Tartalomajánlás felhasználói profilok alapján – automatizáltan

(Content sharing based on user profiles – automated)

Pitlik László, Apertus Nonprofit Kft.

Kivonat: A felhasználók leíró adatai, bármiről szóljanak is ezek, alkalmasak arra, hogy homogén csoportokat lehessen képezni tetszőleges leíró változó egyidejűsége mentén. Amennyiben az ajánlórendszer alapparamétere szerint a cél a csoport-homogenitás növelése, akkor univerzális alapvetésként fogadható el, hogy egy adott személynek adott csoportba tartozás esetén azt a tartalmat kell ajánlani, ami a csoport tartalomfogyasztásának toplistáján a legelőkelőbb helyen áll és az adott felhasználó nem szembesült még az adott tartalommal. Több felhasználói csoport kialakítása alternatív tartalmakat vethet fel, melyek eredőjéhez be kell vezetni a homogenitás index fogalmát, vagyis a csoporttagokra vonatkozó tartalomhasználati arányt. Minél nagyobb ez az indexérték, annál előkelőbb helyre kerül egy ajánlandó tartalom. A leírt folyamat teljes mértékben automatizálható. Az anonimizált demo-rendszer Excel kimutatás-varázslás keretében jött létre.

Kulcsszavak: pivot, olap, csoportképzés, homogenitás index, sorrend

Abstract: The dimensions of the user profiles (whatever these dimensions describe) are capable of building homogeneous groups using the dimensions for filtering users. If content sharing should maximize homogeneity within groups, then such kind of contents should be shared/recommended, which characterize the most members of a group. The top lists concerning user groups can be derived through simple SQL/statistical steps. A given user can be member in different groups, therefore this user can have different contents for recommendation. The different sources for recommendation should be joined into a common set. However, it is necessary to create a homogeneity index to ensure, that the ranking of the element from the top lists for different user groups can be calculated. This kind of estimation of sharable contents can be automated entirely. The demo solution has been executed in Excel based on pivot reports and standard funcions.

Keywords: pivot, olap, building of groups homogeneity index, ranking

# Bevezetés

A kapcsolódó fogalmak közötti rendteremtés érdekében a tartalomajánlást elsőként érdemes megkülönböztetni a kereséstől: a keresés az az aktív, a felhasználó által akaratlagosan felvállalt aktivitás, melyben a tudatosan kiválasztott kulcsszavakhoz illeszkedő tartalmak valamilyen (választott/default) prioritási elvek mentén adott sorrendben felkínálásra kerülnek. Ezzel szemben az ajánlás az ennek megjelenéséig alapvetően passzív célszemély számára azon (logikusan a felhasználó érdeklődésére számot tartható) tartalmak adott, a rendszer (az ajánló) által kialakított prioritási elvek szerinti sorrendben való megjelenítése, melyekre a felhasználótól valamilyen reakciót várunk el (lájkolás, megtekintés/letöltés, stb.). Az ajánlás akkor jó, ha az elvárt reakciókat váltja ki a címzett felhasználóból. A keresés akkor jó, ha az aktív felhasználó saját maga számára az elégedettség felé képes elmozdulni. Az keresés és az ajánlás tehát alapvetően különbözik a felhasználó kezdeti aktivitási szintjében és az eredmény objektivitását illetően, ahol is a keresés elsődlegesen szubjektív elégedettségre vezet, míg az ajánlás sikere objektíven értelmezhető, amennyiben a felhasználó spontán reakcióinak és az elvárásoknak az egybeesését tekintjük sikerkritériumnak. A keresés kapcsán is lehet objektív sikerességről beszélni, amennyiben világos kérdésekre a lehetséges válaszopciók közül a legvalószínűbb (leghelyesebb) válasz racionálisan bizonyítottnak vélhető fellelése megtörténik.

Mindkét esetben lehet az időtényezőt további értékelési finomhangolásra használni: ugyanis nem mindegy, mikor váltódik ki az ajánlás megjelenéséhez képest az elvárt reakció, ill. mennyi idő kell a keresés során adott kérdés megnyugtató megválaszolásához.

Ez a felvezetés mutat rá arra, hogy a BOSTROM[[1]](#footnote-1) (2015) által megfogalmazott gondolat (ajánlás), miszerint bármilyen rendszer befolyásolása előtt a lehetséges állapotok értékelését kell tudni automatizálni, ebben az esetben is igaz: addig nem érdemes a KNUTH[[2]](#footnote-2)-i elvet követni és a tudást forráskódba konvertálni, amíg nem beszélhetünk érdemi hatásossági, majd erre alapozva hatékonysági állapotváltozásokról, melyek alapján a rendszerállapotok értékelni tudása adja.

Az ajánlás jelensége kapcsán az „érdeklődésre számot tarthatóság” képezi jelen cikk fókuszát. Az ajánlások minőségét az utólagos visszamérhetőség objektív értékei adják. Mivel felhasználói reakciók sokfélék lehetnek, így ezek értékelése is alapvetően több-dimenziós illik, hogy legyen:

* minél nagyobb a lájkolás aránya az ajánlott objektumokhoz képest, annál jobb az ajánlórendszer
* minél nagyobb a letöltések aránya az ajánlott objektumokhoz képest, annál jobb az ajánlórendszer
* minél gyorsabban következik be a lájkolás a rendszerbe való belépéshez képest, annál jobb az ajánlórendszer
* minél gyorsabban következik be a letöltés a rendszerbe való belépéshez képest, annál jobb az ajánlórendszer
* minél pozitívabb a hangvétele (vö. szövegelemzés) az esetleges megjegyzéseknek (egy fórum-szerű keretrendszert feltételezve), annál jobb az ajánlórendszer
* minél kisebb a dislike-ok száma (ennek létét feltételezve), annál jobb az ajánló rendszer
* minél jobban közelíti a felhasználói aktivitás nélkül eltelt idő hossza egy adott ajánlott tartalom default elfogyasztási idejét ennek letöltésétől számítva, vagyis minél inkább vélelmezhető az ajánlott tartalom tényleges befogadása, annál jobb az ajánló rendszer
* …

A fenti lista értelemszerűen quasi tetszőlegesen folytatható. A több dimenziós kiértékelés alapját pl. az anti-diszkriminatív hasonlóságelemzések[[3]](#footnote-3) adhatják online automatizmusként.

# Csoport-homogenitás maximalizáló ajánlórendszer

A fenti anti-diszkriminatív értékelés elvét elfogadva a következő kihívás az ajánlhatóság mérhetőségének megteremtése. Ennek egyik triviális alapvetése a csoportképzés és a csoport-homogenitás maximalizálása. Csoport alatt értendő a teljes felhasználói halmaz bármilyen nem véletlenszerű elv mentén való min. két részhalmazra bontása (pl. nemek: férfi vs. nő, korosztályok, stb.). A homogenitás ebben a kontextusban nem más, mint annak célként való elismerése a mindenkori ajánlórendszer-üzemeltető által, hogy az egy csoportba tartozók minél inkább ugyanazon tartalmakat fogyasszák. (Ezzel az elvvel lehet természetesen vitatkozni[[4]](#footnote-4), de ezen stratégiai paraméter kiragadása a jelenlegi szubjektív döntéshozói preferenciák köréből, s átemelése az objektíven vizsgálható stratégia-sikeresség területére ennek a cikknek nem alapfeladata. Ellenben a fentebb felvázolt objektív értékelési rendszer bármilyen ajánlási logika értékét másokkal szemben félreérthetetlenül képes lesz levezetni…)

Csoportot egy adott felhasználó-halmaz leíró statisztikái alapján szinte korlátlan szempont (dimenzió) mentén lehet készíteni (vö. budapesti születésű nők). Az egyre több szempont által jellemzett csoport felhasználólétszáma egyre kevesebb, meredeken közelít a nullához. Vagyis a mennyiség átcsap minőségbe elv egyre kevésbé teljesül. Bár a csoport-homogenitás index önmagában méretfüggetlenítésre képes, vagyis azt jelenti, hogy egy adott tartalom adott csoport tagjai által milyen arányban került „fogyasztásra” már, mégis igaz, hogy azonos homogenitási index esetén további rangsorolási elvekre van szükség (vö. pl. távolugrás-verseny azonos magasságot elért versenyzői közötti különbségtétel szabályrétegei: hányadik kísérletre sikerült, mi volt az előző sikeres magasság, stb.).

A homogenitás index mögött tehát a következő sorrendezési hierarchiaszinten a számosság áll, vagyis a homogenitás % mekkora felhasználóhalmazon állt elő. Ha az azonos homogenitású halmazok elemszáma is azonos, akkor a következő sorrendezési szempont lehet pl. a homogenitás-index trendjének alakulása: minél meredekebben tart a trend az éppen ismert utolsó értékéhez, annál intenzívebb a homogenizálóság, tehát annál inkább ajánlásra méltó az egyik hír a másikhoz képest.

# Demo rendszer

## Adatbázis

Tegyük fel, hogy az alábbi adatvagyon[[5]](#footnote-5) létezik[[6]](#footnote-6):

* objektum: egy-egy felvett képzés neve, id-ja
* a tanuló/kiképzendő azonosítója
* a tanuló/kiképzendő neme
* a tanuló/kiképzendő születési éve
* a tanuló/kiképzendő eddigi munkahelyeinek száma
* …

A rekordok száma legyen 100.000+, a felhasználók száma legyen 10.000+, a képzések száma 100+.

A csoportképzés értelemszerűen kiterjed a nemekre, a születési évekre vagy ezek intervallumaira, ill. a munkahelyek számára (ezek intervallumaira) – valamint ezen opciók potenciális kombinációira.

## Top-listák képzése

Ahhoz, hogy megtudjuk, vajon pl. a nők mely képzéseket vették fel eddig a következő riportra van szükség, mely kimutatás-varázslással Excel-ben előállítható:

* sorok: képzések neve
* oszlopok: nemek (ezen belül is a nők)
* cellák: a rendelkezésre álló rekordok száma tárgyanként és nemenként

A tovább-feldolgozás keretében a nők oszlopában lévő érintettségi értékek sorszámozása következik. Ennek ismeretében a képzések sorba rendezése FKERES() támogatással lehetséges, amennyiben a sorszámokat kézzel definiáljuk pl. 1-től n-ig (ahol n a képzések száma), majd a számított sorszámok között keressük a definiált sorszámokat, s az FKERES() segítségével visszaadjuk a képzések nevét ezek előfordulási sorrendjében. Az így keletkező univerzális riport és függvény-alapú sorrend, ill. top-lista készítő megoldásban az egyes szűrési feltételek (pl. nemek, születési évek, munkahelyek száma, stb.) tetszőleges variációban beállítható és a toplista vizuálisan értelmezhető. Az univerzalitás lényege: bármilyen leíró dimenziók is álljanak rendelkezésre a felhasználókról a csoportképzés jelensége context free módon értelmezhető, vagyis a toplista fogalma is mindenkor létezni fog.

## Egyéni ajánlásokhoz vezető lépések

Amennyiben egy adott személy számára konkrét ajánlásokat kívánunk tenni a csoport-homogenitás maximalizálásának célját követve, akkor a fentebb leírt univerzális top-listakészítő megoldást további statisztikákkal és index-számításokkal kell kiegészíteni.

Statisztikák:

* minden egyes csoport méretét ismerni kell, vagyis a nők csoportjának példájánál maradva tudni kell, vajon mennyi a nők létszáma, amit egy olyan riport képes a kimutatás-varázslás keretei között megadni, ahol
  + a sorok a felhasználó-azonosítók,
  + az oszlopok a nemek,
  + a cellák az adott felhasználó által felvett képzések száma
  + majd a nők oszlopában meg kell számolni pl. a DARAB() függvény segítségével a nem üres, vagyis felvett képzésre utaló számokat tartalmazó cellákat
  + alternatív megoldásként megfelelő lehet a nők által felvett képzések száma, vagyis a nők oszlopában álló számok összege is
* hasonlóan kell eljárni a születési évek és a munkahelyszámok, stb. dimenziók esetében is annak érdekében, hogy a toplistákhoz tartozó mérőszámokat relativálni tudjuk az összes tárgyfelvétel és/vagy az érintettek maximális létszáma alapján

Indexszámítás:

* adott toplistán szereplő adott képzés előfordulásainak számát megadó értéket el kell osztani pl. az összes képzés számával, ami így egy 0 és 1 közé eső indexértéket ad, mely egy fajta homogenitás-indexként értelmezhető,
* mely minél nagyobb, annál inkább igaz, hogy a csoportba tartozók egyre nagyobb arányban szembesültek az adott képzéssel,
* s amennyiben ennek értéke 1, akkor ez a toplista-elem ajánlásra nem alkalmas, mert már mindenki elfogyasztotta az adott tartalmat…

Az adott felhasználó vélelmezhetően valamennyi képzésen már részt vett (ugyanis ahhoz, hogy ebben az adatbázisban jelen legyen, legalább 1 képzést fel kellett vennie). Így a dimenziónkénti toplisták képzései közül azokat, melyek már az adott felhasználó által felvételre kerültek, ki kell zárni. A fennmaradó toplista-elemek a fentebb bevezetett indexértékkel jellemezhetők és ezek alapján, mint elsődleges rendezési szempont alapján sorba rendezhetők. Az indexérték kifejezésének pontossága (pl. csak egész százalékértékek számítása) alapján holtversenyek alakulhatnak ki, melyek feloldására az ismét csak fentebb már bemutatott segéd szabályok (rendezési elvek) kerülhetnek beválasztásra: pl. a százalékos homogenitás index mögötti gyakorisági értékek abszolút nagysága.

Az egyéni ajánlási preferencia-lista első eleme tehát a legnagyobb homogenitás-indexértékkel rendelkező képzés lesz a csoportképzésektől függetlenül. Holtverseny esetén a nagyobb abszolút előfordulást mutató csoport toplista eleme lehet a döntéshozás alapja. De éppen így további sorrendiség-meghatározást támogató adat lehet: az adott toplista-elem előfordulásának száma a dimenzióktól függetlenített halmazban. Hiszen egy adott képzés számos szempontból lehet előkelő helyen, így egy adott képzés több szempontból is bekerülhet pl. a top 10-be. Emellett a top 10-be kerülés mögötti sorszámok átlaga, szórása is sorrendképzési szempont lehet. Vagyis minél nagyobb az átlag, a maximum, a minimum a többször is a top 10-be került képzések esetén, annál előkelőbb helyre való az adott képzés neve a személyes ajánlási listán. A szórás értelemszerűen fordítottan hat: minél kisebb a szórás(a a bekerülési rangsoroknak), annál előkelőbb egy képzés ajánlhatósága.

Mint máris érezhető: a képzés-sorrend olyan sok szempont által befolyásolt jelenség, hogy előbb-utóbb fel kell, hogy merüljön a sorrend-hierarchiák önkényes kijelölése helyett egy anti-diszkriminatív hasonlóságelemzés futtatása, ahol a minden toplista-elem másként egyforma elv sérülése esetén a leginkább ajánlandó képzések sorrendje automatikusan számítható mindennemű emberi által szubjektíven beállított paramétersor nélkül…

Adott ajánló rendszer adott ajánlási terében vizuálisan adott mennyiségű ajánlandó objektum fér el, így a fentebb leírt top-lisztázás során a top x esetén az x nagysága az ajánlandó mennyiség és a felhasználni tervezett dimenziók függvénye.

# Jövőkép

## Még nem létező tartalom ajánlása

Amennyiben adott képzés (tartalom) még nem létezik, vagyis a fenti rendszernek nem eleme, akkor is lehetséges ennek ajánlása, de ez egy egészen más jellegű feladat, mint a tényadatok csoport-statisztikáinak és a homogenitás fogalmának felhasználásával keletkező megoldás.

A kihívás érzékeltetéséhez vegyük alapul a filmeket jellemző humor-érzelem-erőszak indexek rendszerét. Ha egy katalógusban 100 film létezik és a humor-érzelem-erőszak dimenziók pl. 10-10 értéket vehetnek fel, akkor 10^3 filmre van szükség minimum ahhoz, hogy minden egyes film-variációra legyen példa.

Ha egy-egy filmnéző kapcsán ismert az is, mennyire volt elégedett az adott filmmel, s a nézőről sok-sok leíró tulajdonság is ismert (neme, kora, stb.), akkor a sok személy által látott sok film alapján alkotható egy elégedettség termelési függvény, mely megadja: milyen személyek, milyen filmeket nézve mennyire illene, hogy elégedettek legyenek. Feltételezve ezen termelési függvény nagy illeszkedési pontosságát a valósághoz, adott személy(profil) és minden létező film(profil) esetén a várható elégedettség becsülhető. Ha a néző csak új, még soha nem látott filmprofilt szeretne nézni a jövőben az eddigi elégedettséget meghaladó elégedettség-becslés reményében, akkor a még nem látott filmprofilok közül kiszűrhetők azok, melyek ennek a feltételnek megfelelnek.

S itt zárul a kör: mert ezen filmprofilok közül lehet, hogy még egyetlen egy sem létezik, vagyis a konkrét nézőnek ki kell várnia, hogy a szóba jöhető profilok valamelyikének megfelelő filmet valamikor valaki leforgassa…

## Még nem létező csoport kezelése

Ha egy ismert adatvagyonban nem volt még pl. olyan, hogy fiatal, vidéki, férfi, több munkahellyel, stb. – s most megjelenik egy új (eddig nem ismert profilt megtestesítő) felhasználó, s nem akarjuk diszkriminálni azzal, hogy nem ajánlunk neki semmit az általa preferált ajánlási dimenziók mentén, akkor egy nem létező csoportra vonatkozó ajánlási kihívásnak kell megfelelni. (Itt jegyzendő meg: a felhasználó maga is bevonható az ajánlás alapját adó csoport-képző dimenziók és csoportok kialakításába!).

Ebben az esetben lehetséges ajánlási eljárás lehet, ha minden egyes, már ajánlást kapni tudó felhasználó hasonlóságát megkíséreljük kifejezni a szóban forgó felhasználóhoz képest, vagyis minden dimenzió kapcsán igyekszünk leírni a kiválasztott felhasználó és a több felhasználó adatainak távolságát (pl. születési évek távolságának abszolút értéke). Ezekre az eltérésekre a minél kisebb, annál hasonlóbb irányképző szabály vonatkozik. Egy, a távolságokkal jellemzett felhasználói halmaz adatai kapcsán futtatott anti-diszkriminatív hasonlóságelemzés képes a kiválasztott felhasználóhoz leghasonlóbbak felismerésére, majd a nekik ajánlott objektumok ajánlására.

Amennyiben több azonos mértékig hasonló felhasználó is van teljesen eltérő ajánlott objektumokkal, akkor egyéb adat híján ebből a halmazból már lehet választani véletlen szám generátorral is, de lehetséges az is, hogy pl. az egyes hasonló felhasználók egy-egy ajánlás kapcsán hasonlóságelemzéssel számított ajánlási indexét vesszük alapul, s ezek top listáját képezzük minden további modellezés nélkül.

## Tanácsadás – keresés – ajánlás

S végül, most, hogy már átlátható a keresés és az ajánlás különbsége, ill. a nem is létező tartalmak és csoportok kezelésének lehetősége az alapvetésként megfogalmazott csoportképzésre és homogenitás-maximalizálásra berendezkedő rendszer részletei mellett, még egy finomhangolási kihívással érdemes megismerkedni.

A keresés alapvetően egy kérdésre keres egy adekvát választ, s az ajánlás is lefordítható a legnagyobb valószínűséggel elfogadott ajánlás megtalálására. A tanácsadás fogalma egy komplexitás-növelt rendszer fogalma, ahol pl. a felhasználó kimondott/ki-nem-mondott kérdése a következő lehet: milyen képzések láncolata (tartalma, sorrendje) vezethet el egy kiindulási (pl. kompetencia)-állapotból és elvárt (kompetencia)-állapotba?

A keresés/ajánlás képes lehet ebben a komplexitásban a következő és/vagy egy/néhány hasznos elemet fellelni, de nincs semmilyen legitim tudáserőtér arra vonatkozóan, milyen láncreakciók milyen eredményre vezethetnek? Míg pl. a mezőgazdasági termelés kísérletek millióiból immár emberi beavatkozás nélkül képes pl. quasi teljesen szabályozott környezetben paradicsomot termelni, addig ezek a kísérletek a tanulási jelenségek körében quasi teljesen hiányoznak a log-ok kapcsán. Nem véletlen, hogy az MTA szakmódszertani pályázata mind a mai napig azt várja el, hogy adott didaktikai beavatkozás várható hatása előre becsülhető, ill. más megközelítésben bizonyítható legyen…[[7]](#footnote-7)

# Irodalom

…a kapcsolódó háttéranyagok hivatkozásai a szövegközben találhatók…

1. Mottó:…we should work out a solution to the control problem in advance… (Nick Bostrom) <https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are#t-945173> [↑](#footnote-ref-1)
2. [http://miau.gau.hu/miau2009/index\_tki.php3?\_filterText0=\*knuth](http://miau.gau.hu/miau2009/index_tki.php3?_filterText0=*knuth) [↑](#footnote-ref-2)
3. <http://miau.gau.hu/myx-free/> [↑](#footnote-ref-3)
4. <http://miau.gau.hu/miau/kofop/olap_challenges_and_hr_analytics_v1.docx> (cél a munkahelyi inhomogenitás növelésén keresztül a munkahelyek egyenszilárdságának mérése) [↑](#footnote-ref-4)
5. <http://miau.gau.hu/miau/234/olap_challenges_and_hr_analytics_v1.docx> [↑](#footnote-ref-5)
6. <http://miau.gau.hu/miau/234/szemelyes_adatok_kezelesenek_tipushelyzetei_v1.docx> [↑](#footnote-ref-6)
7. <http://miau.gau.hu/miau2009/index.php3?x=e0&string=MTA.szak> [↑](#footnote-ref-7)