HADMÉRNÖK

VIII. Évfolyam 1. szám - 2013. március

Munk Sándor
munk.sandor@uni-nke.hu

AZ INFORMÁCIÓKERESÉS ALAPJAI

Absztrakt

Az elérhető információk mennyiségének ugrásszerű bővülésével az információkeresés feladatai is megnehezedtek, megváltoztak. Az információhoz jutás egyik lehetősége a már meglévő információk közötti keresés. Napjainkban az információk már túlnyomórészt informatikai rendszerekben kerülnek tárolásra és egyre bővülő mértékben azokban is keletkeznek. Az információkeresés számos más alkalmazási terület mellett a védelmi szférában is jelentős lehetőségeket kínál. Az eredményes kutatás és alkalmazás e téren is szükségessé teszi a különböző értelmezések, megközelítések áttekintését, elemzését. Ennek érdekében jelen publikáció az információkeresés alapvető kérdéseinek rendszerezését, egységes keretbe foglalását tűzte ki céljául. Ezen belül: bemutatja a keresés általános értelmezését; meghatározza az informatikai eszközökkel támogatott információkeresés alapfogalmait, főbb típusait és javaslatot tesz az információkeresés általános fogalmára.

With the dramatic expansion of the amount of available information, the tasks of information search, retrieval have changed, became more difficult. One of the possibilities to acquire necessary information is searching between existing information. Today information is mainly stored in IT systems and in growing extent is also created in these systems. Besides a number of other application areas, information search/retrieval provides significant opportunities for the defense sector too. Effective research and application in this field also requires the review and analysis of the variety of different interpretations, approaches. For this reason this publication aims to systematize, and organize into a unified framework the fundamental issues of information search/retrieval. In particular: presents the general interpretation of searching; defines the basic concepts of computer aided information search/retrieval, sets out the its main types, and proposes a comprehensive definition of information search.

***Kulcsszavak**: keresés, információkeresés, szemantikus keresés, informatikai szolgáltatások ~ search, information retrieval, semantic search, IT services*

BEVEZETÉS

Az információk szerepe, jelentősége a történelem során soha sem volt kérdéses. Ami egyes korokban ugrásszerűen megváltozott és ezzel megváltoztatta az információs tevékenységek lehetőségeit, az az írás megjelenésétől a számítógépek és hálózataik megjelenéséig tartó technikai fejlődés. A folyamatosan fejlődő információs technológiák jelentős mértékben, napjainkban már a feldolgozhatóság határain is túl bővítették az elérhető információk körét. Az információrobbanás, a korábban alig ismert és használt előtét tagokkal leírt peta-, exa- és zetta-bájtokkal jellemezhető világméretű információtömeg a mindent elöntő információáradat új megoldásokat követel, különben – mint ahogy azt John Naisbitt jövőkutató már 1982-ben megfogalmazta – "belefulladunk az információkba, miközben éhezünk a tudásra".

A szervezeti, vagy egyéni célok, feladatok megvalósításához szükséges információk megszerzésének egyik lehetősége a már meglévő és valamilyen rögzített formában rendelkezésre álló információk közötti keresés. Ez természetesen elvileg végrehajtható az összes információ átnézésével és a megfelelő információk kiválasztásával, azonban ez már a sumér agyagtábla 'könyvtárakban' sem volt hatékony, megvalósítható megoldás. Ebből következően vált aktuális kutatási és szakterületté az információkeresés és maradt is mindmáig az.

A rendelkezésre álló információk napjainkban már túlnyomórészt informatikai rendszerekben kerülnek tárolásra, sőt egyre növekvő mértékben azokban is keletkeznek. Lassan oda jutunk – egyetértünk-e ezzel, vagy sem; 'jó'-e ez, vagy sem – hogy, ami nem érhető el informatikai eszközökkel, az nincs is.[1] Ebből következően és az információtechnológia lehetőségeinek bővülésére támaszkodva az információkeresés módszerei, megoldásai között is mindenekelőtt az informatikai támogatásra épülőek játszanak jelentős szerepet.

Az informatikai eszközökkel támogatott hagyományos keresés az adatbázis-kereséstől, a szöveges információtárolás és visszakeresésen át, a világháló keresőrendszereiig a 2000-es évek elején eljutott a jelentés és a szövegösszefüggés kérdéseit is figyelembe vevő szemantikus keresésig. Az információkeresés és a szemantikus keresés a védelmi szférában is jelentős szerepet játszik. Csak példaként felsorolva, ide tartozik a nyílt forrású hírszerzés katonai, nemzetbiztonsági, rendőri célú felhasználás; vagy a szervezeti tudás széleskörű elérhetőségének, megosztásának, hasznosításának megoldásai. Mint minden szakterület esetében, a szakirodalomban itt is számos különböző kifejezéssel, értelmezéssel, megközelítéssel találkozhatunk, amely nehezíti az e téren folytatott kutatások eredményeinek hasznosítását, egymásra épülését.

Mindezek alapján jelen publikáció célja az információkeresés alapvető kérdéseinek összegezése, rendszerezése és egységes keretbe foglalása. Ennek érdekében:

- bemutatja a keresés általános fogalmát, értelmezését, meghatározza folyamatát, alapvető összetevőit, valamint az információkeresés meghatározó sajátosságait;
- meghatározza az informatikai eszközökkel segített információkeresés fogalmait, főbb típusait, azok sajátosságait;
- valamint javaslatot tesz az információkeresés általános fogalmára.

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

A KERESÉS ALAPJAI

A keresés széles körben létező, megfigyelhető tevékenység (jelenség), amely nem korlátozódik az emberekre. A helyváltoztatásra képes élőlények fiziológiai szükségleteik kielégítésére általában többféle dolog keresésére kényszerülnek: pld. táplálék, menedék, pihenőhely, fajtársak, stb. Ennek részét képezi a bejárás, az érzékelés, a minősítés, a felismerés és a választás. A továbbiakban a keresés fogalmát témánk szempontjából az emberi tevékenységre szűkítve vizsgáljuk és bemutatjuk a fogalom köznapi értelmezését, röviden, mélyebb elemzés nélkül számba vesszük legfontosabb összetevőit, végül megfogalmazzuk az információkeresés alapvető sajátosságaira vonatkozó megállapításainkat.

A keresés fogalma, értelmezése általában

A *keresés köznapi fogalma* azt a tevékenységet takarja, amikor valakit, vagy valamit – akit, amit nem tudunk, hogy hol van, esetleg azt sem tudjuk, hogy létezik-e – meg akarunk találni. Az értelmező szótár szerint a keres kifejezés témánk szempontjából érdekes jelentése: "I. (Meg)találni igyekszik. 1. <Meghatározott személyt, dolgot, aki, amely elveszett, ill. akiről, amelyről nem tudja (pontosan), hogy hol van> megtalálni igyekszik. … 3. Találni igyekszik <ismeretlen személyt, dolgot, akiről, amiről még azt sem tudja, hogy létezik-e olyan minőségben, amilyenre szüksége van, amilyen neki megfelel>." [1, 858. o.] Ugyanilyen értelmezést hordoz az angol 'search' kifejezés is: "megpróbálni megtalálni valamit gondosan, alaposan végignézve, vagy más módon keresve"[2] [2], illetve "gondosan bejárni, vagy átnézni (helyet, területet, stb.) azért, hogy meg lehessen találni valami hiányzót, elveszettet"[3] [3, 1287. o.]

A keresés már a köznapi értelmezés szerint is két átfogó típusba sorolható aszerint, hogy egy már ismert dolgot, vagyis tulajdonképpen annak aktuális helyét, hollétét keressük, vagy adott tulajdonságoknak megfelelő dolgot/dolgokat keresünk, amilyen/amilyenek nem is biztos, hogy létezik/léteznek. Az elsőre példát nyújtanak a 'keresem a lányomat[a buliban]', 'keresem a mobil telefonomat[a lakásban]', vagy 'keresem a sógorom telefonszámát[, amit valahová felírtam]'. A második csoportba tartoznak a 'keresek egy angolul tudó hallgatót[a tancsoportban', 'keresek aknát [a menetvonalon]', 'keresek legfeljebb 56x45x25 cm méretű kis bőröndöt', vagy 'keresek szinonimákat a krumpli szóra'.

A keresés modellje, összetevői

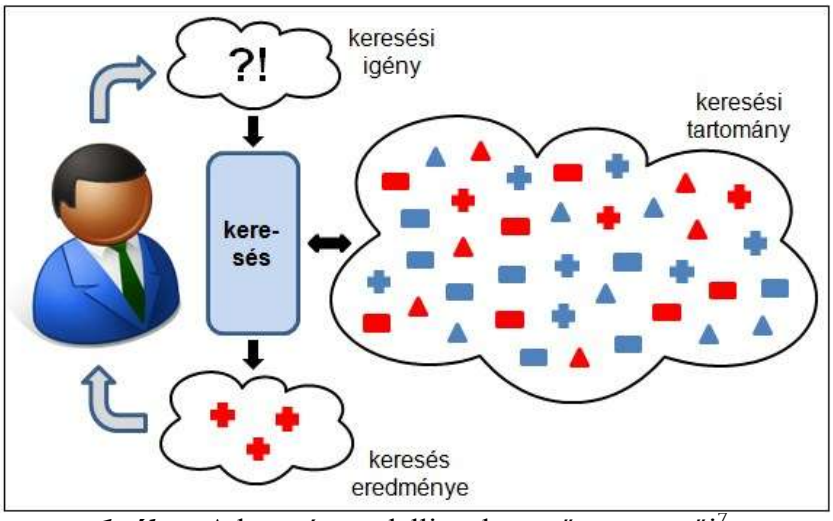
A *keresés fogalmának alapvető összetevői* közé a keresés alanya, a keresés tárgya, valamint a keresési tartomány tartozik, vagyis hogy ki keres, mit keres és hol keresi. Ezen túl a keresésnek mindig van, kell legyen célja, igénye, amelyben a kereső fél meghatározza, hogy mit kíván (meg)találni. Ennek a célnak, igénynek megfelelően indul a keresés, mint tevékenység, amelynek eredménye a keresett – a keresési célnak megfelelő – dolog (dolgok). A keresés részfeladatokra osztható, különböző szereplőkkel megosztható. A kereső fél a keresés során igénybe vehet különböző, a keresést támogató rendszereket, eszközöket, szolgáltatásokat. Végül a konkrét keresések előtt már végrehajthatóak egyes keresési részfeladatok, vagy a kereséseket megkönnyítő előkészítő feladatok.

A *keresés tárgya* lehet egy konkrét személy, tárgy, dolog (amiből egy van), vagy egy adott keresési célnak, feltétel-halmaznak megfelelő dolgok összessége. Az első esetben a keresés akkor ér véget, amikor az adott dolgot megtaláltuk,[4] a második esetben pedig akkor, ha egy adott körből kiválasztottuk a feltételeknek megfelelő dolgokat, vagy találtunk közülük meghatározott számút. A keresés tárgya egyben a *keresés eredménye* is. A keresés

[2] Try to find something by looking or otherwise seeking carefully and thoroughly.
[3] To go or look through (a place, area, etc.) carefully in order to find something missing or lost.
[4] Lehetséges egy időben több konkrét dolog keresése is, ami akkor eredményes, ha valamennyit megtaláltuk.

eredményessége igen/nem skálán leírható azzal, hogy megtaláltuk-e a keresett dolgot, illetve találtunk-e a keresési célnak megfelelő dolgo(ka)t. A második esetben a keresés eredményességének, illetve hatékonyságának fontos jellemzője a pontosság és a teljesség.[5] A pontosság azt mutatja meg, hogy a kiválasztott dolgok közül hány felel meg ténylegesen a keresési igényeknek, a teljesség pedig azt, hogy a feltételeknek megfelelő dolgokból mennyi került ténylegesen kiválasztásra. Ezekkel az információkereséshez kapcsolódóan a későbbiekben részletesebben is foglalkozunk.

A *keresési tartomány*, a keresés helye szintén alapvető összetevő, hiszen teljesen más feladat valamit/valakit megkeresni a szobában, a lakásban, az épületben, vagy a városban; a gyerekek, a felnőtt férfiak, vagy nők között; illetve a feljegyzéseim között, a szervezet iratai között, egy könyvtárban, vagy az Interneten. A keresési tartomány azon dolgok körét határozza meg, amelyek közül ki akarjuk választani a keresettet, vagy a feltételeknek megfelelőket. Ez – halmazok definiálásához hasonlóan – megtehető az átvizsgálandó tárgyak felsorolásával, vagy jellemzőik segítségével történő meghatározással. Anyagi objektumok esetében ilyen jellemző lehet pld. a térbeli (megadott területen történő) elhelyezkedés. Információk keresése esetében a keresési tartomány hagyományos, vagy elektronikus információhordozók meghatározott köre, vagy azok meghatározott részei. Egy adott tartományban történő keresés megvalósítható résztartományokon végrehajtott keresések segítségével, illetve különböző tartományok keresései egyesíthetőek összevont kereséssé[6].



1. ábra. A keresés modellje, alapvető összetevői[7]

A keresés legegyszerűbb formája (anyagi objektumok esetében) a keresési tartomány, egy adott terület személyes bejárása és a keresett dolgok személyes felkutatása, megvizsgálása. Már anyagi objektumok keresése esetében is szükség lehet *segédeszközök* igénybevételére a keresett dolgok felismeréséhez és a keresési feltételeknek történő megfelelés ellenőrzéséhez (pld. aknakeresés, lemerült akkumulátorok keresése, stb.). Számos esetben – pld. nem anyagi dolgok, vagy különösen elektronikus formában tárolt információk keresése esetében – a keresés már nem személyesen, hanem egy *kereső rendszer, keresési szolgáltatás* segítségével kerül végrehajtásra. A kereső rendszer, szolgáltatás számára meghatározott formában rendelkezésre kell bocsátani a keresési igényt, ennek alapján a rendszer, szolgáltatás 'bejárja' a keresési tartományt, hozzáfér az átnézendő dolgokhoz, összeveti azokat a keresési feltétellel, majd az ennek alapján kiválasztott dolgokat, illetve a rájuk vonatkozó információkat (helyüket, elérhetőségüket, stb.) rendelkezésre bocsátja.

[5] Precision, recall.
[6] Federated search.
[7] Saját szerkesztés.

Az információkeresés sajátosságai

Az információkeresés alapvető sajátossága, hogy tárgya – amelyet a keresés alanya megtalálni, megszerezni kíván – *információ*. Az információkeresés kiinduló pontja, motivációja mindig egy információigény. Értelmezésünk szerint az információ emberhez kötődő fogalom, szorosan kapcsolódik az ismeret(ek) és a tudás fogalmához. A *tudás* – viszonylag egységesen elfogadott értelmezés szerint – a közvetlen megismerés és kommunikáció útján megszerzett, valamint a gondolkodás eredményeként létrehozott ismeretek, illetve a gyakorlás útján kialakított speciális képességek (műveletek) összessége, az *ismeretek* pedig a megismerő tevékenység eredményei, a valós vagy elképzelt világ visszatükröződései az emberi tudatban. Ezekre építve az információ a világ egy megragadott aspektusának visszatükröződése, mentális reprezentációja az emberi tudatban. [4, 134-135. o.]

Mivel az információt az emberi tudatban létező dolognak tekintjük, azokat az emberi tudaton kívül kezelni, feldolgozni, velük különböző műveleteket végezni (tárolni, továbbítani, stb.) – így köztük keresni is – csak különböző *információreprezentációk* segítségével lehet.[8] Az információk feldolgozás, vagy későbbi felhasználás (megismerés, átvétel, stb.) céljából anyagi hordozón, egyezményes formában rögzíthetőek. A rögzített információk két nagy csoportba sorolhatóak. Az első csoportot a *szemléletes információreprezentációk* alkotják, amelyek az ember által érzékelhető környezeti hatásokat viszonylag valósághű formában reprodukálják, míg a második csoportba az *absztrakt információreprezentációk* tartoznak, amelyek az előbbiekkel szemben nem hordoznak a valósághoz hasonló jellemzőket, kapcsolatuk az általuk reprezentált dolgokkal önkényes (megállapodáson alapul). [5, 10-11. o.]

Az *adatok* értelmezésünk szerint az absztrakt külső információreprezentációk szinonimáját jelentik, szemben a korábbi és részben még ma is élő értelmezéssel, amely szerint csak a numerikus és logikai adatok, mennyiségi és igaz-hamis jellegű tulajdonságok absztrakciói sorolhatók ide. Napjainkban valójában már nem kérdés a szöveges, vagy multimédiás (rajz-, képi, hang, stb.) adatok létezése, illetve besorolása. Vagyis nincs olyan rögzített információreprezentáció, ami ne lenne adat.

Az adatok (információreprezentációk) strukturáltságuk alapján három nagy csoportba, a strukturált, a félig-strukturált és a nem strukturált adatok körébe sorolhatóak. A *strukturált adatok* közé a relációs adatbázisokban tárolt, tágabb értelemben a táblázatos adatokat sorolják. Ezek olyan adatok (adategyüttesek), amelyek típusokba (egyedtípusokba, osztályokba) sorolt gondolati egységek (egyedek) jellemzőit reprezentálják. Egy típusba tartozó egyedek ugyanazon jellemzőkkel kerülnek leírásra, az egyedeket leíró jellemzők típusa, formátuma és sorrendje meghatározott. Az adatok struktúráját leíró séma és a tényleges adatok egymástól elválnak, utóbbiak a séma nélkül kezelhetetlenek, értelmezhetetlenek.

A *nem strukturált adatok* közé olyan adatok tartoznak, amelyek tetszőleges típusúak lehetnek, nem követnek kötelezően egyetlen formai, sorrendi előírást, vagy szabályt. Ide tartoznak a szöveges adatok, illetve a szemléletes reprezentációk (álló- és mozgóképek, hangfelvételek). A *félig strukturált adatok* struktúrája a strukturált (táblázatos) adatokhoz képest jóval szabadabb, kötetlenebb. Ezek is típusokba sorolt egyedek jellemzőit reprezentálják, azonban az egyes egyedeket leíró jellemzők nem feltétlenül azonosak, a jellemzők típusa, formátuma, sorrendje eltérhet, létezésük nem feltétlenül kötelező. Ide tartoznak mindenekelőtt az XML formátumú dokumentumok és a különböző formalizált üzenetek.

Az információkeresés eredményeként kapott információreprezentációk rendeltetése, hogy értelmezésük segítségével a keresés alanya információhoz jusson, bővítse, módosítsa

[8] Az emberi emlékezetben történő "kereséssel" jelen publikációban nem foglalkozunk.

ismereteit. Ez viszont igaz minden információszerzésre, információfeldolgozásra is, így az információkeresést – bár nem mindig könnyű – el kell határolni az információ előállítástól, amely egy információigényt a keresés eredményeinek további feldolgozásával elégít ki. Ezzel azonban részletesebben majd a következő pontban foglalkozunk.

AZ INFORMATIKAI ESZKÖZÖKKEL SEGÍTETT INFORMÁCIÓKERESÉS ALAPJAI

Az információkeresés kérdései a tudományos vizsgálat tárgyaként intenzívebben a nagytömegű információkban történő kereséshez kapcsolódóan, a XX. század közepén jelentek meg. Természetesen a probléma ennél sokkal régebbi, a könyvtártudomány és előzményei már az ókortól foglalkoztak az írott információforrások gyűjtésének, rendszerezésének és rendelkezésre bocsátásának feladataival. Az igényelt információkat tartalmazó 'könyvtári dokumentumok' előkeresését már akkor különböző megoldások (katalógusok, osztályozási rendszerek, stb.) segítették. [6]

A számítógépek megjelenése más funkciók mellett megteremtette a lehetőségét az információk (pontosabban az azokat hordozó adatok) tárolásának és a tárolt adatok igény szerinti rendelkezésre bocsátásának is. Az információtechnológia fejlődése, a hálózatok megjelenése a hagyományos megoldásokhoz képest elképzelhetetlen mennyiségű információ technikai elérhetőségét teremtette meg. Azonban az információk ebben a tömegben a keresés megfelelő támogatása, megoldásai nélkül egyre kevésbé hasznosultak, egyre kevésbé jutottak el a lehetséges felhasználókig. Ennek megoldására jelentek meg az 1950-es évektől az információkereséshez kapcsolódó szakterületek.

A továbbiakban vizsgálatainkat az informatikai eszközök segítségével történő információkeresésre szűkítjük. Ennek keretében bemutatjuk az információtárolás és visszakeresés alapjait, főbb típusait; összegezzük, elemezzük az információkeresésnek az adatok strukturáltságáról függő sajátosságait; végül megfogalmazzuk az információkeresés általános fogalmát.

Információtárolás és visszakeresés

Az informatikai szakterület, az informatikai megoldások esetében az információkeresés fogalma, feladatai kezdettől szorosan összekapcsolódtak az információtárolás feladataival. Ennek alapját az képezte, hogy a keresési tartományt vagy maguk az informatikai rendszerekben tárolt információk, vagy a hagyományos információhordozók informatikai rendszerekben tárolt leírásai alkották. Az információkeresés így a tárolás és visszakeresés[9] részeként jelent meg.

Az információtárolás és visszakeresés kérdései a hozzáférhető információk körének ugrásszerű bővülésével és ezek pontos és gyors elérése egyre nehezebbé válásával az 1940-es évek második felében jelentek meg és az 1960-as évek első felében kerültek az érdeklődés homlokterébe. Egyre gyakoribbá vált, hogy egy adott információigényt kielégítő, releváns információ nem került felhasználásra, mert feltáratlan maradt, ami sok esetben újbóli előállításukhoz is vezetett. Az információkeresés, mint szélesebb körben elérhető lehetőség, az 1960-as években megjelent adatbázis-kezelő, illetve információ-visszakereső rendszerekhez kapcsolódott.

Az *adatbázisokban történő keresés* az információkeresés egy speciális típusa. *Adatbázis* alatt legáltalánosabb értelemben logikailag összefüggő, meghatározott szerkezetben tárolt adatok gyűjteményét értjük, amelyek a valóság meghatározott aspektusait reprezentálják és rendeltetésük ilyen adatok tárolása és rendelkezésre bocsátása különböző alkalmazások

[9] Storage and retrieval.

számára, ezeken keresztül pedig különböző felhasználói igények kielégítésére. Az *adatbázis-kezelő rendszerek* olyan szoftver rendszerek, amelyek lehetővé teszik adatoknak adatbázisokban történő tárolását, módosítását és elérését.

Az adatbázisok, illetve az azokat kezelő adatbázis-kezelő rendszerek osztályozhatóak a bennük tárolt adatok típusai és az adatbázis struktúrája szerint. Kezdetben a lehetséges adattípusok közé csak a numerikus, a logikai, a dátum-időpont és a korlátozott méretű karaktersorozat típusú adatok tartoztak, amelyek tárolására az 1970-es évek óta szolgál a relációs adatmodell. A lehetséges adattípusok köre később folyamatosan bővült a hosszabb szövegekkel, rajzokkal, multimédia adatokkal (álló- és mozgókép, hang), illetve számos új adatformátummal (pld. XML, vagy RDF dokumentumok).

Az adatbázisok kezelése során végrehajtható műveletek különböző csoportokba sorolhatóak, amelyek különböző utasítások segítségével valósíthatóak meg. Ezek közé tartoznak mindenekelőtt az adatstruktúrák létrehozását, módosítását biztosító adatleíró utasítások; az adatokkal végzett műveleteket (létrehozás, tárolás, módosítás, mozgatás, törlés, stb.) biztosító adatkezelő utasítások; az adatok visszakeresését biztosító lekérdező utasítások; valamint az adatbázis-műveletek végrehajtását irányító tranzakció-vezérlő utasítások. Napjainkban az adatbázisok kezelésének alapvető eszköze a fenti utasításokat magában foglaló SQL nyelv.[10]

Az adatbázis-kezelés egyik fő feladata a meghatározott feltételeknek megfelelő adatok megkeresése és rendelkezésre bocsátása. A keresés alapját a lekérdező nyelven megfogalmazott *adatbázis lekérdezés*[11] ('kérdés') – egy információ igény formális megfogalmazása – alkotja, amelyre az adatbázis-kezelő rendszer 'válaszként' a keresett, a lekérdezésben megfogalmazott körből a megadott feltételeknek megfelelő adatokat válogatja ki, bocsátja rendelkezésre.[12] Az adatbázis lekérdezések (köztük a relációs adatbázisok SQL lekérdezései) túlnyomórészt az információ visszakeresés ún. Boole modelljére épülnek: a feltételek adatelemekre vonatkozó elemi feltételekből felépülő logikai kifejezések. Relációs adatbázisok és az SQL lekérdezés esetében a keresés összekapcsolódhat az új információk (pld. összegző, tömörítő értékek) előállításával.

Az *információ visszakeresés*[13] fogalma az adatbázisokban történő kereséssel ('adat visszakereséssel') szembeállítva, lényegében dokumentumok visszakereséseként került megfogalmazásra. Eszerint az információ visszakeresés az információkeresés egy speciális típusa, "egy információigénynek megfelelő nem strukturált (általában szöveges) anyagok (általában dokumentumok) megtalálása nagy (általában számítógépeken tárolt) gyűjteményekben". [7, 1. o.]

A bibliográfiai adatbázisokban történő keresésből megszülető fogalom eredeti értelmezése szerint az információ visszakeresés célja, rendeltetése nem az információigényben meghatározott információk, hanem az ezen információkat tartalmazó *dokumentumok, információhordozók létezésére, hollétére vonatkozó információk rendelkezésre bocsátása*. "Egy információ visszakereső rendszer nem nyújt információt (nem változtatja meg a felhasználó ismereteit) a kérdés tárgyáról. Egyszerűen csak tájékoztatja a kéréséhez kapcsolódó dokumentumok létezéséről (vagy nem létezéséről) és hollétéről." [8, 1. o.] Ez így van a napjainkban már az informatikai rendszerekben tárolt dokumentumok között kereső rendszerek esetében is.

Az információ visszakeresés folyamata fő vonalaiban hasonló az adatbázis keresés folyamatával: az információ visszakereső rendszer egy lekérdezés formájában megfogalmazott információigényt összevet az elérhető dokumentumokról rendelkezésére álló

[10] Structured Query Language.
[11] Database query.
[12] A lekérdező nyelvek megjelenéséig a lekérdezések megvalósítására egyenként programot, eljárást kellett írni.
[13] Information retrieval.

információkkal és válaszként az igénynek megfelelő dokumentumok listáját (elérhetőségét) bocsátja rendelkezésre. A keresés részleteiben és eredményében azonban jelentős különbségek vannak.

Talán a legfontosabb a megfelelés kritériuma, ami az adatbázis-keresés pontos egyezésével szemben az információ visszakeresés esetében a *relevancia* (fontosság, tárgyhoz tartozás). A relevancia a kategorikus igen/nem helyett sorrendbe rendezett értékekkel jellemezhető, meghatározása jóval nehezebb feladat (amelyre különböző megoldások születtek és jelennek meg napjainkban is). Szorosan kapcsolódik ehhez a második különbség, a relevancia szerinti *rangsorolás*, ami az adatbázis-keresésből hiányzik. A keresés így részleges megfelelőség esetén is szolgáltat eredményt és azok közül kiemeli a 'legjobbakat'. Az információ visszakeresés történhet a dokumentumokat leíró adatok (pld. szerző, cím, kulcsszó, tárgyszó) segítségével, vagy a dokumentumok teljes tartalmának feldolgozásával.

Az eredményes és hatékony információ visszakeresés alapvető feltétele a keresési tartományt képező *dokumentumok előzetes feldolgozása*, amely lehet manuális és automatizált. Az feldolgozás eredményeként létrehozott leíró információk rendeltetése a (későbbi) keresés teljesítményének, gyorsaságának optimalizálása.[14] Ezen információk nélkül a keresés során végig kellene vizsgálni a keresési tartományt, az összes dokumentumot, ami nem hatékony, számos esetben nem is megvalósítható. Az előzetes feldolgozás feladatai az információ visszakereséshez kapcsolódóan számos különböző megnevezés alatt, részben sajátos tartalommal jelennek meg, pld. katalogizálás, indexelés, leíró adatok (metaadatok) előállítása.

Az információkeresés sajátosságai az adatok strukturáltsága szerint

Az informatikai rendszerekben tárolt információk keresése (visszakeresése) az alkalmazott információreprezentációk jellegétől függően különböző sajátosságokkal rendelkező megoldásokra épült, jelentőségük és lehetőségeik az idők során folyamatos változásokon ment keresztül. Ehhez kapcsolódóan az információreprezentációk (adatok) alapvető típusait strukturáltság szempontjából az előző pontban már ismertetett strukturált, strukturálatlan és félig-strukturált adatok, illetve ezek kombinációi alkotják.

A rendelkezésre álló háttértár-kapacitás és az információtárolás lehetőségeinek folyamatos bővülésével egyre több adat keletkezik strukturálatlan és félig strukturált formában. Egyes megállapítások szerint a szervezeti információk 80-85%-a tartozik ebbe a körbe, bár ezt valós felmérések pontosan még nem támasztották alá. Mindesetre a strukturálatlan és félig-strukturált adatok – melyek az adat-univerzum 'sötét anyagát' képezik – elérése, feldolgozása és így a bennük történő információkeresés jelentősége is folyamatosan nő.

A *strukturált adatokban történő keresés* alapját a már említett adatbázis-kezelés, illetve a táblázatkezelés lehetőségei képezik. Hagyományos megoldásai, lehetőségei régóta léteznek, hosszú évek tapasztalatai alapján alakultak ki és biztosítják a relációs adatbázisokban, (számoló)táblázatokban tárolt adatok közötti keresést. Az információ visszakeresés szakirodalma a keresés ezen típusát *adat visszakeresésnek* nevezi és lényegének az objektumok (egyedek) jellemzőire és kapcsolataira vonatkozó tényeket, elképzeléseket hordozó ún. propozicionális (állítás-alapú) információreprezentációk kinyerését tekinti. Az adatok strukturáltságából következően a keresés információigénye könnyebben, bár nem minden esetben könnyen alakítható formalizált lekérdezéssé. Ehhez azonban ismerni kell a keresési tartományt képező adatbázis, táblázat struktúráját (sémáját) és az egyes adatelemek formátumát, értékkészletét.

A strukturált adatokban történő keresés sajátossága, hogy az információigény megfogalmazása során felhasznált elemi keresési szempontok (feltételek) a táblázatos formában tárolt

[14] Az ún. indexelés az adatbázis-kezelésben is létező megoldás a keresések gyorsítására.

adatok egyes mezőinek vizsgálatával közvetlenül, minden előkészítő tevékenység nélkül ellenőrizhetőek. Ezen kívül ezen adatok tartalma és formátuma (talán csak a karaktersorozat mezőktől eltekintve) viszonylag jól meghatározott.

A *strukturálatlan adatokban történő keresés* elsőként a szöveges dokumentumok[15] visszakeresése formájában jelent meg. A vonatkozó szakirodalom információ (vissza)keresés alatt tulajdonképpen mindmáig – a kifejezés köznapi értelmezésétől eltérően – nagyobbrészt dokumentum (vissza)keresést és ezen belül *szöveges dokumentum (vissza)keresést* ért. [6, 7, 8, 9] A következő lépést az 1980-as évek közepén a szintén a strukturálatlan adatok közé sorolt képek, majd később a további multimédia formátumú információreprezentációk (hangfelvételek, videók) keresése képezte. A *kép/hangfelvétel/video keresés (multimédia információk keresése)*[16] adott információigénynek megfelelő képek/hangfelvételek/videók kiválasztása informatikai rendszerekben tárolt képek/hangfelvételek/videók (multimédia dokumentumok) gyűjteményéből.

A strukturálatlan adatok (szöveges és multimédia dokumentumok) keresésének legegyszerűbb módszere az előzőekben már említett, az egyes dokumentumokhoz kapcsolódó *leíró adatokban történő keresésre* épül. A dokumentumok 'természetes' leíró adatai (szöveges dokumentum szerzője, címe, hossza, keletkezés időpontja, stb.; kép készítője, címe, helyszíne, mérete, készítés időpontja, stb.; hangfelvétel, videofelvétel címe, szereplői, hossza, felvétel időpontja, stb.) a dokumentum tárolásával együtt kerülnek rögzítésre. Ezen adatok segítségével lényegében alkalmazhatóak a strukturált adatokban történő keresés módszerei.

A dokumentumok keresése történhet *kategorizáló adatok* (kulcsszavak, tárgyszavak, tartalmi kategóriák, stb.) segítségével, amelyeket megadhat a dokumentum szerzője, készítője, meghatározhatnak a dokumentum gyűjteményt kezelő szakemberek és meghatározhatóak a dokumentumok tartalmának automatizált elemzésével, feldolgozásával. A kategorizáló adatok meghatározhatóak szabadon, vagy előre rögzített kategória-listák felhasználásával. Az egyes kategóriák jelentéstartalmuk, kapcsolataik alapján fogalom-rendszerekbe (taxonómiák, tezauruszok, ontológiák) rendezhetőek, amelyek a keresés során is felhasználhatóak.

A keresés szöveges dokumentumok esetében lehetséges a dokumentumban előforduló valamennyi szó figyelembe vételével. A *teljes szöveges keresés* jellemzően az előforduló szavakra, kifejezésekre, esetleg azoknak a dokumentum adott részeiben (pld. cím, főszöveg, stb.) történő előfordulására, vagy egymáshoz viszonyított elhelyezkedésére vonatkozó feltételek alapján történik.[17] Kisebb dokumentum gyűjtemények esetében a feltétel ellenőrzése történhet a dokumentumok keresés során történő feldolgozásával, nagyobb számú dokumentum esetében azonban már teljes szöveges kereső indexek felhasználása szükséges.

A *strukturálatlan dokumentumok tartalom-alapú keresése* során hatékonyan használhatók fel az információkinyerés egyes korszerű módszerei. Ezek a lehetséges információigényekhez, keresési szempontokhoz igazodó módon határoznak meg, emelnek ki tartalmi elemeket, jellemzőket. Ide tartozik például szöveges dokumentumok esetében a névelem-felismerés[18], a képi és video dokumentumok esetében az arc/személy, illetve tárgy/helyszín felismerés, vagy a hangfelvételek esetében a hang/beszédfelismerés, beszélő/előadó felismerés, illetve zene felismerés.

A *félig strukturált adatokban történő keresés* jelentősége az Interneten elérhető ilyen típusú adatok (dokumentumok) mennyiségének ugrásszerű bővülésével nőtt meg. Ezen adatok a strukturált adatokhoz hasonlóan – mint azt korábban már megjegyeztük – egyedek (entitások) jellemzőit és kapcsolatait reprezentálják, azonban a táblázatos formátumhoz képest jóval rugalmasabb formában, jellemzően fa-struktúrába rendezetten. Az információkat a fa

[15] Pontosabban olyan dokumentumok, amelyek információkat alapvetően szöveges formában tartalmaznak.
[16] Image retrieval, audio retrieval, video retrieval, multimedia information retrieval.
[17] További teljes szöveges keresési eszközök: helyettesítő karakterek, reguláris kifejezések használata.
[18] Named entity recognition, named entity extraction.

egyes csomópontjai (az azokban tárolt adatok) és a csomópontok közötti kapcsolatok hordozzák. A keresés ilyen esetben az információigénynek megfelelő, releváns összefüggő dokumentum-részek megtalálására és rendelkezésre bocsátására irányul.

A félig strukturált adatokban történő keresés módszerei még kevésbé kimunkáltak, mint a két másik csoport esetében, azonban léteznek megoldások, lekérdező nyelvek XML dokumentumokban csomópontok kiválasztására (XPath) és adatok lekérdezésére (XQuery); relációs adatbázisokban XML adatokra is kiterjedő lekérdezésekre (SQL/XML); vagy RDF dokumentumokban tárolt adatok lekérdezésére (SPARQL[19]).

A szakirodalomban találkozhatunk a *részben strukturált adatok* fogalmával is, amelyek olyan adatok, amelyek nagyobb része strukturált, de vannak strukturálatlan összetevőik is. [10] Ez a típus a strukturált és a félig strukturált adatok között helyezkedik el. Ide sorolható számos olyan formalizált dokumentum, amelyben az alapvetően strukturált adatokat kiegészítik szabad szöveges, vagy multimédiás részek (pld. jelentések, jegyzőkönyvek, hibabejelentések, igénylések, orvosi vizsgálati lapok, stb.). Az ilyen dokumentumok tárolhatóak nem strukturált adattípusokat is tartalmazó relációs adatbázisokban és félig strukturált (XML) dokumentumok formájában is. Esetükben a keresés során a strukturált (vagy félig strukturált) és a nem strukturált adatokban történő keresés módszereinek kombinációjára van lehetőség.

Az információkeresés napjainkban *új kihívások* előtt áll. Mint azt érzékeltettük, a különböző strukturáltságú adatokban történő keresés sokszor jelentősen eltérő lehetőségeket biztosít, eltérő módszerekre, megoldásokra épül. Egyre növekszik azonban az igény az egymástól független, vagy elosztott *heterogén forrásokban történő keresés* lehetőségére, amely az egyes adattípusok sajátosságaira épülő, eltérő kereső módszerek összekapcsolását, harmonizálását igényli.

Egy másik kihívás az ún. "*Nagy Adatok*"[20] megjelenése. Az eredetileg a nagymennyiségű, minden korábbinál részletesebb, sokféle – elsősorban nem vagy félig strukturált – adatok kezelését és elemzését leíró kifejezés divatszóvá vált, egységes értelmezése még nem alakult ki. Témánk szempontjából fontos kiemelni, hogy a Nagy Adatnak nem egyedüli, bár kétségkívül alapvető kritériuma a méret. Legalább olyan fontos változatossága és keletkezésének gyorsasága is. [11, 4. o.] Nagy Adat alatt ma széles körben olyan nagy adathalmazt értünk, amely mérete és összetettsége miatt a rendelkezésre álló eszközökkel és módszerekkel nehezen kezelhető (tárolható, kereshető, elemezhető, vizualizálható, stb.). Mindez tehát új keresési módszereket is igényel.

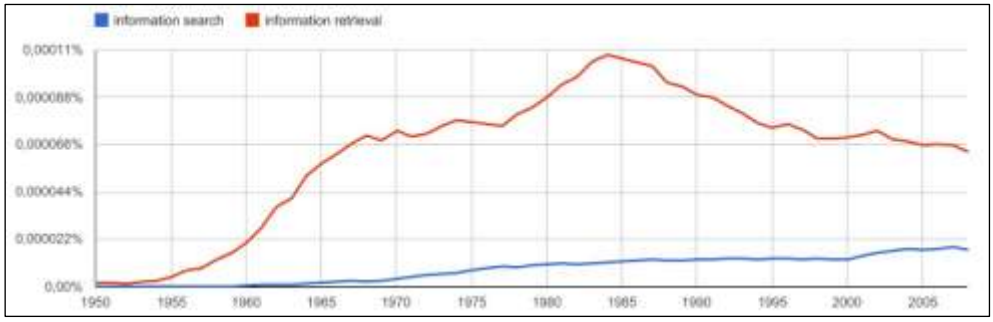
ÖSSZEGZÉS, AZ INFORMÁCIÓKERESÉS FOGALMA, ÉRTELMEZÉSE

Az eddigiekben elmondottak alapján indokoltnak látszik meghatározni az információk kereséséhez kapcsolódó alapfogalmat és annak tartalmát. Ezt megelőzően azonban röviden tekintsük át a 'visszakeresés' és 'információ visszakeresés' kifejezések használatát, értelmezését. A *visszakeresés* köznapi értelemben több jelentést is takar, de a magyar nyelvben közvetlenül az információkereséshez kapcsolódik: "<Adatot, vmely mutató v. utalás alapján> az eredeti helyen (könyvben, táblázatban) megkeres." [12, 481. o.] Angol nyelven az informatikához kapcsolódóan hasonló értelmezéssel találkozhatunk: " 9. (adat) helyének megállapítása és beolvasása tárolóból, pld. képernyőn történő megjelenítésre."[21] [3, 1644-1645. o.] Az információ visszakeresés kifejezés a Magyar Nyelv Értelmező Szótárában nem szerepel, angolul pedig meghatározott információk tárolt adatokból történő visszanyerését

[19] SPARQL Protocol and RDF Query Language.
[20] Big Data.
[21] To locate and read (data) from storage, as for display on a monitor.

értjük alatta: "3. Adatok szisztematikus tárolása és visszanyerése ügyiratból, cédulakatalógusból, vagy egy számítógép memóriájából."[22] [3, 980. o.]

Az *információkeresés és információ visszakeresés* kifejezések használata, mint az előző pontokban elmondottakból is érzékelhető, elsősorban elkülönülő szakterületek terminológiai gyakorlatára, értelmezésére épül. A relációs adatbázisokban történő keresés a széles körben elfogadott értelmezés szerint nem tartozik az információ visszakereséshez (mivel strukturált adatokhoz, relációs adatbázisokhoz kapcsolódik), pedig a köznapi értelmezés tartalmának teljesen megfelel. Mint az a következő ábrából is látható, az információ visszakeresés kifejezés jóval elterjedtebb, mint az információkeresés, bár használata az 1990-es évektől visszaesőben van.



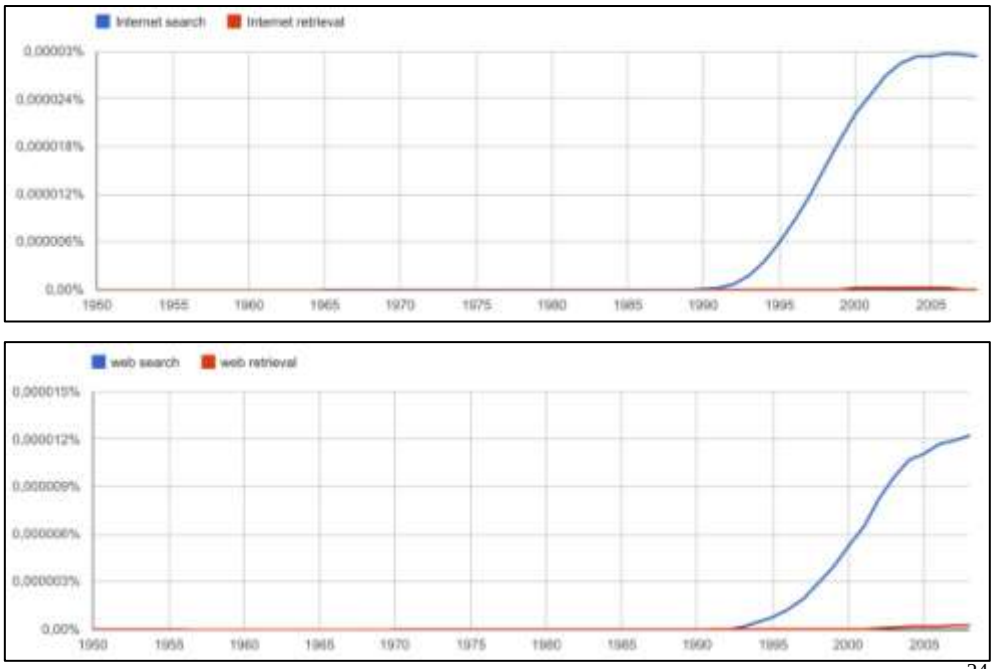
2. ábra. Információkeresés és visszakeresés kifejezések előfordulása 1950-2008 között[23]

A szakirodalomban találkozhatunk a két kifejezés más szempontok szerinti megkülönböztetésével is, amely szerint a keresés egy emberi tevékenység, amelynek során információ visszakereső rendszerek is felhasználásra kerülnek, az információ visszakeresés pedig a keresés technikai megvalósítását jelenti. Az előbbi az emberre és annak az információ visszakereső rendszerhez kapcsolódó viselkedésére, feladataira (keresési stratégia megválasztása, visszanyert információk relevanciájának megítélése), míg az utóbbi a visszakereső rendszer funkcióira, működésére fókuszál. [13, 1517. o.] Bár e megközelítésben van logika, témánk szempontjából nincs előnye és széles körben nem is terjedt el.

Az információkeresés (vagy információ visszakeresés) fogalmát megítélésünk szerint olyan tágan célszerű meghatározni, amely magában foglal minden tevékenységet, amelynek rendeltetése meglévő információreprezentációk (adatok), információforrások kiválasztásával és rendelkezésre bocsátásával hozzájárulni információigények kielégítéséhez. Ennek indokát abban látjuk, hogy a felhasználó szempontjából érdektelen, hogy az információigényeit milyen formátumban – relációs adatbázisokban, vagy szöveges visszakereső rendszerekben – kezelt, tárolt adatok segítségével elégítjük ki és az eltérő lehetőségektől, szolgáltatásoktól eltekintve érdektelenek a megvalósításbeli különbségek is. Ezen felül napjainkban már egyre inkább alapvető jelentőségűvé válik a számos különböző forrást felhasználó összevont keresés.

Ezt követően dönteni kell arról is, hogy a fenti tartalmat hordozó fogalom megnevezése *információkeresés, vagy visszakeresés* legyen. A 2. ábra tartalma alapján az információ visszakeresés lenne a logikus megoldás, azonban napjainkban az információk keresése (vagy visszakeresése) már jellemzően nem egy adott rendszerből, hanem az Interneten, illetve a World Wide Web-en történik. Ehhez kapcsolódóan viszont – mint azt a következő ábra is szemlélteti – a két kifejezés gyakoriságában már más a helyzet. Az angol terminológiában az ok talán abban rejlik, hogy nem az általunk tárolt információkat 'keressük vissza', hanem egy 'ismeretlen helyen' keresünk.

[22] The systematic storage ad recovery of data, as from a file, card catalog, or the memory bank of a computer.
[23] Forrás: Google Ngram Viewer.



3. ábra. Internet és web keresés kifejezések előfordulása 1950-2008 között[24]

A fentiekre építve az *információkeresés fogalma* alatt átfogó értelemben olyan tevékenységet javaslunk érteni, amely információreprezentációk (adatok) meghatározott köréből meghatározott információigény kielégítését segítő információreprezentáció(k), adat(ok) megtalálására, kiválasztására irányul. A definíció nem határozza meg, nem korlátozza a kezelt adatok strukturáltsági típusát, így magában foglalja valamennyi, korábban már említett keresési típust. Az információreprezentációk lehetnek táblázatos formába rendezett (relációs) adatok, dokumentumok teljes szövegű változatai, vagy dokumentumokat leíró tömör reprezentációk ('katalógus cédulák', publikáció referátumok, stb.). Az egyes keresési típusok, változatok a meghatározó sajátosságok beépítésével ebből a fenti fogalomból könnyen származtathatóak. A javasolt fogalom így megfelelő kiinduló alapot szolgáltat az információkereséshez és ezen belül a szemantikus kereséshez kapcsolódó kutatások számára.

Felhasznált irodalom

[1] A magyar nyelv értelmező szótára. Harmadik kötet. H-Kl. – Akadémiai Kiadó, Budapest, 1965.

[2] Oxford Dictionaries Online. – Oxford University Press, 2012. www.dcs.bbk.ac.uk/research/.../bbkcs-00-14.pdf, letöltve: 2012.12.05.

[3] Webster's Encyclopedic Unabridged Dictionary of the English Language. – Gramercy Books, New Work/Avenel, 1996.

[4] MUNK Sándor: Információs színtér, információs környezet, információs infrastruktúra. – Nemzetvédelmi Egyetemi Közlemények, 2002 (VI.) /2. (133-154.o.)

[5] MUNK Sándor: Katonai informatika I. A katonai informatika alapjai. Egyetemi jegyzet. – Zrínyi Miklós Nemzetvédelmi Egyetem, Budapest, 2003.

[6] SINGHAL, A.: Modern information retrieval: a brief overview. – IEEE Data Engineering Bulletin, 2001 (24.)/(4) (35-43. o.)

[24] Forrás: Google Ngram Viewer.

[7] MANNING, Christopher D.–RAGHAVAN, Prabhakar–SCHÜTZE, Hinrich: An introduction to information retrieval. – Cambridge University Press, 2009.

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

[9] BAEZA-YATES, R.–RIBERIO-NETO, B.: Modern information retrieval. – ACM Press, New York, 1999.

[10] KING, P. J. H.–POULOVASSILIS, A.: Enhancing database technology to better manage and exploit partially structured data. Research Report BBKCS-00-14. – Birkbeck University of London, 2000. www.dcs.bbk.ac.uk/research/techreps/2000/bbkcs-00-14.pdf, letöltve: 2012.12.07.

[11] Oracle Information Architecture: An Architect's Guide to Big Data. – Oracle Corporation, 2012.

[12] A magyar nyelv értelmező szótára. Hetedik kötet. U-Zs. – Akadémiai Kiadó, Budapest, 1966.

[13] JANSEN, Bernard J. – RIEH, Soo Yung: The seventeen theoretical constructs of information searching and retrieval. – Journal of the American Society for Information Science and Technology, 2010 (61)/8., 1517-1534. o.