nye szerint is szükséges meghatározni. Lényegében katalogizálási kérdésről van szó, olyan igénnyel, hogy a leirási szabályok elég rugalmasak és redundánsak legyenek az egyedi gyüjtési helyek kivánalmaihoz képest.

Másfelől minden dokumentumtipust jellemez informativ értéke, azaz a benne foglalt információtipusok (pl. statisztikai adatok, müszaki paraméterek, cégnevek stb.) amelyek elemzése szükséges a tájékoztatás számára, illetve amelyek a tároló-visszakereső rendszer ismérvrendszerét adják.

Általánosabb szinten egy-egy dokumentumtipus problémái – a szakirodalmi tájékoztatás szempontjából – ezekben a kérdésekben csucsosodnak ki. Bár Hegyi Nándor és Almásy László könyve ezekkel a kérdésekkel is foglalkozik, de valójában ezen a téren hiányérzetünk marad. Ennek leiró módszerük az oka. Szeretném megismételni, amit fentebb mondottam: az absztraktabb tárgyalásmódot számon lehet-e kérni a szerzőktől? Ez volt-e a feladatuk? Ha a könyvtári dokumentációs szakirodalomnak a feldolgozásra vonatkozó részét tekintjük, megállapithatjuk, hogy ez a fajta tárgyalás még a klasszikus könyvtári dokumentumok esetében is hiányzik. A katalogizálásnak még a könyvek, periodikumok esetében sincs komolyan kimunkált elmélete. Olyan elmélete, amely az információfeldolgozás számára produkálná azokat a tipikus adatstrukturákat, modelleket, amelyeket általánositott formában lehetne a továbbiakban kezelni és tárgyalni. Ugyanigy hiányosak azok a munkák, amelyek a visszakeresés kérdéseit szükebb empiria szintjén tul tárgyalják. Miért lenne éppen a vállalati irodalom kivétel, s miért kellene éppen a legfiatalabb dokumentumtipus esetében ezt megkövetelni?

A szerzőktől talán tulzott követelés lenne mindezt elvárni. S hogy mégis szóvátehető, nem ennek a konkrét könyvnek hibája, hanem az egész magyar könyvtári szakirodalomé.

HORVÁTH Tibor

The SMART Retrieval System. Experiment in Automatic Document Processing. E. Gerard Salton. New Jersey, 1971, Prentice-Hall, Inc.

Az automatizált SMART dokumentumvisszakereső rendszert a Harvard Egyetemen 1961 és 1964 közt tervezték meg. A rendszer a dokumentumokat ill. a keresési igényeket természetes nyelven fogadja el, a szövegek tartalmi elemzését teljesen automatikusan elvégzi, az elemzett dokumentumokat az elemzett kérdésekkel összeveti és kikeresi azokat a tárolt dokumentumokat, amelyek leginkább hasonlitanak a feltett kérdésre.

Más számitógépes visszakereső rendszerektől eltérően a SMART nem hagyatkozik a tárgyszavak vagy indexkifejezések manuális meghatározására, ugyanakkor nem alkalmazza kizárólagosan bizonyos szavak vagy kifejezések szövegbeni előfordulási gyakoriságát, hanem különféle segédeszközöket vesz igénybe a nyelvi elemzés elvégzése során.

Kiválóan alkalmas a SMART rendszer kisérletek elvégzésére. Ezt a lehetőséget számos alkalommal használták ki, s az eredmények jórésze az "Information Storage and Retrieval" report-sorozat köteteiben meg is jelent. Ezek (egybeszerkesztve, frissebb eredményekkel kiegészitve) alkotják e kötet gerincét.

A SMART PROJEKT - HELYZETJELENTÉS ÉS TERVEK

A SMART kisérletek során különféle dokumentumgyüjteményeket használtak fel. Jónéhány kisérletet párhuzamosan valamennyire elvégeztek s a legfontosabb megállapitások a gyüjteménytől függetlennek bizonyultak. Ugy tünik, hogy néhány eredmény minden müszaki jellegü dokumentumvisszakereső rendszerre igaz lehet. Ilyenek:

a) Automatikus tartalom-elemzéshez a referátum mindig hatékonyabb, mint a cim; a teljes szöveg viszont a referátumnál nem annyival hatékonyabb, hogy javasolni lehetne feltétlen előnyben részesitését.
b) A sulyozott tartalmi azonositók hatékonyabbak a sulyozatlanoknál; a dokumentum és a keresőkép hasonlóságának mérése a korrelációs függvény jobb, mint az átfedés vizsgálata.
c) A szókép normalizálása hatékonyabb, ha az alkalmazott szótár redundáns és nem tulzottan müszaki.
d) A szinonimák felismerését biztositó szótár használata jelentősen emeli a visszakeresés hatékonyságát.
e) A frázis-generáló módszerek (akár szótár, akár statisztikai asszociáció alapján) javitják a hatékonyságot, de a nyereség csak ritkán jelentős.
f) A nagymértékben hierarchikus eljárások nem igérnek jelentős eredményeket.
g) Az automatikus szövegfeldolgozás nem lényegesen gyengébb a manuális indexelésnél; nagy és heterogén gyüjtemények esetén határozottan előnyösebb.
h) A kipróbált eljárások minőségi sorrendje (nagyjából): referátumfeldolgozás frázis és szinonima felismeréssel; sulyozott szótő-összevetés és statisztikai asszociáció referátumok alapján; logikai szótő-összevetés sulyozás nélkül; cimfeldolgozás és összevetés átfedés alapján.
i) A felhasználók felőli visszacsatolás különösen ismétlődő keresések esetén igen hasznos.
j) A dokumentumok szétválogatásán alapuló cluster-eljárás lényegesen csökkenti a keresési időt és alig rontja a hatékonyságot.

A SMART rendszer továbbfejlesztésére vonatkozó tervek három fő csoportra oszthatók:

A. Szövegfeldolgozási kisérletek, ezen belül: automatizált szótárszerkesztés; bibliográfiai hivatkozások felhasználása a tartalom azonositására; az eredmények alkalmazhatósága nagyobb gyüjteményekre; összehasonlitás müködő rendszerekkel; összehasonlitás klasszikus tárgyi indexekkel; idegennyelvü input problémája.

B. Visszacsatolási és cluster-technikai kisérletek: optimális stratégia meghatározása a clusterek előállitására ill. a keresésre; visszacsatolási módszerek alkalmazása.

C. Real-time üzemmódu rendszerek: az on-line ("párbeszéd") müködésü rendszerre való áttérés.

A SMART RENDSZER ALKALMAZÁSA

A gyakorlati alkalmazás során a SMART visszakereső rendszert öt alapegységre bontották: 1. nyomtatott szövegek bevitele; 2. dokumentumok csoportositása keresési célokra: cluster-képzés; 3. a keresendő dokumentumok kiválasztása; 4. keresés a dokumentumcsoportokon belül; 5. a keresés kiértékelése. Az egyes tevékenységekkel a kötet tanulmányai részletesen foglalkoznak.

KIÉRTÉKELÉSI PARAMÉTEREK

Az automatikus indexelést alkalmazó dokumentumvisszakereső rendszerek a következő elemekkel jellemezhetők: természetes nyelvü dokumentumok halmaza; természetes nyelvü kérdések halmaza; indexnyelv; transzformáció a dokumentumok nyelvéről az indexnyelvre; transzformáció a kérdések nyelvéről az indexnyelvre; kereső függvény. Ezek alapján felépithető a rendszer modellje. A kiértékelési mutatók elmélete feltételezi ezek mellett a visszakeresett dokumentumok relevancia szerinti – legalább részleges – rendezését.

A visszakereső rendszer teljesitménye kiértékelésének elvi alapja egy összehasonlitás: a teljes gyüjteményből, adott kérdésre, mely tételeket keres ki a rendszer és melyeket keres(ne) ki a kérdést feltevő használó. A relevancia különféle fokai közti megkülönböztetés problémája a visszakeresett tételek rangsorolásával kerülhető meg. Még igy is két különböző, realizálható értékelési szempont érvényesülhet: vagy a visszakereső kérdéseket tekintjük elemeiknek (makroértékelés), vagy a releváns dokumentum és az irreleváns dokumentumok halmaza kapcsolatát (mikroértékelés). Az első kérdés-orientált, a második dokumentum-orientált szempont. A kétféle megközelités összehasonlitására a visszahivás pontosság görbét alkalmazták.

A teljesitménymutatók céljaik szerint három csoportba oszthatók: adott kisérleti szituáción belüli összehasonlitást szolgáló "belső" mutató; a kisérleti eredmények müködő rendszerekre való kiterjesztésének mérőeszközei. Emellett alkalmazhatók az értékelés eltérő szempontjai, s figyelembe veendők a megkivánt tulajdonságok. A rangsoroló rendszerekhez kialakithatók egy-adatos mutatók (log-pontosság, rang-visszahivás, normalizált visszahivás, normalizált pontosság), melyek függetlenek a rangsoron belül a releváns-irreleváns küszöb meghatározásától. Megadhatók a küszöbtől függő teljesitménygörbék is. A kétféle (makro és mikro) átlagolási eljárás eredményei a pontosságvisszahivás görbével vethetők össze, s az eredmények kiterjeszthetők eltérő általánosságu gyüjteményekre ill. a keresés során a clustertechnika alkalmazására is.

NYELVI ELEMZÉS

A nyelvi elemzés célja a természetes nyelvü dokumentum vagy kereső kérdés számára tartalmi azonositók meghatározása automatikus módszerekkel. Ennek során számos alapprobléma merül fel: a csak szintaktikai funkciót hordozó szavakat törölni kell; a szinonimákat fel kell ismerni; a hononimák eltérő jelentését azonositani kell; a szintaktikai parafrázisokat szintén; lehetőleg figyelembe kell venni a szövegen belüli indirekt utalásokat; fel kell tárni a szavak közti implicit kapcsolatokat; vigyázni kell a szavak jelentésének esetleges időbeli változására. Az elemzés elvégzéséhez négyféle szótár látszik szükségesnek: negativ szótár; tézaurusz; frázis-szótár; a kifejezések hierarchikus (fa-tipusu) elrendezése. A tézauruszokon belül is több tipust különböztetünk meg: szinonima-szótár és végződés-jegyzék; frázis-szótárak; fogalmi hierarchia. A tézaurusz szerkesztésében az automatizálás különféle mértékben érvényesithető.

A SMART rendszer a következő, a dokumentum-elemzés szempontjából fontos lehetőségekkel rendelkezik: szótövek meghatározása; fogalmak azonositása számmal; fogalmak hierarchikus elrendezése; statisztikai asszociációs módszerek; szintaktikai elemzés; statisztikai frázis-felismerés; kérdés-dokumentum összevetés. A rendszer biztositja a különféle feldolgozási eljárások értékelését és összehasonlitását. Ennek során elsősorban a következő problémákat vizsgálták: a rendelkezésre álló dokumentum (szöveg) hossza; összevetés-függvények és sulyozott kifejezések; hierarchikus kiterjesztés; manuális indexelés.

Néhány éve megindultak a kisérletek idegen (nem-angol) nyelvü dokumentumok automatikus feldolgozására is. A legtöbb eredmény eddig a német nyelvvel kapcsolatban született. Próbálkoztak többnyelvü tézaurusz kialakitásával is. Egy vegyes nyelvü dokumentumgyüjteményen lefolytatott kisérletsorozat alapján ugy tünik, hogy a SMART módszerei más nyelvekre is alkalmazhatók.

CLUSTER-GENERÁLÁS ÉS KERESÉS

A visszakeresési hatékonyság maximalizálása helyett, a felmerülő költségek miatt, az optimális megoldás látszik célszerünek. Ez pedig, főleg nagy és viszonylag heterogén adattár esetén, nem a teljes adattár, hanem csak annak bizonyos részei keresését teszi indokolttá. A SMART rendszer ezt a következőképpen oldja meg:

a) Minden egyes dokumentum tartalmi azonositóját az összes többivel összevetve a hasonló azonositóju (tartalmu) dokumentumokat csoportokba, clusterekbe vonja össze.
b) Minden clusterhez meghatároz egy reprezentativ elemet, az un. centroid vektort.
c) A keresést két lépésben végzi: először a keresőképet összeveti valamennyi centroid vektorral, majd a keresést már csak azokban a clusterekben végzi el, amelyek centroidja hasonlit a keresőképhez.

A cluster-képzéshez megfelelő algoritmus áll rendelkezésre, amely az eljárást automatizálhatóvá teszi. Az algoritmus kiértékeléséhez szükséges muta-

tók figyelembe veszik az "átnézett" clusterek számát, ez ugyanis befolyásolja a rendszer teljesitményét.

Egy másik (gyors) algoritmus a gyüjtemény tételeinek automatikus osztályozását célozza. Ezt két gyüjteményen is kipróbálták. Az eredmények kiértékelésére statisztikai szignifikancia-próbát alkalmaztak.

A cluster-képzés egy további módszere volna a keresőképből és az ehhez más módon visszakeresett dokumentumokból való kiindulás. Ezt is alkalmazták kisérletileg, és igy összehasonlitást tehettek a clusteres ill. a teljeskörü keresés teljesitménye között.

VISSZACSATOLÁS

Dokumentum-visszakereső rendszer hatékonyságának értékeléséhez el kell különiteni azokat a változókat, amelyek a rendszer viselkedését befolyásolják. Az e célból felállitott modell a visszakereső rendszer három alapfunkcióját határozta meg: indexelés, a keresőkép megfogalmazása és a keresőkép-dokumentum összevetés. Ezek közül a legnagyobb változatosság a keresőkép megfogalmazásában lehet (ezt ugyanis a használó végzi). Éppen ez adja a visszakeresés eredményére vonatkozó használói véleményt a visszacsatolás jelentőségét a keresőkép folyamatos javitásának folyamatában. E módositási eljárás algoritmizálása és automatizálása szintén megtörtént. A relevancia-visszacsatolás egyike a leghatékonyabb módszereknek, amelyeknek célja, hogy a felhasználót bevonja az információkeresés folyamatába.

A megfelelően felhasznált visszacsatolás egyuttal alapja lehet a felhasználó és a gép közti párbeszédnek, az interaktiv üzemmódnak, amelynek során a visszakeresést a használó folyamatosan ellenőrzi és irányitja. A visszacsatolást alkalmazó kereső eljárások hatékonyságának mérésére szolgáló mutatórendszert kidolgozták és kipróbálták.

Ez a lehetőség annál is inkább jelentős, mert az automatikus vagy félautomatikus információkereső rendszerek jelenlegi vagy a közeli jövőben várható teljesitménye lényegesen kisebb, mint azt sok potenciális használó képzeli. Ugy tünik, hogy a nyelvi elemzés vagy az adatszervezés terén várható változások sem idéznek majd elő gyökeres javulást. A legjelentősebb mérvü fejlődés az interkativ kereső technika alkalmazásától várható, ennek során ugyanis a gép szinte "kikényszeriti" a használóból igényeinek egyre pontosabb megfogalmazását. Az interaktiv keresés jól párositható a jelenlegi, viszonylag merev fileszervezés helyett egy dinamikus file alkalmazásával.

Érdekes kisérletet folytattak a visszacsatolás egy másik alkalmazási lehetőségére vonatkozólag is: ez nem egyszerüen módositja a keresőképet a visszajelzés eredménye alapján, hanem attól függően két vagy több uj keresőképet ad meg, melyek kissé különböznek az eredetitől, s a használó ezek közül választ. A kisérletek bizonyos foku teljesitmény-növekedésre mutatnak.

Egy további kisérleti eljárás a negativ relevancia visszacsatoláson alapul, amely akkor is müködik, ha az eredeti keresőkép egyetlen releváns dokumentumot sem eredményezett.

Biztató eredményeket mutat az a kisérlet is, amely a visszacsatolás alapján elhatárolt releváns ill. irreleváns dokumentumok közti szignifikáns hasonlóságból indul ki, s a megismételt-módositott keresőképet ennek alapján állitja elő. Az eljárás egyelőre további kipróbálás alatt áll.

A visszacsatolás alkalmazása nem szükségképpen merül ki a keresőkép iterativ módositásában. Alternativ lehetőség az indexelt dokumentumbázis (un. dokumentumtér) ill. a dokumentumok tartalmi leirása (un. dokumentumvektor) módositása, transzformációja is. A SMART rendszerben mindkettőre fejlesztettek ki algoritmust s alkalmazták kisérleti körülmények közt. Mindkét esetben a folyó, ismételt keresés mellett biztositották a korábbi keresési eredmények felhasználásának lehetőségét is.

MÜKÖDÉSI JELLEMZŐK

A SMART-kisérletsorozat – elsősorban távlati alkalmazás szempontjából – jelentős része az üzemszerüen müködő rendszerek tanulmányozása és összevetése a SMART eredményeivel, tapasztalataival. Ilyen vizsgálat volt az interaktiv keresés és az automatikus információ-kijelzést tartalmazó módszerek vizsgálata, a relevancia-itéletek és az információkereső rendszerek teljesitmény-értékelése közti kapcsolat elemzése, végül a manuális ill. az automatikus indexelési módszerek összehasonlitása. A SMART-rendszer alapelveinek, módszereinek gyakorlati alkalmazását azonban még jónéhány vizsgálatnak kell megelőznie, főleg a felmerülő költségek és a teljesitmény viszonyát illetően.

SÁRDY Péter

Fejér Megyei Könyvtáros. Fejér megye könyvtárosainak tájékoztatója. Kiadja a Vörösmarty Mihály Megyei Könyvtár, Székesfehérvár.

Könyvtári szakirodalmunk egyik mostohagyermeke a (ma már pontatlanul) "hálózati hiradó"-nak nevezett müfaj. Bevallottan "helyi érdekünek" készül, s ezt a könyvtáros-kollegák annyira komolyan veszik, hogy tudomásom szerint egyetlen intézményben sem található meg ezeknek a kiadványoknak többé-kevésbé teljes sorozata. Természetesen azt, hogy e szigoru elbirálás igazságos-e, csak a hiradók elemzése döntheti el.

Az un. hálózati hiradók nagyjából két csoportra oszthatók: a területi és a szakterületi hatókörüekre. Az előbbinek klasszikus példája a megyei hiradó, az utóbbié a szakmai (müszaki, mezőgazdasági, orvostudományi, stb.) hálózat hiradói. Miután pedig az élet gyakran produkál nehezen beskatulyázható jelenségeket, itt is találunk átmeneti kategóriákat. Példaként elég a hálózatnyi létszámmal müködő nemzeti könyvtárunk "házi" (de nemcsak házon belül érdekes) hiradójára vagy egyes egyetemi könyvtáraink hasonló kiadványaira utalni (pl. "Müszaki Egyetemi Könyvtáros"), mely utóbbiak a központi könyvtáron kivül a tanszéki, kari és egyéb hálózati könyvtárak problémáit is müsorra tüzik. A szakmai hálózati hiradók a nagy szakkönyvtári-információs szervezetek fejlő-