Gayer Zoltán

Ízléselemzés és Big Data

Megjósolható-e egy nő vásárlási szokásaiból, hogy terhes? Mint Dessewffy és Láng (2015) tanulmányában láthattuk, igen, megjósolható! S ha ez sikerült, akkor milyen hasonló, nagy kérdések jósolhatók még meg? Megjósolható-e, hogy valaki kire fog szavazni? Hogy miként fog alakulni az ízlése? Hogyan fog viselkedni pánikhelyzetben, pl. amikor ég a ház? Megjósolható, hogy mikor kap el egy betegséget? Megjósolható-e, hogy meddig fog élni?

A válasz ezekre a kérdésekre az, hogy igen, és ez több okból kötődik a Big Data jelenségéhez. A Dessewffy és Láng tanulmányában felvetett kérdésekhez egy konkrét Big Datá-s projektünk néhány érdekes eredményének ismertetésével szeretnék hozzájárulni, de mielőtt így tennék, egy kiegészítést fűznék a tanulmányhoz.

Jóslásokra, előrejelzésekre természetesen mind a Big Datára építő, mind a reprezentatív mintás kutatások is képesek. De miben áll a különbség e kettő között? Mert az világos, hogy szociológiai módszerekkel eddig is megjósolható volt, hogy az emberek hogyan fognak viselkedni, például hogy kire fognak szavazni.

Amiben a különbség megmutatkozik a Big Datán alapuló és a reprezentatív mintás kutatásokban, az az, hogy míg a reprezentatív mintán alapuló közvélemény-kutatások révén a szociológiailag meghatározott (nem, kor, iskolai végzettség és egyéb, puhább kategóriák szerinti) csoportok valószínű viselkedése jósolható meg, addig a Big Datán alapuló módszerek esetén az *egyéné is*. Ez adódik az adatfelvétel különbségéből, hiszen a Big Datá-s projektek rendszerint olyan adatokon nyugszanak, amelyekben az egyén beazonosítható, míg a reprezentatív kutatásokban nem, sőt az, hogy az egyén ne legyen beazonosítható, etikai elvárás is a reprezentatív kutatásokkal szemben. Ennek az ígéretnek a betartását maguk a kutatók garantálják, a kérdőívekre válaszolók pedig – joggal – elvárják. Ezzel szemben a Big Data alapjául szolgáló rendszerekben (közösségi oldalak adatbázisai, vásárlási adatbázisok, ügyféladatbázisok) jellemzően nemhogy nem tiltott, hanem a folyamatnak egyenesen természetes része a beazonosíthatóság, hiszen legtöbbször maga a szolgáltatás is ezen alapul. Vagyis a legtöbb, létező nagy adatbázis általában beazonosítható felhasználóhoz, ügyfélhez kötődik. Dessewffyék öt pontját tehát kiegészíthetjük egy hatodikkal: egyének, és nem csak kategóriák viselkedését értjük meg.

Ennek következménye, hogy az egyénről akár olyan szenzitív információkkal is rendelkezhetünk, amelyeket ő nem is feltétlenül osztana meg másokkal, sőt akár ő maga nem is tud róluk. Ilyen például a tanulmányban idézett, terhességgel kapcsolatos példa.[1]

Esettanulmány

Ezen gondolatok mellett a szerzőpáros tanulmányához olyan példákkal szeretnék hozzájárulni, amelyeket saját adatbázison alapuló kutatásainkból kaptunk, s amelyek némelyike igen meglepő volt számunkra is.

A LInKE rendszerben hálózati, klaszterező elemzéssel vizsgáljuk az emberek like-olási szokásait.[2] Az eredeti hipotézisünk az volt, hogy ha kellően nagy számú embernek megismerjük a hozzá pozitívan kötődő linkjeit (amelyeket like-olt, megosztott, ajánlott), akkor az URL-ekből egy hálózatot lehet építeni, és e hálózatot egy iteratív eljárással olyanná tudjuk alakítani, amelyben a pontok ízléscsomópontokká, ízlésklaszterekké állnak össze. Az eljárásban – amelyben a pontok (node-ok) az URL-ek, az élek pedig azt mutatják meg, hány embernek volt egyaránt köze két webcímhez – azokat a pontokat kötjük össze és olyan súlyú éllel, amelyeket kettő vagy több ember egyaránt like-olt, megosztott vagy ajánlott. Az így létrejövő hálózat egy extrém sűrű, igen nagy elemszámú hálózattá vált, amelynek a hálózati elemzése nem kevés munkát igényelt.[3] Hipotézisünk tehát az volt, hogy a Facebookon évek alatt kitett like-ok, megosztások és ajánlások, amelyek mindegyike egy-egy URL-hez köthető, a hálózati elemzésben ízlésmintázatokat mutatnak meg.

Ezt a hipotézisünket vizsgálatunk részben alátámasztotta. A jelen cikk írásának időpontjában a mintegy 8500 felhasználó, akiknek az adatait vizsgáljuk, összesen 8,9 millió like-kal rendelkezik, ezekből mintegy 4,4 millió az egyedi URL. E helyütt nincs lehetőség a klaszterek részletes, tudományos igényű elemzésére, csak rögzíteném azokat a megfigyeléseket, amelyeket e klaszterek kapcsán tettünk, s amelyek számunkra is megmutatták a nagy adatbázison alapuló hálózatelemzések elképesztő erejét.

Az eljárás, amelyet pillanatnyilag két-három óránként lefuttatunk a rendszerünkben,[4] több mint hatezer klasztert hoz létre az egészen kicsi, néhány tagú klasztertől elindulva az egészen nagyokig, amelyeknek több tízezer URL a tagja.[5] Az URL-ek egymáshoz kötődése révén mindegyik klaszter maga is egy-egy külön alhálózatot alkot. Nagyon sok módszer kipróbálása után jutottunk arra az állapotra, ahol megláttuk, hogy a klaszterek kezdenek értelmes csoportokat alkotni, de amikor idejutottunk, sok meglepetés ért minket.

1 Természetesen ez maga is felvet rendkívül sok és sokrétű etikai problémát, de erre e cikk keretében nincs mód részletesen kitérni.

2 A LInKE egy olyan facebookos applikáció, amelyben a felhasználóktól (az ő beleegyezésükkel) megkapjuk like-jaikat, megosztásaikat és ajánlásaikat, majd ezeket elemezzük, aminek alapján adunk nekik egy személyre szabott tartalmat. A hálózatelemzési munkát Vassy Zsolt hálózatkutató vezetésével végezzük.

3 Az elemzés több lépésben redukálja az adatokat. Előklaszterezés után végezzük el a hálózatépítést, majd az alklasztereken belül modularitásalapú, Luvain-féle klaszterező eljárással alakítunk ki csoportokat.

4 Ez egy üzleti vállalkozás is, így nem garantálom, hogy az olvasó akár e cikk megjelenésének a pillanatában is működőképes rendszert talál.

5 A klaszterek mérete attól függ, hány embernél talál meg közös mintázatot az eljárás. Minél kisebb a klaszter mérete, annál inkább lehet szó ismerősi, illetve rokoni, családi közösségeken alapuló közös like-olási mintázatokról. A nagyjából hatezer klaszter közül a legtöbb kisebb méretű, vélhetően ismerősi hálózatokban létrejövő összefüggéseket mutat meg, ugyanakkor négyszáz nagyobb méretű (50 URL feletti), sok felhasználóhoz köthető klaszter keletkezik.

Kutatóknak, szociológusoknak nem egyszerű az egyének ízlés vagy életstílus szerinti kategorizálása – holott ha valamire, akkor legtöbbször éppen ezekre a finom, puha kategóriákra van szükségünk a társadalmi folyamatok megértéséhez. A szociológiában már a kilencvenes évek előtt lezajlott az a fordulat, amelyben a kemény változók (nem, kor, iskolai végzettség, település) mellett vagy helyett egyre inkább az ún. puha változók, életstílus- és státuszcsoportok válnak az elemzés alapjává (vö. Veres et al. 2010). Egy kérdőíves kutatásban rengeteg kérdést fel kell tenni ahhoz, hogy a személyek státuszát, életstílusát, ízlését értelmezni tudjuk. Veresék kutatásában például bemutatják, hogyan igazították az ún. ESOMAR-rendszerben (vö. AAPOR 2011) szereplő termékek birtoklását mutató kérdéseket az akkori jelenhez, és vettek a rendszerbe új, a korhoz igazított termékeket.[6] A státusz mérésére még talán kevésbé, de az életstíluscsoportok beazonosítására már sokkal inkább alkalmas lehet a facebookos tevékenység. Bár nyilvánvalóan igen bonyolult kérdés, és nincs is mód részletesen kifejteni, hogy valaminek a like-olása, megosztása a Facebookon pontosan mit is jelent, de a rendszerünkből látszik, hogy egy ember átlagosan 1000 ilyen attitűdkérdésre „válaszol" facebookos tevékenysége révén.[7]

Zárójelbe téve tehát a like-ok jelentését, nyilvánvaló, hogy éppen az olyan életstíluscsoportok megtalálására lehet alkalmas a facebbokos tevékenység elemzése,[8] amelyek alapja a szabadidővel kapcsolatos attitűdök, a vásárlással, fogyasztással kapcsolatos attitűdök, illetve a márkákhoz való viszony attitűdjei (Veres et al. 2010: 14).

Tulajdonképpen az egyik legnagyobb meglepetés az volt, hogy hipotézisünk működött, és valóban, az URL-ek több mint négymilliós halmazából a többlépcsős, iterációs eljárásokból álló algoritmus révén kialakuló klaszterek legtöbbje már ránézésre is értelmes csoportot alkot.[9] Így jött létre (a teljesség igénye nélkül, csak felsorolva néhányat) az

- Anime-klaszter (2416 URL),
- Emós klaszter (1872 URL),
- Magyar labdarúgás topic klasztere (1354 URL),
- Fradi-klaszter (1200 URL),
- Egészséges étkezés klasztere (1160 URL),
- Trollfoci-klaszter (960 URL),

6 Hagyományos színes televízió, LCD, plazma TV, music center, DVD-lejátszó, videokamera, kettő vagy több autó, digitális fényképezőgép, asztali számítógép, laptop, notebook, kézi számítógép (iPad, PDA), multifunkciós tűzhely (sütés, mikrohullám, grill), hétvégi ház, nyaraló, automata mosógép, mosogatógép, elektromos kézi fúró, sarokcsiszoló, mikrohullámú sütő, benzinmotoros fűnyíró gép, fagyasztószekrény vagy legalább 3 polcos fagyasztó a kombinált hűtőben, kettő vagy több hordozható rádió CD-lejátszóval, otthoni szauna.

7 Egy felhasználótól a Facebookon keresztül átlagosan 1000 adat, azaz like, megosztás és ajánlás érkezik meg a LInKE-hez. Vagyis a „mintánkban" szereplő átlagos Facebook-felhasználó eddigi facebookos életében átlagosan 1000 tevékenységet végzett.

8 Módszertanilag persze nagyon fontos kérdés, hogy a like-olások rendszerét nem csak az egyszerű kedvelések alakítják, hanem a Facebook-felhasználók interperszonális és csoportviszonyai is. Világos, hogy egy-egy like bizonyos esetekben sokkal inkább egy-egy személyhez fűződő viszony megerősítésének céljával jön létre, vagyis része a berne-i simogatások rendszerének, mégis a nagy számok mögött sokkal több attitűdjellegű bejegyzés húzódik meg, és a hálózati elemzés révén éppen ezeket az összefüggéseket tudjuk kimutatni.

9 Maga az eljárás, vagyis az URL-ek node-okként kezelése, egy véletlen eloszlásával indul, amelyből az iterációs eljárás révén mozognak a pontok egymás felé. Az iterációs eljárás vége akkor következik be, amikor már nincs számottevő elmozdulás a pontokban, illetve a csoportokban.

…de keletkeztek ilyen URL-csoportok is, mint a

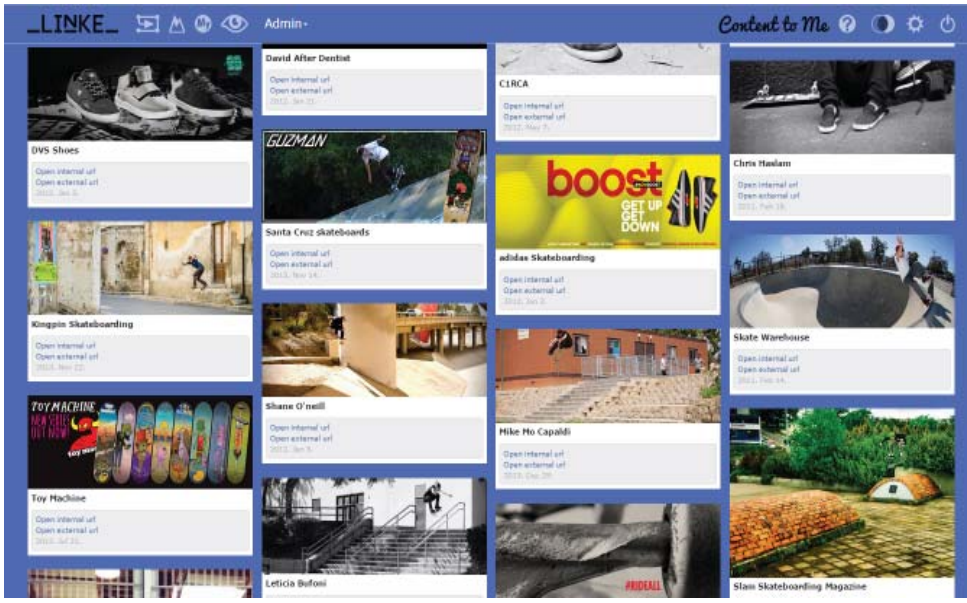
- Budapesti Műszaki Egyetemisták klasztere (606 URL),
- Ellenzéki klaszterek (több is),
- Mini morris fan-ek klasztere (512 URL),
- Real Madrid-klaszter (484 URL)

…és még rengeteg, pl. metálos, skateboardos, biciklis, kutyás, lovas, VFX, Budapest, rally, de akad például olyan klaszter is, amelyben tudományos érdekességek találhatók, ahogyan vannak cseh, szlovák, erdélyi magyar klaszterek is stb.

Az eljárás tehát értelmes csoportokba rendezte az URL-eket anélkül, hogy előre kijelöltük volna, mit keresünk. Hozzá kell tennünk, hogy maga az eljárás nem tud semmi kvalitatív szempontot, nem tudja, mit keressen, egyszerűen úgy hozza létre a csoportokat, hogy az együtt like-olások hálózatában keresi a minél homogénebb, azaz összetartozó linkeket. Ez utóbbi, hogy ti. nincs kijelölve, mit keresünk, nagyon fontos szempont az ilyen elemzések során, ugyanis így, hasonlóan a kvalitatív eljárásokhoz, olyan eredményeket kapunk, amelyekre nem számítottunk. Sőt magának az eljárásnak a lényegéhez tartozik, hogy színtisztán mátrixalapú algebrai műveletekkel jut el olyan felfedezésekhez, hogy milyen mintázatok húzódnak meg az emberek like-olási szokásai mögött.

A példa kedvéért nézzük meg az *1. ábrát*, amelyen egy skateboardos klaszter linkjei láthatók. Látszik, hogy az elemzés összetartozó linkeket hozott ki, hiszen a linkek témája közel mindegyik esetben a skateboard, amely linkek pedig nem ilyen témájúak, azok az így viselkedő (like-oló) emberek egyéb ízlésválasztásait mutatják meg.

1. ábra. Skateboardos klaszter URL tagjai a LInKE-rendszerben



Forrás: www.li-n-ke.com, adminisztrátori felületről látható oldal

Magukat a klasztereket külön-külön is lehet elemezni, és meghatározni, hogy az összes közül mely tagnak van a klaszterben kiemelkedő szerepe.[10] Így – elkanyarodva a fenti skateboardos példától – kifejezetten érdekes volt, hogy a Fradi iránt elkötelezett felhasználók URL-hálózatának középpontjában a focis oldalak mellett milyen komoly szerepe van a *Trónok harca*-sorozatnak, a *Gyűrűk ura*-filmeknek és az androidos telefonnak. Ezek a klaszterek tehát önmagukban alkalmasak annak kimutatására, hogy egy bizonyos ízléscsoportba tartozó személyek milyen egyéb tevékenységet, jelenséget, eszközt vagy terméket kedvelnek.

A BME-s klaszterből például, amely vélhetően műszaki egyetemisták szokásaiból állt össze, megtudjuk, hogy (a műszaki egyetemhez köthető különféle linkek mellett) kedvelik a műszaki egyetem közelében található hentest,[11] és pizzát a pizzásh-tól[12] rendelhetnek.

Napasztmek

Hogy az eljárás szociológiai relevanciáját is kiemeljem, érdemes megnézni az egyik „meglepetésklasztert", amit én „napasztmek"-klaszternek neveztem, de valójában lehetett volna „fatyok" is a neve.

2. ábra. „Napasztmek"-klaszter URL tagjai a LInKE-rendszerben

Forrás: www.li-n-ke.com/napasztmek

10 A legfontosabb node-ok nem feltétlenül azonosak a leginkább like-olt témákkal, a fontosságot többféle mutatóval lehet mérni. Mi a *betweenness centrality* mutatóját használjuk. Ez az érték azt mutatja meg, hogy a hálózatban mely ponton keresztül fut a legtöbb legrövidebb út két tetszőleges pont között.

11 https://www.facebook.com/pages/A-Hentes/119239174753603?sk=timeline.

12 http://pizzasch.sch.bme.hu/main/order.

Ahogy az a *2. ábrá*n látható legtöbb bejegyzésből látszik, ez egy olyan témacsoport, amelyben olyan weblapok, leginkább Facebook-oldalak szerepelnek, amelyek szerzői a magyar nyelvet sajátosan használják: a betűket úgy keverik össze, hogy az még viccesen értelmezhető legyen. Ilyen oldalt már láthatott bárki, én is találkoztam ilyennel korábban, de azt, hogy ez egy sajátos szubkultúra, csak az elemzés mutatta meg. Hiszen azzal, hogy klaszterré összeállt ez a témacsoport, az derült ki, hogy bizonyos emberek számára ez egy sajátos stílusirányzat, amelyet különféle honlapokon előszeretettel like-olnak. Vagyis azért fontos az ilyesfajta feltáró eljárás alkalmazása Big Datá-s környezetben, mert így olyan eredményekhez jutunk, *amelyeket előre nem definiáltunk,* és nem is sejtettük, hogy léteznek. A normál kérdőíves kutatásokon alapuló klaszterező eljárásokhoz képes két fontos tényezőt kell említenünk, amely miatt ott erre az eredményre nem tettünk volna szert.

Egyrészt ahhoz, hogy egy ilyen szubkultúrát bemérjünk, tudnunk kell, hogy maga a szubkultúra létezik. Emellett azt is tudnunk kell, hogy mely honlapok kötődnek hozzá, és ebben az esetben sok ilyen honlapra rákérdezve[13] majdnem ki is lehetne mutatni, hogy a szubkultúra valóban létezik. Valójában azonban vélhetően még így sem tudnánk kimutatni, ugyanis ha ez valóban egy kisebb csoportot érintő sajátos szemléletmód, akkor egy 1000 fős mintába nem is kerül be kellő mennyiségben olyan ember, akik révén a klaszter létrejönne. Tehát még egyszer: a hagyományos kérdőíves módszer problémája, hogy nem rendelkezünk az e csoport felleléséhez szükséges előzetes tudással, másrészt, ha rendelkeznénk is, nem tennénk szert kellő mennyiségű adatra.

Ezért nem mutathatunk ki ilyen és ehhez hasonló szubkultúrákat tudatosan megtervezett kutatásokkal. A „Napasztmek" témához hasonlóan érdekes, előre nem várt témacsoportot többet is kaptunk: slammer csoportot, gördeszkás szubkultúrát, olyan zenekarok rajongóit, akiknek még csak a stílusát sem ismertük stb. És ezzel át is tértünk az adatbányászati projektek másik fontos természetére: nem várt eredményeket szülnek, és ezzel olyan tudással látják el a társadalomtudományt, amelyet az máshogyan nem, vagy csak nagy időkéséssel tudna megszerezni.

Ultrák

Egy másik példám egy fociultrákhoz köthető klaszter. Tudni kell, hogy a futballrajongók között az ultráknak nevezett drukkerek egészen sajátos, rejtőzködő életmódot folytatnak. Az ultráknak ugyanis nem áll érdekükben, hogy mindenfelé (például a Facebookon) hirdessék ultra mivoltukat, az ugyanis sokszor olyan illegális cselekedetekkel is összefügghet, amelyeket a rendőrség üldöz. Így az ultra érzelmű emberek közül sokan kifejezetten visszafogottan like-olnak ultraoldalakat, tartva egy hatósági beazonosítástól. Éppen ezért volt érdekes, hogy rendszerünkben több ultracsoport is létrejött, és az egyik egy vidéki város szurkolóihoz kötődött.[14]

Ezen URL-csoport kapcsán már önmagában az is érdekes volt, hogy egyáltalán létrejött, hiszen az érintett tagok, mint mondtam, ritkán teszik közzé, milyen akciókban vettek részt.

13 „Tetszik Önnek ez a honlap, vagy ennek a honlapnak a tartalma?"

14 A város nevét adatvédelmi okokból nem közöljük.

Ugyanakkor mégis létrejött az URL-hálózat, mert más tevékenységükben mégiscsak hasonlítottak egymásra a tagok, így tulajdonképpen – mint Dessewffyék példájában – a rendszer anélkül azonosította be őket, hogy az mindenki számára nyilvánvaló lett volna. Külön érdekesség, hogy ezen URL-csoport középpontjában milyen webcím volt található: a város egyik nightklubja.

Egyéb érdekességek

Spammelő oldalak

A rendszer a fenti – leginkább ízlésmintázaton alapuló – csoportok mellett még nagyon sok érdekes csoportot is beazonosított. Így például igencsak meglepetésszerű volt, hogy külön URL-csoportba rendeződtek azok az oldalak, amelyek arra épülnek, hogy rendkívül bulváros, megbotránkoztató vagy undorító tartalmakat kínálnak, ha azonban a felhasználó átugrik az oldalukra, akkor már maga is csak úgy tud kattintani, hogy az oldalt like-olja, és ezzel megosztja. Ilyen „cseles" oldallal bármelyikünk találkozhatott Facebookozása során, leginkább egyszer. Mert az első ilyen esemény után egy átlagos felhasználó észreveszi a csapdát, óvatosabbá válik és többször nem lép bele. Ennek ellenére vannak olyan – mondjuk így: naiv – emberek, akik ezeknek a linkeknek többször is beugranak, így jött létre egy olyan URL-klaszter, amelyben egymás mellett megtalálhatók a rosszindulatú, ezért kiszűrendő oldalak. Úgy döntöttünk, hogy az eljárást magát arra fogjuk használni, hogy ezeket az oldalakat szűrjük, és ehhez egy olyan „ízléscsoport" tevékenysége járul hozzá, melynek tagjai nem kellőképpen óvatosak internetezésük során.

Facebook-akciós oldalak

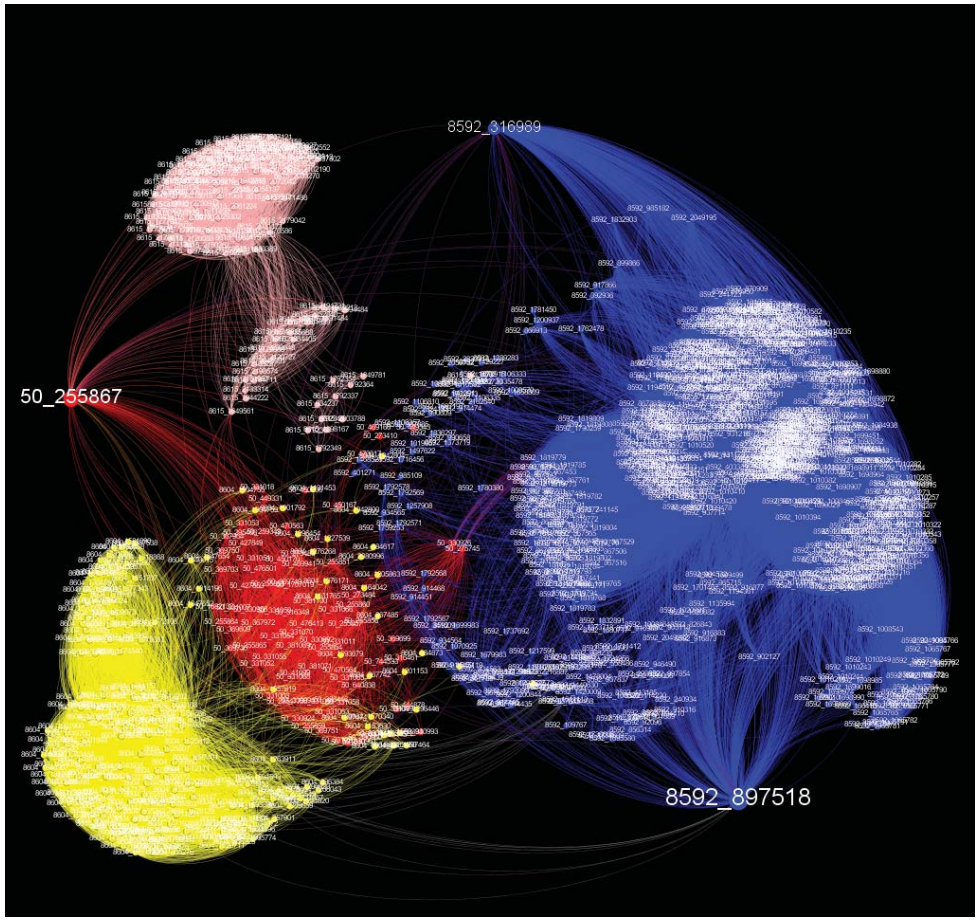
Szintén külön kiugrottak azok az oldalak is, amelyek a Facebookon nyereményjátékokat hirdetnek. E csoportba csupa olyan link tartozik, amelyen arra buzdítják a felhasználót, hogy kommenteljen a poszt alá, ezzel is terjesztve a nyereményjátékot. Nos, mint magából a csoportból látszik, vannak olyan felhasználók, akik előszeretettel vesznek részt nyereményjátékokban, és ezt felfoghatjuk ízlésnek is, valójában egyfajta beállítódást jelez. Így tulajdonképpen megint csak egy újabb szűrési lehetőséget hoznak létre: rendszerünkből az ő tevékenységük (ízlésük, habitusuk) révén tudjuk szűrni a nem releváns tartalmat. Ugyanakkor látni kell, hogy ha valaki szereti a nyereményjátékokat, akkor tartalmi ajánlásként is örül neki, így azoknak, akik e csoportban érintettek, rendszerünk továbbra is megmutatja, milyen más nyereményjátékok futnak aktuálisan.

Négy gamerklaszter összefüggéseinek elemzése

Végül még két példát, illetve elemzési lehetőséget villantanék fel, amelyek több csoport egymáshoz kapcsolódásának kimutatására épülnek. A *3. ábrá*n négy olyan klaszternek a linkjeit ábrázoljuk, amelyek mindegyike online vagy pedig konzolos játékokkal kapcsolatos.

Anélkül, hogy a részletesebb elemzésbe belemennénk, az ábrára nézve két dolgot vehetünk észre. Egyrészt jól látszik, hogy mindegyik URL-csoportnak[15] van egy belső szerkezete, így meghatározható, mely URL-ek játszanak döntő szerepet egy-egy klaszterben. Másrészt ebből a „vizuális elemzésből" is látszik, hogy vannak olyan URL-ek, amelyeket többféle gamercsoportba, gamerstílusba tartozó személy is like-olt.

3. ábra. Négy gamerklaszter egymáshoz való viszonya a LInKE hálózatában



Forrás: saját ábra, készítette Vassy Zsolt

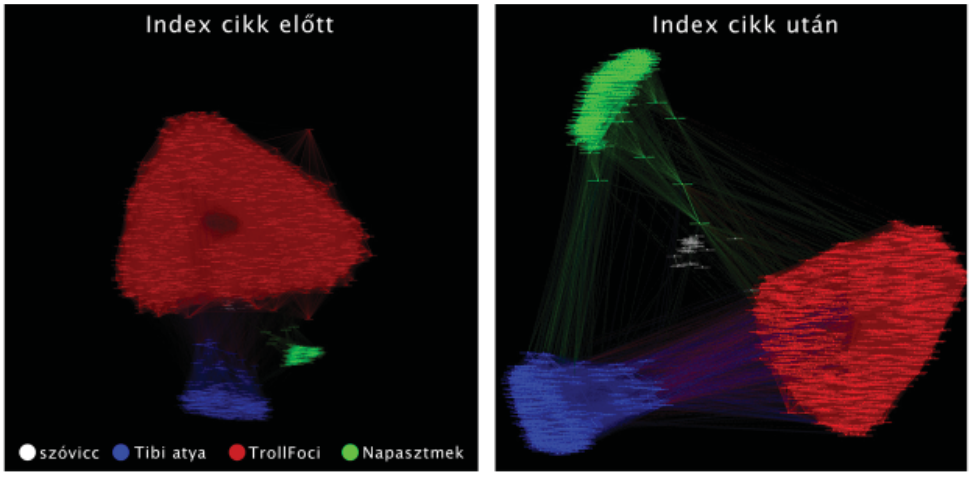
Egy ilyen elemzéssel tehát ki lehet mutatni, mely elemek kötnek össze különféle típusú, de egyaránt ellenzéki csoportokat, mi az összekötő elem az eltérő irányú futballrajongói cso-

15 A csoportokat eltérő színekkel ábrázoltuk. A pontok itt is az URL-ek jelzik, a vonalak pedig azt, hogy mely URL-eket like-olták egyaránt bizonyos felhasználók. (A színek a nyomtatott a folyóirat nyomtatott változatában a szürke árnyalatait jelentik. *A szerk.*).

portok között, a különféle autóimádók (rally, smoke, old timer, sebesség) körében milyen oldal az, amelyet egyaránt látogatnak, becsülnek, és így tovább. Vagyis egy ilyen elemzés tökéletesen ki tudja mutatni azokat a „modus vivendi" oldalakat, amelyek sokrétű, de egyazon érdeklődésű csoportokat összekötnek.

Egy másik kíséreltünkben az index.hu-val együttműködve megvizsgáltuk, hogy egy cikk megjelenése előtt és után hogyan alakul négy viccklaszter egymáshoz való viszonya. Az alábbi ábrán az látszik, hogy a cikk megjelenése után a trollfocis, szóvicces, Napasztmekes és Tibi Atyás viccek hálózatában az indexen favorizált szóviccek csoportja hogyan vált el a nagy, piros színnel jelzett Trollfocis klasztertől, illetve hogy milyen összekötő elemek jöttek létre az amúgy meglehetősen elkülönülő klaszterek között.

4. ábra. Négy viccklaszter egymáshoz való viszonya a LInKE-rendszerben



Forrás: saját ábra, készítette Vassy Zsolt

Lehetőségek

A nagy adatbázisok elemzésében elképesztőek a lehetőségek. Ismerek olyan példát, amikor egy banknak a kártyatranzakciós adatait elemezték hasonló, hálózati, klaszterező algoritmusokkal. Ott is, mint a LInKE-ben, a rendszer semmi mást nem tudott a felhasználóról, csak azt, hogy melyik kártyához mikor, milyen összegű vásárlás kötődött. A hálózati eljárás végigfuttatása után alapvetően két nagy csoport bontakozott ki a kártyahasználók között. Amikor megnézték, hogy mi jellemző a két csoportba tartozó kártyaügyfelekre, milyen egyéb tulajdonságban egyeznek meg (amit az eredeti elemzés nem ismert, mégis kimutatott), akkor kiderült, az egyik csoportba azok tartoznak, akik amúgy jó adósai a banknak, a másikba pedig a rossz adósok (a fizetéssel néha megkéső, rosszul teljesítők). Itt is, mint más esetekben, egy bizonyos típusú adatmintázat (vásárlások mintázata) kvalitatív különbséget mutat ki egy másik területre vonatkozóan (jó adósság, rossz adósság).

Anélkül, hogy sokat kockáztatnánk, megjósolható, hogy például ha a gyógyszerfelírásokkal kapcsolatos adatokat vinnénk be egy hálózati, klaszterező rendszerbe, akkor ki lehetne mutatni, kik azok az orvosok, akiket valamilyen gyógyszergyár arra motivál, hogy az ő gyógyszereiket írják fel a betegeknek, vagyis kiket korrumpálnak a gyógyszergyárak. Ez vélhetően egyszerűen már egy vizuális elemzésben is megmutatkozna, mint amire példát egy korábbi kutatásunkban is mutattunk (Gayer és Balogh 2011).

A példák száma végtelen lehet, így például úgy gondolom, nyilvánvaló, hogy megfelelő adat-előkészítés után egy rendőrségi adatbázist felhasználva a klaszterező eljárással ki lehet mutatni a bűnelkövetők helyét vagy egyéb közös tulajdonságaikat, ami az elkövetők fellelését segítené stb.

A LInKE rendszerben tapasztalt lényegkiemelés után azt mondhatjuk, hogy magukra a hálózatelemző, iteratív algoritmusokra jellemző, hogy ezek révén szinte bármilyen, hálózatba rendezhető adattömegből kijön a lényeg, a titok.

Távlatok

Az eddig felvázolt lehetőségekben összességében az az igazán ijesztő, hogy a dolog nem áll meg egy adatbázisnál, hanem adódhat a lehetőség, hogy adatbázisokat akár össze is kapcsoljunk. Mi lenne, ha például egy ízlésmintázatot kimutató elemzés mellé bekapcsolható lenne, hogy merre fog járni az illető adott időpontban, például egy másik, mondjuk telefonos adatokon alapuló adatbázis alapján (lásd például Barabási 2010). Minél több rendszert kapcsolunk össze, a nagy rendszer „intelligenciája" annál fejlettebb lesz. Meg tudjuk mondani, hogy merre fog járni, ott mit fog nézni, arra hogyan fog reagálni. Ez tehát az egyénre szabott ajánlatok és reklámok helye és még ezer más mindené.

De ha még tovább megyünk ennek a kérdésnek a végiggondolásában, feltehető a kérdés, hogy mikor jön létre egy olyan gigantikus adatbázis, amely a teljes világunkat értelmezi. Mi lesz ennek a határa? Mi az akadálya annak, hogy hálózatelméletileg értelmezzük az összes kérdésre adott összes választ, amelyet az interneten el lehet érni? Ha ez megtörténik, akkor a nyelvi elemzés kihozhat egy Turing-teszten átmenő intelligenciát? Hiszen minden algoritmizálható. A kontextus, a korábbi kérdések, az így létrejövő „hangulat". Mi a határ? Jelenleg azt gondolhatnánk, hogy a gépek mérete, hiszen nincs olyan gép, amely egy ekkora adatbázist értelmezni tudna, és itt jön a képbe Moore törvénye.[16] Ha a Moore-törvényből indulunk ki, akkor azt gondolhatnánk, hogy az ilyen számítógépek létrejötte még messze van. Azonban látni kell, hogy ezeknek az eljárásoknak az ereje már a gépek egymáshoz kötöttségében van, vagyis azt gondolhatjuk, hogy a mesterséges intelligencia a gépek egymáshoz kapcsolásából jöhet létre.

16 Gordon E. Moore jóslata 1965-ből még arról szólt, hogy a gépek teljesítménye másfél évenként megduplázódik, amit 1975-ben módosított úgy, hogy kétévénte történik meg a duplázódás (lásd: https://en.wikipedia.org/wiki/Moore's_law). Ezt a törvényt az azóta eltelt idő igazolta, a gépek CPU-teljesítménye és háttértároló képessége az évek során nagyjából ilyen ütemben fejlődött. Így tulajdonképpen akár kiszámítható is lenne, hogy a gépek mikor érik el azt a szintet, hogy ki tudjanak számolni ilyen bonyolult kérdéseket.

Hivatkozott irodalom

AAPOR (The American Association for Public Opinion Research) (2011): *Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys.* Interneten: https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR_Standard-Definitions-Final-Dispositions-of-Case-Codes-and-Outcome-Rates-for-Surveys.pdf (letöltve: 2015. 06. 15.).

Barabási Albert László (2010): *Villanások*. Budapest: Libri.

Dessewffy Tibor és Láng László (2015): Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika* (92–93): 157–170.

Gayer Zoltán és Balog Barabás Tibor (2011): Adatbiztonság, adattudatosság a közösségi hálózatokban. *Médiakutató* 12(3). Interneten: http://www.mediakutato.hu/cikk/2011_03_osz/01_adatbiztonsag_adattudatossag (letöltve: 2015. 06. 15.).

Veres Zoltán, Andics Jenő, Hetesi Erzsébet, Kovács Péter, Prónay Szabolcs és Vajda Beáta (2010): Életstílus alapú fogyasztói szegmentumkutatás. In *Életstílus alapú fogyasztói szegmensek Magyarországon*. Veres Zoltán (szerk.). Szeged: Szegedi Tudományegyetem, Gazdaságtudományi Kar, Üzleti Tudományok Intézete, 7–190.