TÁJÉKOZTATÁSI MUNKA ÁLTALÁBAN

Az információk redundanciájának, relevanciájának, s használati értékének összefüggései az információkereső rendszerekben

A tudományos kutatás története és fejlődése során gyakran merülnek fel diszkrepanciák (ellentmondások, eltolódások) az elméleti megállapítások és az empirikus megfigyelések között, s *néha az elméleti megfontolások alternatívái között is találunk ellentéteket, amelyek nem ritkán új eredményekhez vezetnek.* Talán a társadalomtudományokban fordulnak elő leginkább ilyen esetek. Az információtudományban, amelynek orientációja elsősorban technológiai, s kevésbé elméleti, főként az információs rendszerek működését érintő problémakörök terén várható paradigmatikusnak tűnő ellentmondások felbukkanása az idetartozó tételekben.

Az operatív információs rendszerekről kialakult az elgondolásoknak, megállapításoknak egy olyan sora, amely egymással jól összeegyeztethető, koherens tételeket tartalmaz. Más nézetek, megfontolások többé-kevésbé könnyen beilleszthetők ezek közé, részben azért, mert az össze nem egyeztethető vonások nem durván szembetűnőek, részben pedig azért, mert még senki sem elemezte őket tüzetesebben.

Ebből a szempontból figyelmet érdemelnek azok az elgondolások, amelyek egyrészt *a dokumentumokban mutatkozó redundanciára, ismétlődésre, tartalmi átfedésre,* másrészt az *egyes dokumentumok értékelésére,* s az információs rendszer teljesítményének megítélésére vonatkoznak.

Elég széles körben elterjedt nézet, hogy a dokumentum-termelés tekintélyes része viszonylag triviális, felszínes, ismétlődő közléseket tartalmaz, amelyek valójában nem sokkal gyarapítják az irodalmat, s csak nagy munkaterhelést jelentenek, komolyabb megtérülés nélkül.

Ennek a nézetnek talán legnyíltabb szószólója R. SHAW, de ilyen vagy olyan formában képviselői közé tartozik ZIMAN, PRICE, SOERGEL, és sok más szakember is.

Az átfedés, a tartalmi (fogalmi) ismétlődés tehát (mint negatív kritérium) fontos szerepet játszik a bibliográfiai file-ok, input-stratégiák kialakításában, a dokumentum-gyüjtemények karbantartásában és más hasonló kérdésekben, – vagyis *„operatív doktrina"*-nak tekinthető.

E doktrina másik oldalán találhatjuk, mint egyértelműen pozitív kritériumot: *a relevancia-értéket,* amely megbízhatóan jelzi az információs rendszer teljesítményének minőségét. Ha a redundanciát (helyesebben a tartalmi ismétlődéseket) ugyanúgy megmérjük, értékeljük, mint a relevanciát, akkor elképzelhető, hogy mindkét kritériumot párhuzamosan kell vagy lehet használni az információs rendszerek teljesítményének vizsgálatára. S a relevancia és a fogalmi átfedés összehasonlító mérése, elemzése egyúttal lehetőséget ad arra is, hogy mélyebben megvilágítsuk azt a *bonyolult összefüggést, amely az információk redundanciája, relevanciája és tényleges felhasználói értéke között jelentkezik.*

Mindezeknek a problémáknak, feltevéseknek konkrét elemzése során az első lépés *a fogalmi átfedés mérési módszerének kidolgozása* volt, majd annak alapján megvizsgáltak egy 455 dokumentumból álló kisérleti gyűjteményt (mintát), amelynek tételei egy szűkebb tárgykörre (a légiforgalom irányítása számítógéppel) vonatkoztak. A mintát részben az ESRO rendszer NASA-STAR file-jában, részben a DIALOG rendszer INSPEC és NTIS file-jaiban való keresés alapján válogatták ki.

A fogalmi átfedések mérése az indexelésben használt tárgyszavak vizsgálata alapján történt, ami módot adott arra, hogy *rámutassanak az indexelés bizonytalanságaira,* sőt szeszélyeire is. A fogalmi ismétlődés és a relevancia viszonyának tisztázása érdekében tíz dokumentumnak egészen részletes szövegelemzését is elvégezték.

A vizsgálatok alapján az alábbi következtetéseket vonták le.

A fogalmi átfedés a dokumentumok között kimutatható, mérhető; a mérési eredmények koherens statisztikai tulajdonságokat jeleznek, s felhasználhatók más jellegzetességek előrejelzésére is, vagyis hasznosabb, értékesebb dokumentumok kiválasztásának elősegítésére egy meghatározott dokumentum-halmazból. Mégis – mindezek a jellegzetességek, mérési eredmények nem meggyőzőek, tehát *érvényük gyenge* ahhoz, hogy a gyakorlati alkalmazás tekintetében komolyabban figyelembe vehetnénk őket.

Ez a gyengeség két forrásból származik. Az egyik magában az irodalomban rejlik, mivel az irodalom tüzetes vizsgálata azt mutatja, hogy *a szélsőséges, teljes redundancia valójában igen ritkán fordul elő!* Ez a megállapítás ellentmond az információs szakemberek széles körében elterjedt hiedelemnek, bár egy szűk szakterületről származó kisérleti dokumentumállomány vizsgálatán alapszik! S ha azt látjuk, hogy egy egészen speciális témakör több mint 400 dokumentumában a kirívó önismétlésnek mindössze két esetét találjuk, s nagyobb fokú redundanciát mutató más dokumentumcsoportokra példák nem fordulnak elő, – akkor az

előbbi megállapítást meggyőzőnek kell tartani. (Annál is inkább, mert a vizsgálatokat végző kutatók maguk is azt várták, hogy a fogalmi átfedések sokkal nagyobb mértékűek lesznek!)

Az indexelésben rejlik a másik ok, ami miatt az átfedések mérésének eredményei gyakorlatilag alig alkalmazhatók. Az elemzések során megállapították, hogy egymással tartalmilag csaknem megegyező, vagy *egymáshoz hasonló dokumentumok viszonylag kevés közös tárgyszót tartalmaztak*, tehát indexelésük eltérő volt, – s ha ez így van, akkor ezen az alapon az átfedések mérése teljesen bizonytalan vállalkozás.

Az indexelés következetlenségei közismertek. A jelenlegi indexelési eljárások nem olyanok, hogy megbízhatóan támogathatnák az információkereső rendszerek részletekbe menő válogatási folyamatait vagy más ilyen gyakorlati „beavatkozásait".

Mindebből (első pillanatra) azt a tanulságot vonhatnánk le, hogy alaposan meg kell javítani az indexelés szinvonalát, jóval magasabbra kell emelni a mércét. De nem valószínű, hogy ez megérné a fáradságot. Az a költség ugyanis, amit az indexelés szinvonalának jelentős megjavítására kellene fordítani, valószinűleg messze meghaladná azt a hasznot, amit azzal érnénk el, hogy pontosan meg tudjuk állapítani a fogalmi átfedéseket a kikeresett dokumentumok egy csoportján belül. Ez részben azért van így, mert – amint erre már utalás történt – a redundancia, a tartalmilag azonos vagy közel azonos dokumentumok előfordulása kétségtelenül igen ritka.

A tartalmi, *fogalmi átfedések jellegzetes esetei* a következőkben foglalhatók össze:

a nyílt önismétlés (plagizálás) egy-két egészen ritka esete;

a lényeges átfedés néhány esete, ami több tényezőnek tulajdonítható, pl. annak, hogy több kommunikációs csatornát használnak fel ugyanannak a kutatási eredménynek közlésére, kissé megváltoztatott tartalommal (kongresszusi előadás, folyóiratcikkek a fő tárgykört, s a kapcsolódó tárgyköröket érintő szaklapokban stb.);

az átfedések főként *a dokumentumokat alkotó fejezetek, részletek szintjén* nyilvánulnak meg, de azok a dokumentumok, amelyeknek néhány fejezete közös tartalmi mondanivalót hordoz, más fejezetek tartalmában jelentősen eltérhetnek egymástól. (A dokumentumok szövegének mély elemzése egyébként azt jelezte, hogy a tartalmi ismétlődést mutató dokumentumokban is volt legalább egy olyan gondolat, amely csak ott szerepelt, tehát a felhasználó szempontjából releváns volt!)

Ezek a szempontok arra engednek következtetni, hogy a vizsgált probléma valószínűleg nem oldható meg a szokásos dokumentum-elemzés szintjén, hanem csak *mélyrehatóbb szövegelemzéssel*. Ehhez olyan jellegű teljes szövegfeldolgozásra van szükség, mint amilyet SOERGEL vetett fel (a *Dokumentation und Organisation des Wissens* című művében).

Mindezekből a jövőt illetően *két irányzat* körvonalazódik elénk világosabban. Az egyik arra vállalkozhatna, hogy *megerősítse azokat a kisérleti általánosításokat*, amelyeket eddig az egymást keresztező tárgykörök összehasonlítása útján állapítottak meg. Röviden arról van szó, hogy ami érvényes a műszaki irodalom egy szűk témakörére, az nem biztos, hogy érvényes a műszaki irodalom egyéb tárgyköreire, vagy az orvostudományra, s az alaptudományokra. Ezt a kérdést kisérletileg kell tovább vizsgálni.

A munkálatok másik irányát a *teljes szövegelemzés* jelentené – addig a határig, ameddig még gazdaságos. Ennek a tevékenységnek az eredménye sokkal több lenne, mint a fogalmi átfedés mérési rendszerének kidolgozása. Csaknem bizonyos, hogy megadná az alapot *a ténykereső (fact retrieval) rendszerek működéséhez*, amelyeket sok dokumentalista úgy tekint mint szakmájának csaknem végső feladatát.

/CLEVERDON, C. W – KISS, J. S.: Redundancy, relevance, and value to the user in the outputs of information retrieval systems = The Journal of Documentation, 32. köt. 3. sz. 1976. p. 159–173./

(Györe Pál)