Kontrollált és nem kontrollált szótárak összekapcsolásának lehetőségei (az LCSH és a Delicious példáján keresztül)

A folkszonómia egy a közösségi taggelés, felhasználói osztályozás eredményeként létrejövő kontrollálatlan szótár a digitális információk rendszerezésére, kategorizálására (bővebben l. http://tmt.omikk.bme.hu/show_news.html?id=5120&issue_id=503 [A ref.]). A cikk úttörőnek mondható annyiban, hogy elsőként vet össze egy „kollaboratív taggelési rendszert", a webes tartalmak indexelésére létrehozott *Delicious*t (http://delicious.com/) egy létező kontrollált szótárral, a washingtoni *Kongresszusi Könyvtár* tárgyszórendszerével (*Library of Congress Subject Headings* = *LCSH*). (Ez utóbbit alkalmazza újabban a *Google Book Search* is.) A végső cél annak vizsgálata, hogy miként lehetséges a folkszonómiák alapvető hiányosságainak – úgymint pontatlanságok, kétértelműségek, félreérthetőségek, az ezekből fakadó kaotikusság – kiküszöbölése azáltal, hogy egy szakemberek által felépített, ellenőrzött szótárhoz kapcsoljuk őket.

A szerzők három kérdésre keresik a választ:

1. Milyen mértékű átfedés van a két szótár között?
2. Miként oszlanak meg a folkszonómia tagjei az LCSH teljes hierarchikus struktúrájában?
3. Mennyiben valósítható meg a két rendszer összekapcsolása?

A cikk által részletesen ismertetett módszer segítségével a szerzők a Delicious rendszerében adott időpontban legfrissebben indexelt 4552 weboldalhoz tartozó, egynél többször előforduló, szám szerint 299 tag előfordulásait vizsgálják a 291 ezer besorolási tételt (értelemszerűen ugyanennyi főtárgyszót) tartalmazó LCSH-ben, pontosabban a Kongresszusi Könyvtár 1985–2005 között épített egységesített tárgyszólistájában (subject authority file). (A vizsgálatból az altárgyszavakat kizárták.) Az elemzés érdekében, kihasználva a tárgyszóként alkalmazott fogalmak közötti hierarchikus kapcsolatokat (l. általánosabb fogalom – speciálisabb fogalom) az LCSH deszkriptorait és nemdeszkriptorait egy 28 136, elviekben egymástól független fogalomkörből, „fából" álló fastruktúrában ábrázolták, amelyek törzsei egy e célból létrehozott, „fiktív"gyökércsomópontban találkoznak. (Erre azért volt szükség, mert az LCSH-t eredetileg nem tezaurusznak tervezték.)

A tagek használatára vonatkozó elemzések során a szerzők az összes (388) angol nyelvű felhasználói taget vették alapul, míg az összehasonlítás során a már említett 299 taget vizsgálták.

Kiderült, hogy a Delicious rendszerében található tagek használati gyakoriságára jellemző, hogy kisszámú taget használnak sokan, s a tagek jelentős részét csak alig néhányan. Vagyis egy koordináta-rendszerben, ahol *y* tengelyen a használati gyakoriság, *x* tengelyen pedig az egyes tagek felsorolása található, a gyakoriság egy inverz *J* formát mutat.

A vizsgálat fényt derített továbbá arra, hogy a használt tagek mintegy 11%-a többszavas kifejezés. Miután a rendszer nem engedi, hogy szóközt használjunk, a felhasználók három különböző módon jelenítették meg a több szóból álló kifejezéseket: szóközelhagyással (opensource), kötőjellel (web-services) és aláhúzás beiktatásával (rare_book).

24 olyan szót találtak, amely egyes számban és többes számban is jelen van a Deliciousban. 11 szópár esetén az egyes számú (pl. blog), ugyancsak 11 szópár esetén a többes számú alak (pl. movies) használata volt a jellemzőbb. 13 szó két eltérő nyelvtani formában is jelen van a szótárban (pl. podcast – podcasting), négy szó eltérő helyesírással jelenik meg, például: cataloging – cataloguing), egy rövidítés és egy betűszó is található a szótárban a feloldásával együtt: fr – french vagy france, illetve os – operatingsystem.

A két szótár közötti átfedések vizsgálatakor a „teljes szóegyezéseket" vették számba. Ilyen egyezésnek számít a „computer" kifejezésnek a „computer networks" szókapcsolatban való előfordulása, míglen a „network" szó nem mutat „teljes szóegyezést" többes számú alakjával, a „netwoks"-szel. Ennek az elvnek az alkalmazására azért volt szükség, hogy a hibás egyezéseket (pl. work–network) kiszűrjék, jóllehet nyilvánvaló, hogy ebből adódóan a ténylegesnél kevesebb előfordulást tudtak kimutatni.

Az eredmény: a tagek 60,9%-a (a 299-ből 182) mutat teljes szóegyezést az LCSH tárgyszavaival. Tíz olyan taget találtak, amely mindössze egyetlen tárgyszóval mutat egyezést, ilyenek például: „rss"–„RSS Feed", „metadata"–„Metadata". Egy olyan tag volt (a „language" szó), amely 2680 tárgyszóban szerepelt. Gyakori előfordulást mutatnak még a következő kifejezések: „people", „game", „art", „literature". Az esetek nagy többségében egyetlen tag sokszori előfordulásáról van szó. A teljes szóegyezést nem mutató kifejezések jelentős része, 45,1%-a műszaki jellegű (pl.: folksonomy, mysql). 35,4%-uk többszavas kifejezés. Ezek egy része az LCSH egyes tárgyszavaihoz igencsak közelálló terminus (pl.: rare_books–rare books), ennek ellenére – a „teljes szóegyezés" elve alapján – csakúgy nem vették őket számításba, mint a tartalmi szempontból ugyancsak egyezést mutató „blog"–„weblog" vagy „css"–„Cascading Style Sheets" szópárokat.

A következőkben megvizsgálták a 182 egyezést mutató tag megoszlását az LCSH hierarchikus fogalmi (fa)struktúrájában. 15 olyan taget találtak, amely mindössze két fában (fogalomkörben) található meg. A „literature" kifejezést ugyanakkor 160 fában találták meg, amely a szerzők szerint bizonyítja az egyes fák (fogalomkörök) egymástól való függetlenségének hipotézisét. A tagek többsége egyenként nagyjából 60 különböző fogalomkörbe tartozik.

A „literature" szó 1280 tárgyszóegyezést mutat, s egy-egy fában átlagosan nyolcszor fordul elő. Összehasonlításképpen: a „library" kifejezés 219 tárgyszóban található meg, és 45 fogalomkörben bukkanhatunk rá.

Ami a fentiekben feltett, a szótárak közötti átfedésre vonatkozó kérdést illeti, egyfelől látnunk kell, hogy az újonnan keletkező szakkifejezések némi késedelemmel kerülnek be az LCSH-ba, míg – ahogy az várható – a folkszonómiákba gyorsan beépülnek. A vizsgálat során kiderült, hogy a tagek 45%-a specifikus szakkifejezés, például egy szoftveralkalmazás vagy hardver neve. Ezek nagyrészt nem találhatóak meg az LCSH-ban, ezzel szemben a generikusabb kifejezések (úm. education, language) számos fogalomkörön belül előfordulnak ugyanott. További vizsgálatokat igényelnek a többszavas kifejezések és a különböző nyelvtani alakokban előforduló szavak előfordulásai.

A felhasználók tagjei 61%-ban vannak jelen adott formájukban az LCSH-ban, ugyanakkor további 10% „potenciális előfordulással" számolhatunk, ha megoldjuk a többszavas kifejezések és a toldalékkal ellátott szavak problémáját. A formai mellett a szemantikai zavar is kiküszöbölendő. Egy példa rá az „ajax" tag, amely a Delicious-ben az „Asynchronic JavaScript and XML" technológiára utal, míg a Kongresszusi Könyvtár szótárában a görög mitológiai hősre.

A második kérdés a következőképpen hangzott: Miként oszlanak meg a folkszonómia tagjei az LCSH teljes hierarchikus struktúrájában?

A számítógép által készített szóegyezési vizsgálatok eredményeinek kiértékelésekor a szerzők megállapították, hogy a megoszlás aszimmetrikus: a tagek 96%-a 67-nél kevesebb fogalomkörben (fában) fordul elő, míg a maradék 4% előfordulási mutatója 68 és 160 között mozog a fogalomkörök tekintetében. Kiküszöbölve ugyanakkor az egyes tárgyszavak többszöri jelenlétét az LCSH rendszerében, jóval kevesebb tag-előfordulást regisztrálhatunk. A fenti példák esetén a „literature" és a „library" szavak előfordulása 1280-ról, illetve 219-ről 1067-re (17%-os csökkenés), illetve 127-re (42%-os csökkenés) redukálódik.

A tanulmány megállapította, hogy a fogalomkörönkénti megoszlás esetén a *Zipf-törvény* érvényesül, vagyis egy szó előfordulási gyakoriságának (*f*) és az előfordulások száma szerinti listán elfoglalt helyezésének (*r*) szorzata egy konstans értéket (*K*) eredményez, azaz $f \times r = K$. Ez más szóval azt jelenti, hogy az előfordulási gyakoriságok értékei eleinte meredeken zuhannak, majd – az alacsonyabb értékek esetén – kisebb mértékű csökkenést mutatnak. A Zipf-törvény segítségével leírható a fogalomkörök relevanciája a tageket illetően, vagyis a több előfordulást tartalmazó fogalomkörök relevanciája nagy valószínűséggel nagyobb (helyezésük az adott ranglistán előrébb található), mint a kevesebb előfordulást mutatóké, ugyanakkor számolni kell a fogalomkörök méreteivel (vagy-

is a bennük található csomópontok/tárgyszavak számával) is.

A harmadik kérdésre keresve a választ, azt a potenciált kell közelebbről szemügyre vennünk, amely a két rendszer összekapcsolásában rejlik. A jelenlegi felhasználói taggelésen alapuló rendszerek alkalmasak bár a webforrások bizonyos fokú rendszerezésére, hátrányuk, hogy a keresési lehetőségeik meglehetősen behatároltak. Nem kereshetünk egy időben több tagre, nem áll módunkban logikai operátorokat használni, és a találatok megjelenítésekor a sorrendet nem a relevancia, hanem a kronológiai szempont határozza meg. Megfontolandó volna tehát e szótárak kontroláltabbá tétele az LCSH valamilyen módon történő integrálásával. Így például olyan esetekben, ahol több tag (szó) hoz releváns találatot, érdemes volna inkább a Kongresszusi Könyvtár tárgyszavait alkalmazni. Hosszabb távon megvalósítható volna a taggelési rendszerek hozzákapcsolása egyéb digitális információs rendszerekhez, így például az OPAC-okhoz. A Google Book Search, amelynek találati oldalai releváns bibliográfiai leírásokat is tartalmaznak, újabban az LCSH vonatkozó tárgyszavait is felhasználja – a találati oldalba integrálható – hivatkozások generálására. Hasonlóképpen, az LCSH „közvetítésével" a kollaboratív taggelési rendszerekben található webes forrásokat integrálni lehetne a könyvtárak OPAC-jaiba.

A két rendszer, a nem kontrollált és a kontrollált szótár összekapcsolásának pozitív hozadéka volna a két indexelési szemlélet: a felhasználóközpontú és a rendszerközpontú megközelítés termékeny egyesülése. Az összekapcsolást ugyanakkor több tényező is megnehezíti: az LCSH-ban meg nem található műszaki kifejezések gyakorisága a tagek között, továbbá a több szóból álló kifejezések, ragozott alakok, rövidítések, betűszavak, amelyek nem feleltethetők meg LCSH-tárgyszavaknak. Ezek a problémák megoldhatók a kifejezések normalizálásával: a ragozott szavak lemmatizálásával, a többszavas kifejezések előre definiált formában való rögzítésével egy adatbázison vagy szótáron belül.

A felhasználók által adott tárgyszavak, tagek sok esetben eléggé együgyűek ahhoz, hogy a visszakereshetőséget biztosítsák. A folkszonómiáknak az LCSH-hoz hasonló kontrollált szótárakhoz való kapcsolása nagymértékben javíthat ezen a helyzeten.

/YI, Kwan – CHAN, Lois Mai: Linking folksonomy to Library of Congress subject headings: an exploratory study. = Journal of Documentation, 65. köt. 6. sz. 2009. p. 872–900./

(Dancs Szabolcs)