

Almát a körtével?

Tudománymetriai összehasonlítások szakterületek között

Schubert András

MTA Könyvtár és Információs Központ, Tudománypolitikai és Tudományelemzési Osztály, Budapest

Közismert tény, hogy a tudománymetriai mutatószámok mindegyike erősen függ a szakterülettől. Sokan ezért eleve idegenkednek e mutatószámok szakterületek közötti bármiféle összehasonlításától. A szerző áttekinti azokat a lehetőségeket, amelyek a legfontosabb tudománymetriai mutatószámok: a publikációs szám, az idézettség vagy a h-index normalizálására és ezen keresztül a szakterületek közötti összehasonlítására felhasználhatók. *Orv. Hetil.*, 2016, 157(16), 631–634.

Kulcsszavak: tudománymetria, mutatószámok, szakterületek közötti összehasonlítás

Apples with oranges? Comparison of scientometric indicators between fields

It is well known that all scientometric indicators strongly depend on research fields. Therefore, there is a certain reluctance to make any cross-field comparison of these indicators. The paper reviews the possibilities to normalize the most important scientometric indicators: publication counts, citation rate or h-index, thus making them suitable for cross-field comparison.

Keywords: scientometrics, indicators, cross-field comparison

Schubert, A. [Apples with oranges? Comparison of scientometric indicators between fields]. *Orv. Hetil.*, 2016, 157(16), 631–634.

(Beérkezett: 2016. február 7.; elfogadva: 2016. február 25.)

„Az összehasonlítások utálatosak” – tartja az évszázados bölcsesség [1]. Az évszázados patina azonban nem feltétlenül jelent megfellebbezhetetlen igazságot. Egy másik – számomra régóta irányadó és emlékeim szerint (más forrásra nem akadtam) *Samuel Johnson*tól származó – mondás szerint „minden hasonlat félrevezető, de semmi sem kevésbé félrevezető, mint egy jó hasonlat”. Ami pedig a címben szereplő párosítást illeti, én személy szerint ki nem állhatom a körtét, tehát számomra az összehasonlítás triviális: a legrosszabb alma is jobb, mint a legjobb körte.

A lényeg talán az, hogy nem a dolgokat magukat kell összehasonlítani, hanem meghatározott minőségi vagy mennyiségi tulajdonságaikat. Aligha vitatja bárki, hogy össze lehet hasonlítani egy alma és egy körte színét vagy súlyát. A következő kérdés az összehasonlítás relevanciája. Milyen következtetést vonhatunk le az összehasonlítás eredményéből?

Ezeket a szempontokat kell szem előtt tartani a tudománymetriai összehasonlítások esetében is. Meg kell találni a mérhető és összemérhető tulajdonságokat, és gondosan meg kell fontolni az összehasonlítás eredményéből levonható következtetéseket.

A mutatószámok szakterületek közötti összehasonlításának módszere általában az, hogy a mutatószámokat a saját szakterületükre vonatkozó sztenderdekhez kell mérni, és az így normalizált mutatószámokat lehet azután összemérni.

Publikációs termelékenység

A publikációs termelékenység szakterületek közötti összehasonlításának a legfőbb akadálya az, hogy – bármilyen meglepő – a legkevésbé erre a mutatószámra állnak rendelkezésre megbízható szakterületi sztenderdek. Ennek oka főképpen az, hogy a nagy bibliográfiai adatbázi-

sokban a legutóbbi időkig a szerzőket csak a családnév és egy-két névbetű jelölte, és ennek alapján a személyeket nem lehetett egyértelműen azonosítani. A publikációs produktivitási listák élén általában a *Wang, Y., Li, J., Tanaka, K.* vagy *Smith, J.* nevek szerepeltek, esetenként több ezer publikációval.

A problémát csak részben orvosolták azok a próbálkozások, amelyek a neveket automatikus, szövegelemzési módszerekkel igyekeztek azonosítani. Igen kifinomult módszereket alkottak, amelyeket több-kevesebb sikerrel próbáltak ki az orvosi szakirodalom nagy adatbázisain is [2, 3]. Konkrét eredmények azonban legfeljebb egy-két szűk szakterületen, illetve néhány kisebb országra vonatkozóan születtek.

Gyökeres változást hozhatnak azok a szerzői azonosítók (Researcher ID, ORCID, Scopus Author ID), amelyek a legutóbbi években kezdenek meghonosodni a bibliográfiai adatbázisokban. Ha ezek valóban beépülnek a nagy nemzetközi és multidiszciplináris adatbázisokba (Web of Science, Scopus), akkor reménykedhetünk széles körű, nagy adatmennyiségre alapozott elemzések elkészültében, amelyek alapul szolgálhatnak a szerzői produktivitás szakterületek közötti összehasonlításához is.

Idézettség

Az idézetelemzés kezdetei óta nyilvánvaló, hogy a tudományos cikkekre kapott idézetek száma rendkívül erősen függ a szakterülettől. Ezt alapvető szempontként kell figyelembe venni a kutatói teljesítmény idézettség alapján történő megítélése során is. *Garfield* már évtizedes tapasztalatként szűri le könyvében [4] a tanulságot: „A helyett, hogy közvetlenül hasonlítani össze egy matematikust mondjuk egy biokémikussal, először mind a kettőt saját kollégáik között kell rangsorolnunk, és utána a saját rangsorukban elfoglalt helyezésüket hasonlíthatjuk össze.”

Általánosan véve cikkeket hasonló cikkekkal, kutatókat hasonló kutatókkal, intézményeket hasonló intézményekkel, folyóiratokat hasonló folyóiratokkal kell összehasonlítani. A „hasonlóság” pontos meghatározása, illetve a hasonló elemek kiválasztása az a kérdés, amelyre a tudománymetriai kutatások évtizedek óta keresik a legjobb választ.

A legkézenfekvőbb megközelítés a létező szakterületi besorolásokból indul ki. A nagy bibliográfiai adatbázisok (Web of Science, Scopus, MEDLINE/PubMed) mind

1. táblázat | A különböző korú publikációk átlagos idézettsége az egyes szakterületeken (A Thomson-Reuters Essential Science Indicators 2016. januári adatai alapján)

| Szakterület | A publikálás óta eltelt évek száma | | | | | | | | | | |
|----------------------------------|------------------------------------|------|------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Anyagtudományok | 0,49 | 2,80 | 5,56 | 8,35 | 10,63 | 12,85 | 13,96 | 14,99 | 16,71 | 16,13 | 17,07 |
| Biológia és biokémia | 0,48 | 2,89 | 6,57 | 10,45 | 14,24 | 18,19 | 22,15 | 24,88 | 27,86 | 31,14 | 34,15 |
| Farmakológia és toxikológia | 0,37 | 2,31 | 5,23 | 8,17 | 10,87 | 14,13 | 16,73 | 19,62 | 22,03 | 24,81 | 25,82 |
| Fizika | 0,50 | 2,58 | 5,19 | 7,98 | 9,91 | 12,28 | 14,12 | 15,62 | 16,19 | 17,24 | 18,68 |
| Földtudományok | 0,42 | 2,04 | 4,71 | 7,53 | 10,59 | 12,91 | 15,84 | 17,60 | 19,16 | 22,13 | 23,58 |
| Ideg- és viselkedéstudományok | 0,50 | 2,98 | 6,85 | 11,10 | 15,40 | 19,83 | 23,86 | 27,10 | 31,50 | 34,01 | 37,82 |
| Immunológia | 0,56 | 3,41 | 7,76 | 12,01 | 16,84 | 21,41 | 26,09 | 29,74 | 33,56 | 36,71 | 40,79 |
| Kémia | 0,61 | 3,36 | 6,73 | 10,25 | 12,81 | 15,37 | 17,14 | 19,26 | 20,04 | 21,86 | 23,61 |
| Klinikai orvostudomány | 0,40 | 2,23 | 5,13 | 8,16 | 11,07 | 14,07 | 16,93 | 19,22 | 21,88 | 25,15 | 27,84 |
| Környezettudományok, ökológia | 0,39 | 2,08 | 4,93 | 8,21 | 11,16 | 14,56 | 17,11 | 20,71 | 23,41 | 26,09 | 29,28 |
| Közgazdaság-tudomány | 0,19 | 0,87 | 2,22 | 3,82 | 5,89 | 7,97 | 9,92 | 11,60 | 14,42 | 17,38 | 19,53 |
| Matematika | 0,13 | 0,65 | 1,49 | 2,50 | 3,59 | 4,72 | 5,57 | 6,42 | 7,08 | 7,80 | 8,44 |
| Mezőgazdaság-tudomány | 0,27 | 1,43 | 3,26 | 5,12 | 6,94 | 8,95 | 10,52 | 12,08 | 14,58 | 16,66 | 18,11 |
| Mikrobiológia | 0,43 | 2,80 | 6,22 | 9,37 | 12,94 | 17,70 | 20,89 | 23,34 | 26,87 | 30,04 | 33,98 |
| Molekuláris biológia és genetika | 0,60 | 3,99 | 9,55 | 15,24 | 22,03 | 28,13 | 34,08 | 39,05 | 46,02 | 49,58 | 54,07 |
| Műszaki tudományok | 0,24 | 1,41 | 3,11 | 4,71 | 6,46 | 7,95 | 9,19 | 9,57 | 10,77 | 11,15 | 11,13 |
| Növény- és állattan | 0,29 | 1,52 | 3,46 | 5,44 | 7,59 | 9,82 | 11,63 | 13,09 | 14,93 | 17,22 | 18,57 |
| Pszichiátria és pszichológia | 0,33 | 1,76 | 4,12 | 6,88 | 10,11 | 13,39 | 16,11 | 19,56 | 23,08 | 25,77 | 28,88 |
| Számítástudomány | 0,18 | 1,06 | 2,48 | 3,96 | 5,73 | 7,29 | 8,90 | 9,51 | 10,55 | 7,45 | 7,86 |
| Társadalomtudományok | 0,23 | 0,94 | 2,24 | 3,68 | 5,30 | 7,06 | 8,51 | 10,06 | 12,40 | 14,26 | 15,71 |
| Űrtudományok | 0,86 | 4,48 | 8,64 | 12,55 | 16,29 | 19,93 | 22,69 | 23,24 | 26,85 | 29,37 | 29,71 |

tartalmaznak szakterületi kategóriákat, és ezek alapján meg lehet határozni különféle publikációs és idézettségi időtartamokra vonatkozó idézettségi sztenderdek. Az adatbázisok kiegészítő szolgáltatásai készen kínálnak ilyen táblázatokat. Példaképpen az 1. táblázat egy ilyen összeállítást mutat be a Web of Science-hez kapcsolódó Essential Science Indicators (ESI) nyomán.

A szakterületi besorolásokon alapuló idézettségi normalizálás legfőbb gyengesége a besorolások esetlegessége. Az ESI esetében első látásra feltűnő a szakterületek méret és specifitása szerinti heterogenitása. A fizika és az immunológia nyilvánvalóan nincs egy „súlycsoportban”, ugyanígy a klinikai orvostudomány és a mikrobiológia sem. Hasonló problémákkal terhelt a többi hasonló szakterületi kategóriarendszer is, ráadásul a különböző adatbázisok rendszerei nem kompatibilisek.

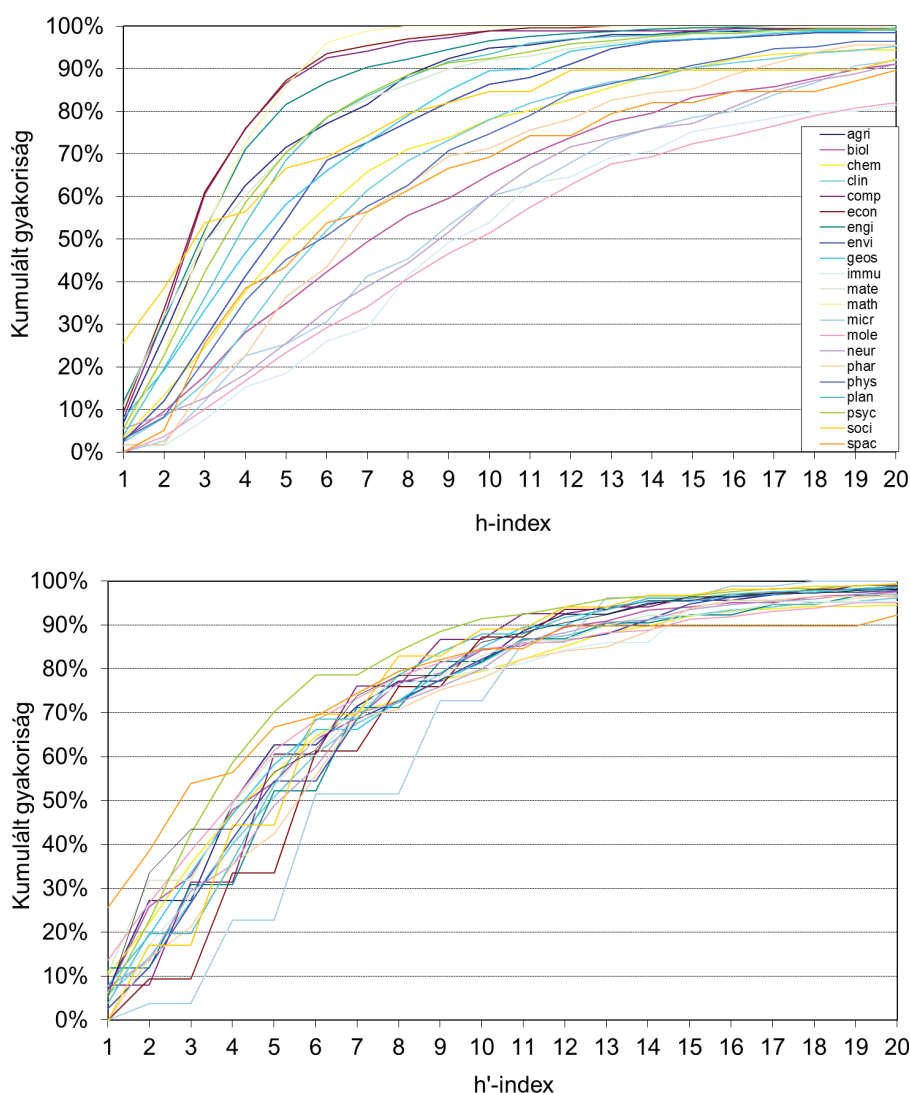
Majd’ negyedszázaddal ezelőtti cikkünkben [5] három módszert javasoltunk az összehasonlítási alapul szolgáló sztenderdek kiválasztására.

1. Az egyes cikkek ténylegesen kapott idézettségét hasonlíthatjuk az ugyanabban a folyóiratban, ugyanabban

az évben megjelent többi cikk idézettségéhez (a várható idézettséghez). A kettő aránya, a relatív idézettség megadja, hogy egy cikk vagy egy cikkhalmaz (például egy szerző vagy egy intézmény közleményei) mennyivel idézettebb vagy kevésbé idézett a közlő folyóiratok átlagánál.

2. Abban az esetben, ha a vizsgált közlemények nem a szakterület specifikus szakfolyóirataiban, hanem általánosabb, multidiszciplináris folyóiratokban jelentek meg, pontosabb eredményt adhat, ha a cikkek irodalomjegyzékében szereplő folyóiratok képezik az összehasonlítás alapját. Egy *Nature*-ben megjelent geológiai cikk irodalomjegyzékében feltehetően számos geológiai folyóirat található; ezek alapján valószínűleg realisabb képet kapunk a cikk várható idézettségére, mint ha a *Nature* összes cikke alapján számolunk. Ugyancsak az irodalomjegyzékekben szereplő folyóiratok alapján tudjuk megítélni a folyóiratok relatív idézettségét.

3. A legmunkaigényesebb, de legtestreszabottabb összehasonlítási alap a vizsgált cikkhez bibliográfiailag csatolt cikkek halmaza. A bibliográfiai csatolás [6] azt jelen-



1. ábra | A folyóiratok normalizálatlan (h) és a normalizált (h') h -indexének eloszlása az ESI szakterületi besorolása szerint [12]

ti, hogy azokat a cikkeket válogatjuk ki, amelyeknek az irodalomjegyzéke a legjobban hasonlít a vizsgált cikkéhez. A hasonló irodalomjegyzék jó eséllyel szoros tematikai hasonlóságot tükröz. Ez az eljárás megfelel a legújabb nagy népszerűségnek örvendő cikkszintű értékelés („article-level metrics”, „altmetrics”) alapkritériumának.

Lényegében mindmáig ezekre a módszertani elvekre épülnek az idézettség szakterületek közötti összehasonlítási módszerei. A cikkünket idéző száznál több publikáció között van a téma legidézettebb [7] és e munka írásáig legfrissebb [8] cikke egyaránt.

H-index

Egy publikációhalmaz h-indexe függ a publikációk számától és idézettségétől is. A h-index szakterületi normalizálását tehát mindkét összetevő figyelembevételével lehet csak elvégezni. Glänzel [9] teremtette meg az elméleti alapjait annak a közelítésnek, amelynek első sikeres alkalmazását Csajbók és mtsai cikkében [10] találhatjuk. E szerint a h-index (h), a mintanagyság (n, a cikkek száma) és az átlagos idézettség (x) között a $h = c \cdot n^{1/3} x^{2/3}$; összefüggés áll fenn, ahol c 1 nagyságrendű pozitív állandó.

Ennek az összefüggésnek az alapján a h-indexet normalizálni lehet az egyes szakterületek jellemző publikációs- és idézettségadatainak ismeretében. A normalizált h' indexet a teljes sokaság átlagos idézettsége (x_0) és a szakterületi átlag (x) ismeretében a $h' = (x_0/x)^{2/3}$ képlet segítségével számíthatjuk ki [11, 12].

Az 1. ábrán bemutatjuk a folyóiratok normalizálatlan (h) és a normalizált (h') h-indexének eloszlását az ESI szakterületi besorolása szerint. Feltűnő, hogy a normalizálás után az indexek eloszlása lényegesen közelebb kerül egymáshoz, mint előtte volt, alátámasztva a módszer eredményességét.

Zárszó

Kétségtelen, hogy a tudománymetriai mutatószámok mindegyike – a legegyszerűbb publikációs számtól kezdve a különféle idézettségi mutatószámokon keresztül a legkifinomultabb hálózati mérőszámokig – erősen szakterületfüggő. Ez azonban nem jelenti azt, hogy nem lehet módot találni a szakterületek közötti összehasonlításra. A legelső kérdés természetesen mindig az kell, hogy legyen, hogy van-e célja, és ha igen, akkor mi a célja az összehasonlításnak. Csak a cél pontos ismeretében találhatjuk ugyanis meg a célnak legjobban megfelelő eszközt. A folyóiratok szakterületek közötti összehasonlítása például ugyan jó modell lehet egy módszertani kísérletben, de nehéz elképzelni olyan gyakorlati problé-

mát, amikor különböző szakterületeken működő folyóiratokat kellene egymással összemérni.

A tudománymetriai mutatószámok szakterületi normalizálásának és ezen keresztül a szakterületek közötti összehasonlítás lehetőségének nagy irodalma van, és a nagy bibliográfiai adatbázisok hatalmas adattömegeinek feldolgozása egyre bővíti a lehetőségeket. Ezzel azonban a választás nehézsége is egyre fokozódik. Szakértelem és körültekintés kell az adott feladat megoldására legalkalmasabb eszköz kiválasztásához és alkalmazásához.

Anyagi támogatás: A szerző köszöni az FP7-SSH-2013-2 #613202 (IMPACT-EV) projekt támogatását.

A cikk végleges változatát a szerző elolvasta és jóvá hagyta.

Érdekltségek: A szerzőnek nincsenek érdekltségei.

Irodalom

- [1] Primary source of the phrase “comparisons are odious” is John Lydgate: Debate between the Horse, Goose, and Sheep, cca. 1440. „Odyous of olde been comparisonis, and of comparisonis engendyrd is haterede.” <http://www.phrases.org.uk/meanings/Comparisons-are-odious.html>
- [2] Torvik, V. I., Smalheiser, N. R.: Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data, 2009, 3(3), 1–29.
- [3] Lin, W., Dogan, R. I., Kim, S., et al.: Author name disambiguation for PubMed. J. Assoc. Inform. Sci. Technol., 2014, 65(4), 765–781.
- [4] Garfield, E.: Citation indexing. Its theory and applications in science, technology, and humanities. Wiley, New York, 1979.
- [5] Schubert, A., Braun, T.: Reference standards for citation based assessments. Scientometrics, 1993, 26(1), 21–35.
- [6] Kessler, M. M.: Bibliographic coupling between scientific papers. American Document, 1963, 14(1), 10–25.
- [7] Radicchi, F., Fortunato, S., Castellano, C.: Universality of citation distributions: Toward an objective measure of scientific impact. Proc. Natl. Acad. Sci. U.S.A., 2008, 105(45), 17268–17272.
- [8] Colliander, C.: A novel approach to citation normalization: A similarity-based method for creating reference sets. J. Assoc. Inform. Sci. Technol., 2015, 66(3), 489–500.
- [9] Glänzel, W.: On the h-index – A mathematical approach to a new measure of publication activity and citation impact. Scientometrics, 2006, 67(2), 315–321.
- [10] Csajbók, E., Berhidí, A., Vasas, L., et al.: Hirsch-index for countries based on Essential Science Indicators data. Scientometrics, 2007, 73(1), 91–117.
- [11] Iglesias, J. E., Pecharromán, C.: Scaling the h-index for different scientific ISI fields. Scientometrics, 2007, 73(3), 303–320.
- [12] Schubert, A.: Rescaling the h-index. Scientometrics, 2015, 102(2), 1647–1653.

(Schubert András,
e-mail: schuba@iif.hu)