

Információs adatbázis hatékonyságának becslése

Az adatbázisokból végzett manuális vagy számítógépes információkeresés hatékonyságának új halmazelméleti eljárásával a megadott képletekkel a keresés teljességi és pontossági mutatója, ill. az adatbázis relevanciamutatója becsülhető meg.

Napjainkban egyre karakteresebb a gazdaság piaci orientációja, a "mi éri meg és mennyiért" szemlélet válik uralkodóvá. Ebben a környezetben fontos lenne tudnunk – felhasználóknak és szolgáltatóknak egyaránt –, vajon egy adott információs adatbázis milyen hatékonysággal használható, legyen az akár hagyományos, akár online módon elérhető számítógépes adatbázis.

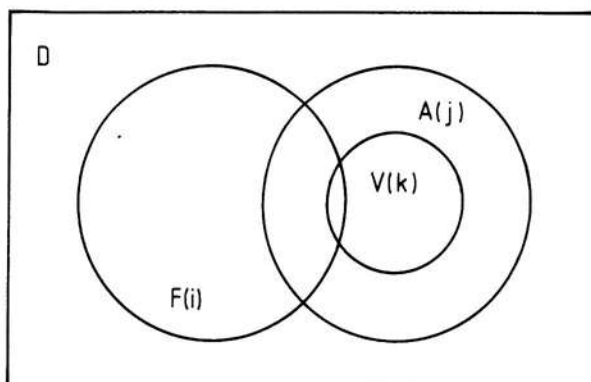
Az alábbiakban közreadott módszer egy lehetőség az információs szolgáltatás várható hatékonyságának, potenciális hasznának, azaz használati értékének becslésére. Érdekes felsorolni azokat a tényezőket, amelyek egy adott információs áru használati értékét befolyásolják.

- ▶ *nóvumfok* (újszerűség),
- ▶ *aktualitás* (időszerűség),
- ▶ *relevancia* (adekvátság),
- ▶ *emészthetőség* (érthetőség),
- ▶ *kapcsolhatóság* (illeszthetőség már ismert rendszerekhez, elméletekhez).

Egy információs adatbázisból történő szolgáltatás minőségének két fontos, jól ismert mutatója a keresőkérdésre kapott találatok száma és azok relevanciája. A találatok száma a keresőkérdés bővítésével növekszik, illetve a keresőkérdés szűkülésével csökken. Releváns az a találat, ill. az a dokumentum, amely tartalmilag választ jelent a feltett keresőkérdésre, vagyis amely számunkra egy adott szempontból fontos, érdekes információt tartalmaz. Az ettől eltérő válaszok jelentik az információs zajt, adják az érdektelen dokumentumot.

Az információs adatbázis használatának, a keresés hatékonyságának ismert mutatói:

- ▶ *teljesség* (recall = R), amely azt mutatja meg, hogy az adatbázis összes releváns dokumentumából mennyit találtunk meg;
 - ▶ *pontosság* (precision = P), amely a megtalált releváns dokumentumokat viszonyítja az összes megtalált dokumentum számához.
- Halmazelméleti jelöléssel:



ahol:

- D – az összes információs áru halmaza;
- $F(i)$ – az i -edik téma szempontjából fontos, releváns információs áruk halmaza;
- $A(j)$ – a j -edik adatbázis információs áruinak halmaza;
- $V(k)$ – a k -edik visszakeresésnél (szolgáltatásnál) megtalált információs áruk halmaza.

$$F(i), A(j) \in D \\ \text{és } V(k) \in A(j).$$

Teljesség:

$$R = \frac{V(k) \cap F(i)}{A(j) \cap F(i)},$$

azaz a k -edik visszakeresésnél megtalált $|V(k)|$ az i -edik téma $|F(i)|$ szempontjából releváns tételeinek (a képlet számlálója), valamint a j -edik adatbázis $|A(j)|$ az i -edik téma $|F(i)|$ szempontjából releváns tételeinek (nevező) hányadosa.

Pontosság:

$$P = \frac{V(k) \cap F(i)}{V(k)},$$

azaz a k -adik visszakeresésnél megtalált $|V(k)|$ az i -edik téma $|F(i)|$ szempontból releváns tételeinek és az összes megtalált tétel $|V(k)|$ aránya.

A teljesség és pontosság egymás rovására érvényesülő, komplementer jellegű mutatók.

Ha a fenti mutatók felhasználásával számításokat akarunk végezni egy adott szolgáltatás (keresés) hatékonyságának jellemzésére, a következő problémával kerülünk szembe. Nem tudjuk, mennyi lehet az adott adatbázisban a kérdéses szempontból releváns tételek száma összesen, tehát az $A(j) \cap F(i)$, a teljesség képletében leírt nevező nagysága. A pontos számbavétel nem járható út, mert ugyan ki vállalkozna arra, hogy az adatbázis összes tételét egyenként, külön-külön értékelje.

Hasonló gond a pontosság kiszámításában nem jelentkezik, hiszen egy adott keresőkérdésre kapott válaszok megszámlálhatók, és a tételek értékelése alapján a köztük lévő releváns információk száma is pontosan meghatározható.

Egy adatbázis használati értékét, potenciális hasznát mérhetnénk az adatbázis *relevanciamutatójával*:

$$AR = \frac{A(j) \cap F(i)}{A(j)}, \text{ amely az adatbázisban található}$$

összes releváns tétel arányát mutatná.

Az adatbázis relevanciamutatójának becslésére a következő módszer javasolható. Az adatbázisból n darab m elemű véletlen halmazt, $|V(k)|$, választunk ki. A $V(k)$ halmazok elemei közül kiszelektáljuk az $F(i)$ szempontból releváns tételeket.

Egy adott $V(k)$ visszakeresés pontossága számolható:

$$P_k = \frac{V(k) \cap F(i)}{m},$$

n darab $V(k)$ halmaz átlagos pontossága:

$$\bar{P} = \frac{1}{n} \sum_{k=1}^n P_k,$$

ahol $\bar{P}$ 95,5%-os valószínűséggel beleesik az adatbázis relevanciamutatója (AR)

$$\pm \frac{2 \cdot \sqrt{\sum_{k=1}^n \left(\frac{P_k - \bar{P}}{n} \right)^2}}{n} \quad \text{széles konfidenciasáv-}$$

jába.

Tehát jó közelítéssel

$$\bar{P} = \frac{A(j) \cap F(i)}{A(j)},$$

a véletlenszerűen visszakeresett halmazok átlagos pontossága egyenlő az adatbázis relevanciamutatójával.

Az $\frac{A(j) \cap F(i)}{A(j)}$ hányados segítségével

az $A(j)$ ismeretében az $A(j) \cap F(i)$ és ezáltal a teljességi mutató is számítható.

Beérkezett: 1988. VII. 26-án.

JELENTKEZÉSI FELHÍVÁS

Az Országos Műszaki Információs Központ és Könyvtár, valamint az Idegennyelvi Továbbképző Központ közös szervezésében 1989. szeptember 27-én

angol nyelvű informatikai tanfolyamot indítunk

információs szakemberek és információfelhasználók részére. A tanfolyam 60 órás, ebből 36 óra nyelvi kondicionálás (meglevő ismeretek felrészítése, beszéd-készség javítása), 24 óra pedig angol nyelvű informatikai tárgyú előadás. A nyelvórákra és az előadásokra heti 2x4 órában, délelőtti órákban kerül sor.

A tanfolyam helye: az OMIKK tanácsterme (1088 Budapest, Múzeum u. 27. I. em.). A tanfolyam elvégzése után a hallgatók látogatási bizonyítványt kapnak.

Az informatikai tárgyú nyelvi gyakorlatok témakörei:

- A műszaki tudományos információk forrásai
- Szakirodalmi információ mint a vezetés erőforrása
- A mikrográfia szerepe a dokumentumkezelésben
- Iparvállalatok és kutatóintézetek műszaki-tudományos információs rendszerei

- Mikroszámítógépek az információs és könyvtári munkában
- Adatbázis-kezelő rendszerek
- Az NTMIR és más nemzetközi információs rendszerek
- Információs technológiák a 80-as években

A részvételt középfokú angol nyelvvizsgával vagy annak megfelelő szintű nyelvismerettel rendelkező informatikai érdeklődésű szakembereknek ajánljuk.

A részvételi díj 4000 Ft, melynek átutalása csak a tanfolyami jelentkezés elfogadása után válik esedékessé (német és orosz nyelvű informatikai tanfolyamokat legközelebb 1990 tavaszán indítunk).

Jelentkezési határidő: 1989. szeptember 15.

Cím:

OMIKK oktatási és módszertani osztály,
Nagy Gábor
 1088 Budapest, Múzeum u. 17.
 Tel.: 382-935, 382-300/111