer

A lekérdezést összeállító alrendszer rendeltetése az információs igény alkalmazott informatikai rendszernek megfelelő formátumú reprezentációban történő előállításának támogatása. A feladat jelentősége az igény

[26] Tulajdonképpen konkordancia = egy könyv, vagy írásmű (legfontosabb) szavainak ábécébe szedett jegyzéke.

[27] Content-based image retrieval (CBIR), content-based audio retrieval (CBAR), feature extraction.

minél pontosabb reprezentációjának kialakításában és a lekérdezés összeállításának minél hatékonyabb megkönnyítésében rejlik. A keresés kezdetekor a keresőnek általában csak részben körvonalazott elképzelései vannak, így a lekérdezést összeállító alrendszernek kell segítenie információigénye értelmezésében és kifejezésében.

A lekérdezés (kereső kérdés)[28] az információkeresési igény meghatározott formátumban (kereső nyelven) történő megfogalmazása. A lekérdezés típusa az információkereső rendszertől, az alkalmazott modelltől és a kereshető dokumentumok típusától függően többféle lehet. Ezek közé tartoznak többek között a kereső kifejezésekre épülő, a természetes nyelvi és a példa alapján történő lekérdezések. A kereső kifejezésekre épülő lekérdezés lehet egyszerűen kereső szavak listája és lehet ezek speciális keresési operátorokkal kiegészített logikai kifejezése. A kereső szavak lehetnek szabadon választhatóak és köthetőek a dokumentumok feldolgozása során alkalmazott, vagy összegyűlt indexkifejezések köréhez. A természetes nyelvi (szabad szöveges) lekérdezés[29] esetében a megadott szöveg — a dokumentumokhoz hasonlóan — leíró kifejezések meghatározásához szolgál alapul, vagy valamennyi szava kereső kifejezésként kerül felhasználásra. Végül a példa alapján történő lekérdezés[30] esetében teljes dokumentumot, vagy dokumentum-részletet lehet megadni és a kereső alrendszer majd ehhez 'hasonló' dokumentumokat fog keresni. Ez utóbbi típus jellemzően média dokumentumok keresése során használatos.

A lekérdezések összeállítása az adott rendszer által biztosított felhasználói felület segítségével történik, amely egyben az eredmények megjelenítésének is eszköze. A lekérdezések összeállításának öt hagyományos módja közé a következők sorolhatóak: parancsnyelvi, űrlap-alapú, menü rendszerű, közvetlen manipulációs[31] és természetes nyelvi. [4, 278. o.] Ezek mellett az ember-gép interakció fejlődésével fokozatosan jelennek meg a más modalitásokra (hang-, gesztus-, stb.), illetve multimodalitásra épülő megoldások is.

[28] Query.
[29] Natural language query, free text query.
[30] Query by example.
[31] Az adott objektumokkal történő – pld. képernyőn mutatóeszközzel megvalósított – interakció révén.

A többszempontú (fazettás) lekérdezés[32] a többszempontú osztályozáshoz [11, 12-15. o.] kapcsolódó – a keresés hatékonyságát növelő – speciális megoldás, amelyben a kereső kifejezések egymástól független, önálló osztályozási szempontok szerinti csoportokban adhatóak meg és ennek megfelelően is kerülnek felhasználásra. A többszempontú keresés hatékonyan segíti a feltáró jellegű, irányított, fokozatosan pontosodó keresést. Ezt lekérdezést összeállító alrendszer a különböző szempontok szerinti választási lehetőségek (listák, taxonómiák, stb.) felkínálásával támogathatja.

A lekérdezés kiegészítése, finomítása[33] a keresési hatékonyság növelésének egyik eszköze elsősorban a kereső kifejezésekre épülő és természetes nyelvi lekérdezések esetében. A funkció lényege, hogy a felhasználó által megfogalmazott lekérdezés a keresés végrehajtása előtt, vagy a keresés folyamatában módosításra kerül. A lekérdezés módosításának legegyszerűbb változata a korábbi lekérdezések felhasználásával, vagy nyelvi eszközökkel (pld. további nyelvtani alakváltozatok beillesztésével) történő bővítés, bővítési javaslat. A lekérdezés emellett bővíthető szinonimákkal, fogalmi kapcsolatban álló kifejezésekkel létező, vagy a dokumentumkör alapján, nyelvi és statisztikai úton generált thesaurusok segítségével. [12, 189. o.]

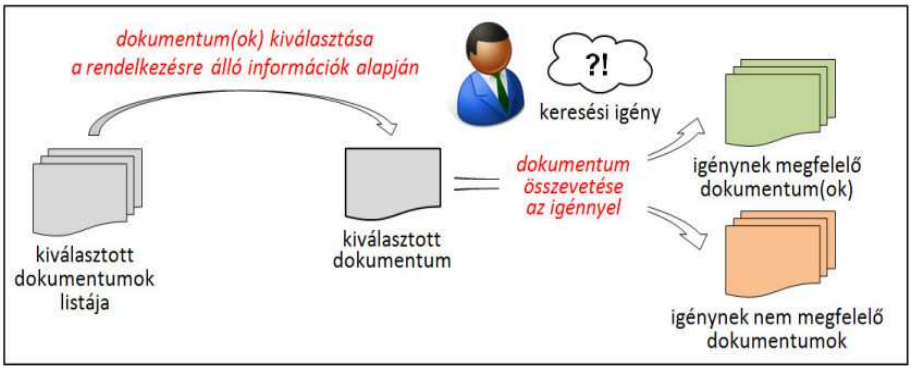
A lekérdezés finomításának alapja lehet a felhasználói visszajelzés, a kapott eredmények relevanciájának értékelése[34] is. A megoldás alapgondolata, hogy a felhasználó első lekérdezése — információk hiányában — csak egy kezdeti próbálkozás és annak eredményei, pontosabban azok relevanciái felhasználásra kerülnek a lekérdezés iteratív finomítására. A finomítás jelentkezhet a lekérdezés bővítésében, vagy a kereső kifejezések súlyozásának módosításában. A visszacsatolás lehet explicit (vagyis a felhasználó által megadott), vagy implicit (vagyis a rendszer által kialakított). Ez utóbbinak két megoldása a lokális (az eredmény halmazra kiterjedő) és a globális (a teljes dokumentumkörre kiterjedő) elemzés. [12, 178-188. o.]

Megjelenítő alrendszer

[32] Faceted query.
[33] Query expansion, query refinement.
[34] User relevance feedback.

A megjelenítő alrendszer rendeltetése, hogy a kereső személy, a felhasználó rendelkezésére bocsássa a kereső alrendszer által kiválasztott, relevánsnak ítélt dokumentumok listáját és — napjaink információkereső rendszereiben már — támogassa az egyes dokumentumok elérését, azok érdemi összevetését a keresési igénnyel. Az alrendszernek alapvető szerepe van abban, hogy a felhasználó milyen eredményességgel és hatékonysággal jut el a számára releváns dokumentumokhoz.

A felkínált eredmények az információkereső rendszer felhasználói felületén érhetőek el, jellemzően a lekérdezés összeállításának lehetőségeivel együtt. Az alrendszer által a felhasználó rendelkezésére bocsátott – jellemzően részekre tagolt – listából a felhasználó egyenként kiválasztja a valószínűsíthetően releváns dokumentumokat, hozzáfér azok teljes tartalmához és ezt összevetve információigényével, dönt a relevanciájáról.



6. ábra: Keresési eredmények feldolgozása

A keresés céljától függően az eredmények feldolgozása befejeződhet az első releváns dokumentumnál, vagy kiterjedhet több releváns dokumentum megtalálásáig. A felajánlott dokumentumlista feldolgozása során, annak eredményei alapján felmerülhet a lekérdezés kiegészítése, finomítása is, amely aztán új — a remények szerint a keresési igényhez jobban illeszkedő — dokumentumlistát eredményez. A folyamat a felhasználó döntésétől függően fejeződik be.

A dokumentum helyettesítő[35] a felkínált dokumentumok listájának elemi összetevője, egy dokumentum alapvető adatainak, tartalma szemléltetésének összessége. A dokumentum helyettesítője tartalmazza a dokumentum címét és fontos leíró jellemzőit (szerző, keletkezési dátum, dokumentumtípus, méret, elérhetőség helye, stb.). Ezek lényegében a könyvtári katalóguscédulák tartalmának megfelelői. A relevánsnak feltételezett dokumentumok hatékony kiválasztásához ezek az adatok általában nem elegendőek, szükség van a tartalom rövid érzékeltetésére is.

A tartalmi kivonatok[36] szöveges dokumentumok esetében a publikációk absztraktjainak szerepét töltik be. A kivonatok napjainkban általában a lekérdezésben szereplő kereső kifejezéseket tartalmazzák szövegkörnyezetükben (esetenként a dokumentum több helyéről összeválogatva), figyelemfelhívó szövegkiemeléssel kiemelve.

A tartalom megítélését természetesen segíti a hosszabb kivonat, ami dinamikusan is megjeleníthető (pld. a kivonatra 'mutatással'). A megfelelő tartalmi kivonat konkrét információ keresése esetén önmagában megadhatja a keresett választ. Képi dokumentumok esetében pedig a tartalmi kivonat egyszerűen maga a dokumentum, pontosabban annak kisebbített, alacsonyabb felbontású képe.

A keresési eredmények csoportosítása a keresés hatékonysága növelésének egyik lehetséges megoldása (részletesebben lásd pld. [13]). Ennek során a dokumentumlista rendező elve nem egyedül a számított relevancia, ezen kívül a dokumentumok a keresést könnyítő csoportokba is besorolásra kerülnek.

A csoportok lehetnek tartalmi hasonlóság alapján képzett klaszterek, előre meghatározott kategóriák szerinti besorolások, vagy a többszempontú lekérdezés során már említett 'fazetták'. A különböző csoportok elősegítik, hogy a felhasználó a számára fontos dokumentumokat válassza ki. A csoportosítás mellett napjainkban már megjelentek az eredménylista értelmezését, felhasználását segítő vizualizációs megoldások is.

Összegzés, következtetések

[35] Document surrogate.
[36] Abstract, extract, excerpt, snippet.

Jelen publikáció a strukturálatlan (elsősorban szöveges, ezenkívül multimédiás) információk közötti keresést megvalósító információ visszakereső (továbbiakban röviden információkereső) rendszerek alapvető tulajdonságait összegezte, rendszerezte, főbb összetevőiket és azok rendeltetését, megoldásait azonosította. A publikációban foglalt legfontosabb megállapítások a következőkben összegezhetőek.

Az információkereső rendszerek magukban foglalhatják a keresési tartományt képező dokumentumokat (amikor valójában információtároló és visszakereső rendszerről beszélhetünk) és lehetnek azoktól független megoldások, amilyenekkel napjainkban gyakrabban találkozhatunk. Az informatikai eszközökkel támogatott információkeresés szerepe a XX. század végére jelentősen megnőtt, ennek megfelelően a dokumentumok emberi kezelése, értelmezése helyébe részben, vagy teljesen az informatikai eszközök által végzett műveletek léptek.

Az információkereső rendszerek – mint minden rendszer és ezen belül minden informatikai rendszer – jól elkülöníthető funkciójú alrendszerekre tagolhatóak. A szakirodalomban található felosztásokat is felhasználva jelen publikáció egy öt alrendszerre tagolt felépítést alkalmaz. Napjaink információkereső rendszereiben a konkrét keresés már (többnyire) nem az eredeti dokumentumok, hanem azok reprezentációi között történik. A keresési tartományt képező ('kereshető') dokumentumokat a keresés során a tartalmat és néhány más fontos jellemzőt megjelenítő (meta)adatok, ún. index kifejezések helyettesítik. Az index kifejezések származhatnak a dokumentumból, meghatározhatóak automatizált módon, azonban a tartalom 'jó' jellemzése bonyolult feladat.

Az információkereső rendszer kulcsfontosságú, központi összetevője a kereső alrendszer, amelynek rendeltetése a megfelelő formában megfogalmazott konkrét információigény (lekérdezés) és a már feldolgozott dokumentumokról összegyűjtött, összeállított információk alapján az igénynek megfelelő dokumentumok listájának meghatározása. Az idők során számos különböző keresési modell jelent meg, amelyek a matematikai alapok szerint három nagyobb (halmazelméleti, algebrai, valószínűségelméleti) csoportba sorolhatóak. Minden egyes csoporton belül több megoldás létezik, amelyek egyre fejlődő szolgáltatásokat nyújtanak alkalmazóik számára.

Az információkeresés feltételét a dokumentum betöltő és a dokumentum feldolgozó alrendszerek teremtik meg, amelyek közül az előbbi a konkrét keresést megelőzően aktív módon megkeresi, vagy passzív módon átveszi a keresési tartományt képező (új, vagy megváltozott) dokumentumokat, kiküszöböli az eltérő formátumokból származó különbözőségeket, míg az utóbbi a dokumentum alapján elkészíti a keresés során használható reprezentációt (leírást). Ennek során egyes leíró jellemzők könnyebben előállíthatóak, míg a tartalmat jellemző indexkifejezések meghatározása (az indexelés) az információkereső rendszerek eredményességét alapvetően befolyásoló feladat.

A lekérdezést összeállító és az eredményeket megjelenítő alrendszerek az információkereső rendszer felhasználói felületének részét képezik. Az előbbi rendeltetése, hogy lehetővé tegye, segítse a felhasználó információigényének minél pontosabb megfogalmazását az információkereső rendszer követelményeinek megfelelően. Az utóbbi rendeltetése pedig, hogy a felhasználó rendelkezésére bocsássa a relevánsnak ítélt dokumentumok listáját és támogassa az egyes dokumentumok elérését, azok érdemi összevetését a keresési igénnyel. Ennek során szükség lehet a lekérdezés kiegészítésére, finomítására is.

Mint azt a bevezetőben is említettük, napjaink információkereső rendszereinek, megoldásainak egyre inkább elengedhetetlen összetevője a szemantikus (jelentésorientált, tartalmi megközelítésű) technológiák alkalmazása. Ezek nélkül technikai eszközökkel nem lehet hatékonyan támogatni az alapvetően jelentés-alapú, tartalmi információigények kielégítését. Ezek a szemantikus technológiák (a szemantikus keresés megoldásai) megjelenhetnek az információkereső rendszerek valamennyi összetevőjében, alrendszerében, illetve a keresési tartományt képező dokumentumokban is. Így a jelen publikációban foglaltak reményeink szerint megfelelő átfogó keretet nyújtanak a szemantikus keresés megoldásai vizsgálatához, az információkereső rendszerekben elfoglalt helyük, szerepük meghatározásához.

Felhasznált irodalom

[1] Munk Sándor: Az információkeresés alapjai. – Hadmérnök, 2013 (VIII.)/1. (242-254. o.)

[2] Rijsnbergen, C. J. van: Information retrieval. 2nd edition. – Butterworth, London, 1979.

[3] Hiemstra, Djoerd: Information Retrieval models. – In. Goker, Ayse-Davies, John (szerk.): Information retrieval: Searching in the 21st Century. John Wiley and Sons, 2009.

[4] Baeza-Yates, Ricardo–Riberio-Neto, Berthier: Modern information retrieval. – ACM Press, New York, 1999.

[5] ISO/IEC 11179-1, Information Technology – Metadata Registries (MDR) – Part 1: Framework. Second Edition. – ISO, Geneva, 2004

[6] Understanding Metadata. – National Information Standards Organization, Bethesda, 2004

[7] Koltay Tibor–Prókai Margit: Terminológiai változások a XX-XXI. századi könyvtártudományban. – Magyar Terminológia, 2010 (III.)/2. (269-284. o.)

[8] Long, Fuhui–Zhang, Hongjiang–Feng, David Dagan: Fundamentals of Content-Based Image Retrieval. – In. Multimedia Information Retrieval and Management, Technological Fundamentals and Applications, Springer, 2003

[9] Mitrović, Dalibor–Zeppelzauer, Matthias–Breiteneder, Christian: Features for Content-Based Audio Retrieval. – Advances in Computers, 2010 (78.)/71-150.

[10] Champclaux, Yael–Dkaki, Taoufiq–Mothe, Josiane: An information retrieval models taxonomy based on an analogy between cognitive science and information retrieval. – In. Colloque Veille Stratégique Scientifique et Technologique (VSST 2010), Toulouse, 2010.

[11] Hemerly, Jess: Classification. – In. Glushko, Robert J.-Borgman, Christine L. (szerk.): Intellectual Foundations for Information Organization and Information Retrieval. Draft. 2011. http://people.ischool.berkeley.edu/~glushko/IFIOIR/Chapter6-20100917.pdf

[12] Manning, Cristopher D.-Raghavan, Prabhakar-Schütze, Hinrich: An Introduction to Information Retrieval. Online Edition. – Cambridge University Press, Cambridge, 2009. http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[13] Käki, Mika: Findex : Search Results Categories Help Users when Document Ranking Fails. – In. Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, 2005. április 2-7., 131-140. o.