

AZ INFORMÁCIÓKERESÉS MINT PRÓBÁLGATÁSOK ÚTJÁN VÉGBEMENŐ KÖZELÍTÉSI FOLYAMAT

Don R. SWANSON

A szerző *Information Retrieval as a Trial-and-Error Process* c. cikkének szemelvényeit (*Library Quarterly*, 47. vol. 1977. 2. no. 128–148.p.) KÖVENDI Dénes fordította.

Az információkeresési rendszereket általában annak alapján értékelik, hogy mennyiben képesek kiválasztani a releváns dokumentumokat és kiszűrni az irrelevánsakat. Magának a relevancia fogalmának azonban eddig nem szenteltek kellő figyelmet. Az alábbiakban az utóbbi évtized néhány fontos visszakeresési vizsgálata kapcsán megvizsgáljuk, hogy milyen következményekkel járhat ez a mulasztás. Az elemzett kísérletek mindegyikében egy dokumentumgyűjteményt vetettek egybe egy tesztkérdés-csoporttal. A keresést egy vagy több ellenőrzendő indexelési eljárás segítette. Valamennyi kísérlet lényege annak a meghatározása volt, hogy az egyes keresésekben hány releváns dokumentum maradt feltáratlan.

A Cranfield II. kísérletben¹ a tulajdonképpeni visszakeresési tesztek előtt igyekeztek a kísérleti gyűjtemények valamennyi releváns dokumentumát azonosítani, közvetlenül megvizsgálva mindegyik dokumentumot mindegyik kérdés szempontjából. A jelen cikk szerzője bírálatában² rámutatott arra, hogy ez a szűrési eljárás feltáratlanul hagyja a releváns dokumentumoknak kevesebb mint 90%-át. Igyekezett is némi világot vetni ennek a titokzatos jelenségnek az okaira: a relevancia fogalmával kapcsolatos problémáknak bizonyára volt szerepük benne.

Lancaster MEDLARS-tesztje³ olyan módszerrel határozta meg a keresés teljességét (a visszakeresett releváns dokumentumok arányát az összes releváns dokumentumhoz képest), amely félrevezető lehet: a visszakeresett dokumentumok számát azon dokumentumok csoportjához viszonyította, amelyeket maga a kereső előzőleg relevánsnak minősített. Csakhogy részben már ez a csoport is olyan dokumentumokból állt össze, amelyeket előzőleg mások – tehát nem a kereső – kiválogattak mint esetleg relevánsakat. Az ilyen dokumentumokat könnyebb relevánsnak elismerni – s ennél fogva talán könnyebb indexelni és visszakeresni is –, mint azokat a releváns dokumentumokat, amelyek nem mentek keresztül ilyen előzetes szűrőn.

Salton néhány száz dokumentumból álló kísérleti gyűjtemény alapján hasonlította össze a MEDLARS-t egy teljesen automatikus rendszerrel, referátumok szövegének visszakeresése terén. Azt állapította meg, hogy az ilyen automatizált rendszerek pótolni tudják a manuális indexelésre épülőket.^{4,5} Csakhogy egyelőre semmi sem bizonyítja, hogy

Salton következtetése nagy dokumentumgyűjteményekre is érvényes. Lancaster vizsgálataiból (függetlenül a fent említett kifogásainktól) az következik, hogy a MEDLARS rendszer képes kiválasztani néhány tucat releváns dokumentumot mintegy 800 000 irreleváns közül; a Saltonéhoz hasonló kis kísérleti gyűjtemények alapján egyetlen konkurrens rendszerről sem lehet megállapítani, hogy képes ilyen nagymérvű diszkriminációra.

Mindezekben a kísérletekben látszólag kizárólag a kereső megítélése alapján határozták meg, hogy a dokumentum releváns-e a mérés szempontjából. A relevancia megítélését azonban voltaképpen mindegyik kísérletben áthárították valaki másra, előzetes szűrés formájában. Ez a következetlenség ajánlatossá teszi a relevancia fogalmának behatóbb vizsgálatát.

A cikkíró a relevanciának két egészen különböző jelentését vizsgálja részletesebben. A relevanciaítélet lehet egyrészt a kereső alkotó, szubjektív gondolati aktusa, amely azt fejezi ki, hogy a dokumentum kielégít-e valamilyen, az ő kérésében foglalt információszükségletet. Másrészt tükrözheti a keresőnek azt a megállapítását is, hogy az adott dokumentum „ugyanarra a témára” vonatkozik, mint az ő kérése. Ez nem jelenti szükségszerűen azt is, hogy kielégíti a kérése alapjául szolgáló információszükségletet. [Lényegében a relevancia és a pertinencia különbsége ez – utóbbi kifejezést a cikkíró nem használja. (–a ford.)] Az utóbbi esetben a kereső relevanciaítéletét bizonyos mértékben előre megfogalmazhatja más is (az első esetben nem). Az információkeresési kísérletek e kétféle értelmezési szempontjának a megkülönböztetése, amint látni fogjuk, némi világot vethet az említett nehézségek forrásaira.

Köztudomású, hogy az információ visszakeresése voltaképpen próbálgatások útján végbemenő közelítés. Ajánlatos ezt a szempontot rendszeresebben érvényesíteni. A keresési igény ugyanis semmi egyéb, mint azoknak a sajtóságoknak a feltételezése, amelyekkel az általunk óhajtott dokumentumnak rendelkeznie kell. A rendszer választásán alapuló próbálgatás arra használjuk fel, hogy módosítsuk eredeti feltevésünket, próbálkozás végett. Modellként vehetjük Popper leírását a tudományos ismeretek megszerzéséről és növekedéséről.⁶ Az információkeresési rendszert ezek szerint úgy kell kialakítani, hogy a maximális rugalmasságot és a gyors választ biztosítsa magában a próbálgatásos megközelítési folyamatban. Három specifikus gyakorlati intézkedést javasolhatunk ennek a rugalmasságnak az elérése érdekében.

Az első a használó és az információkeresési rendszer közötti párbeszéd kiterjesztését és tökéletesítését célozza. A számítógépes interaktív rendszerek másodpercek alatt közlik a dokumentum bibliográfiai adatait (szerző, cím), magához a dokumentumhoz azonban esetleg csak napok, sőt hetek múlva juthat hozzá a használó. Szükség volna hát egy közbülső szolgáltatásra: olyan rendszerre, amely a jelen autoreferátumhoz hasonló részletes referátumokat percek alatt bocsátana a használó rendelkezésére. Felbecsülhetetlen segítséget jelentene a kérés pontosításához és módosításához az a lehetőség, hogy gyorsan kiválaszthatjuk vagy elvethetjük az egyes címekeket, rezüméket, referátumokat, jellemző szemelvényeket, s végül magukat a cikkeket.

A második javaslat a könyvtárak „böngészhetőségére” vonatkozik. A polcokon való böngészés az elemi példája a próbálgatásos megközelítésre épülő információkeresésnek, csak hogy a könyvtárak ésszerűtlen kölcsönzési politikája akadályozza ezt az eljárást.

A jelenlegi gyakorlat ugyanis mintha épp azt akarná biztosítani, hogy a leggyakrabban használt, feltehetőleg tehát legértékesebb anyagot ne lehessen megtalálni a polcon. Javíthatjuk a könyvtár böngészhetőségét, tehát hasznosságát, ha a kölcsönzési határidőt a könyv használatának várható gyakoriságától függően állapítjuk meg. Bizonyos határokon belül tehát csak rövidebb időre szabad kikölcsönözni a gyakran használt műveket. Ezt a módszert nem rossz eredménnyel próbálta ki a University of Lancaster könyvtára.

A harmadik javaslat a hivatkozásokat kívánja használhatóbbá tenni. A szerzők minden egyes tételről világosan jelezzék, hogy miért utalnak rá. Így javíthatjuk a próbálgatásos megközelítésen alapuló keresés lehetőségét minden olyan folyamatban, amely egy tanulmány hivatkozásaihoz kapcsolódik. A jelen cikk végén található hivatkozások mutatják e javaslat megvalósításának egyik formáját: mindegyik idézett műről megemlítjük, hogy milyen speciális köze van tanulmányunk gondolatmenetéhez.

[Alább az első hat hivatkozást közöljük, amely magának az autoreferátumnak a szövegéhez fűződik. A szerző csillaggal emeli ki azokat, amelyek tanulmánya szempontjából központi jelentőségűek: e hat utalás mindegyike csillagot kapott, ezt mi külön nem jelezzük. (– a ford.)]

HIVATKOZÁSOK

1. CLEVERDON, Cyril – MILLS, J. – KEEN, Michael: ASLIB Cranfield Research Project; factors determining the performance of indexing systems. Vol.1. Test design. College of Aeronautics, Cranfield, Bedfordshire. 1966.
Ez a munka egyike annak a három nagy, információkeresési rendszerekre vonatkozó kísérleti vizsgálatnak, amelyekről a fentiekben megállapítottuk, hogy bizonyos korlátaik vagy következetlenségeik feltehetően olyan nehézségekből erednek, amelyek a relevancia fogalmára épülő minden értékelési módszerre jellemzők.
2. SWANSON, Don R.: Some unexplained aspects of the Cranfield tests of indexing performance factors. = *Library Quarterly*. 41.vol. 1971. 223–228. p.
A cikk bírálja a Cranfield II. vizsgálatnak azt a módszerét, hogy a visszakeresési kísérlet előtt mindegyik kérés szempontjából minden releváns dokumentumot azonosítanak. Magának a kísérletnek az adatai alapján mutatja ki, hogy így a dokumentum és a kérés közötti releváns kapcsolatok mintegy 90%-a elvész. Sohasem tárták fel ennek a hatalmas információvesztésnek a titkát. A szerző feltevése szerint elsősorban a relevancia fogalmával kapcsolatos alapvető nehézségekről van itt szó.
3. LANCASTER, F. W.: Evaluation of the MEDLARS Demand Search Service. Library of Medicine, U. S. Department of Health, Education, and Welfare. Washington, D. C. 1968.
Felhívja a figyelmet arra, milyen nagy különbségek vannak a különböző kérésekre kapott visszakeresési eredmények között (lásd a 129. p. grafikonját). Ezek a különbségek véleményünk szerint a relevancia fogalmával kapcsolatos alapvető nehézségekből erednek.
4. SALTON, Gerard: A comparison between manual and automatic indexing methods. = *American Documentation*. 20.vol. 1969. 61–71. p.
Ez a munka is egyike annak a három nagy, információkeresési rendszerekre vonatkozó kísérleti vizsgálatnak, amelyekről a fentiekben megállapítottuk, hogy bizonyos korlátaik feltehetően olyan nehézségekből erednek, amelyek a relevancia fogalmára épülő minden értékelési módszerre jellemzők. Ez a kísérlet felhasználja Lancasternek a teljességre vonatkozó adatait, így erre

is érvényesek amannak a korlátai. A kísérlet valóban képet ad a SMART és a MEDLARS rendszer teljesítményének viszonyáról kis dokumentumgyűjtemények esetében, semmiféle következtetést nem lehet viszont levonni belőle a SMART rendszernek igen nagy gyűjteményekre való alkalmazásáról.

5. SALTON, Gerard: A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). = J. of the American Society for Information Science. 23. vol. 1972. 75–84. p.

Lásd a 4. jegyzetben található megjegyzéseket. Ismét megállapítja, hogy a teljesen automatizált dokumentumrendszerek pótolni tudják a manuális rendszereket. Az újabb összehasonlítás egy 450 referátumból álló gyűjteményre épül, ez pedig messze kisebb annál, amely bármely következtetést megengedne arra nézve, milyen nagy a „zaj” (a visszakeresett irreleváns dokumentumok aránya) nagy gyűjtemények esetében. Megjegyzéseket fűzök ahhoz, hogy e kísérletben a relevanciátétel a keresőről más személyekre hárították át.

6. POPPER, Karl R.: Conjectures and refutations. The growth of scientific knowledge. Harper Torchbooks, New York. 1968.

Popper filozófiája, kivált az 1. fejezetben (33–59.p.) kifejtett formájában, ösztönzést ad arra, hogy az információkeresést próbálgatásos megközelítésnek tekintsük, így alapul szolgál e tanulmány fő gondolatához.

* * *

Nemcsak tartalmi jelentősége miatt közöljük e tanulmány autoreferátumának teljes fordítását, hanem azért is, mert illusztrálja a szerző egyik érdekes javaslatát: az információkeresési folyamat eredményesebbé tételére a bibliográfiai adatok és a teljes dokumentum rendelkezésre bocsátása közötti közbülső láncszemként rendszeresíteni kellene a (jelen esetben a szöveg 1/8-át kitevő) részletes referátum rendelkezésre bocsátását is. Hasonló újító módszertani javaslat indokolja a szerző hivatkozási módszerének bemutatását.