Tartalomajánló rendszer finomhangolása hasonlóságelemzésekre támaszkodva

(Fine tuning of controlling system in content sharing/recommendation based on similarity analyses)

Pitlik László

Kivonat: A tartalomajánlás logikája a véletlen választástól a hasonlóságok legkülönfélébb formájáig szinte tetszőleges komplexitással alakítható ki. Az ajánlás mikéntje maga tehát egy tendenciájában egyre sikeresebben üzemelő ajánlórendszernek csak az egyik pillére. A másik pillér az ajánlások elfogadottságának mérésére alapozó finomhangoló vezérlés, mely által a statisztikák alapján egyre kevésbé elfogadott tartalmakhoz vezető logikák egyre hátrébb szorulnak a vizualizációs keretek által korlátozott, egyidejűleg megjeleníthető helyekért folytatott harcban. Míg a jobb elfogadottsági arányokat produkáló ajánlási szempontok egyre előkelőbb helyet kapnak. Ebben a megközelítésben egyelőre maga a tartalom semmilyen módon nem kerül elemzésre, ahogy az ajánlási logikák is lehetnek context free jelleggel kialakított megoldások.

Kulcsszavak: magatartásminta, ajánlási elv, anti-diszkriminációs modell

Abstract: Content sharing can be executed through a quasi-unlimited amount of methods (from random logic to analyzing e.g. similarities) having diverse complexity level. The recommendation principles are just one pillar of a content sharing system delivering tendentious better and better offers during its working process. The other pillar is a controlling subsystem ensuring that the news from the less acceptable methods for recommendations should have a decreased competition position, while the products from the better algorithms should have better and better ranking positions concerning the limited spots of a surface for visualizing recommendations. The recommended content as such plays no role in this context-free approximation.

Keywords: behavior pattern, recommendation principle, antidiscrimination model

# Bevezetés

Az immár több tanulmány alapját adó, több, mint 100.000 felhasználóról szóló adatvagyon alapján eddig felvázolásra került, miként lehet egy context free tartalomajánló rendszert kialakítani a személyek leíró attribútumai alapján, ahol a lényegi elemek a tetszőleges komplexitással kialakított csoportokba tartozás, mint triviális (többlépcsős, szűrés-alapú) hasonlósági alakzatok és a tartalomfogyasztási relatív gyakoriságok voltak. Az alábbi dokumentumok alapján a kiindulási helyzet minden releváns részlete értelmezhető:

* <http://miau.gau.hu/miau/234/szemelyes_adatok_kezelesenek_tipushelyzetei_v1.docx>
* <http://miau.gau.hu/miau/234/olap_challenges_and_hr_analytics_v1.docx>
* <http://miau.gau.hu/miau/235/content_sharing_v1.docx>
* <http://miau.gau.hu/miau/236/recommendation_controlling_v1.xlsx>

A teljes KÖFOP-projekt elemzés-módszertani dokumentációja itt érhető el: <http://miau.gau.hu/miau/kofop>

# Kulcsszavak

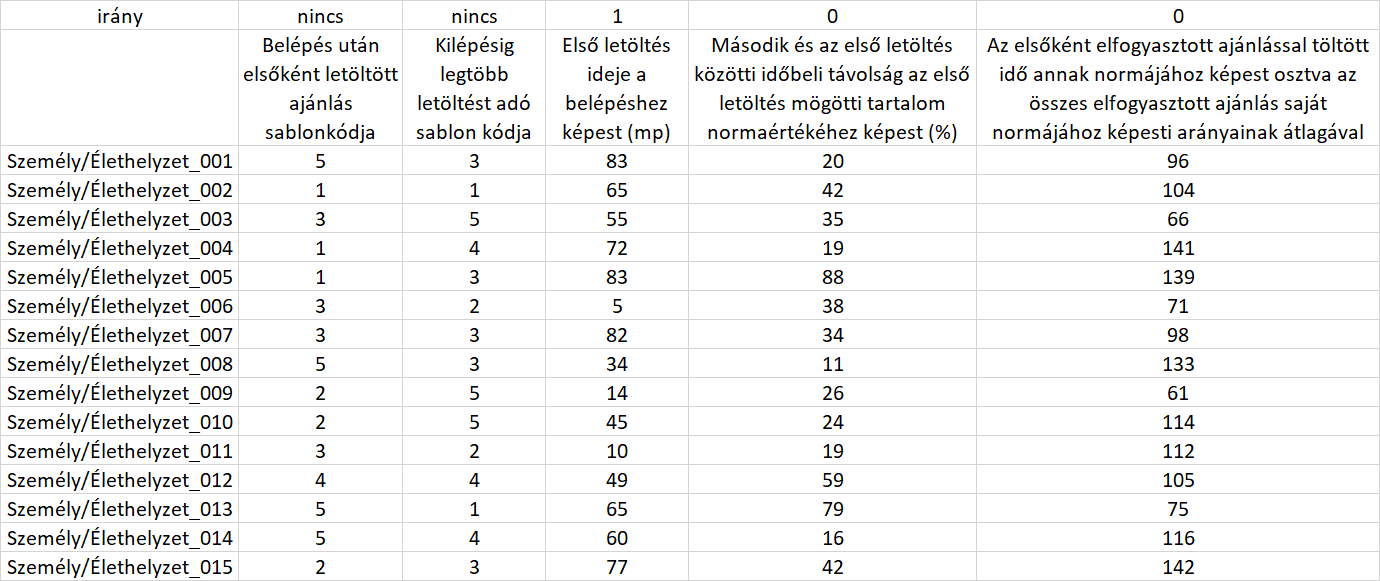
* Sablon: a tartalmak ajánlásához felhasznált módszertan, kezelés
* Tartalom: ajánlható objektum (alapesetben pl. videó) – szinonimái: hír, ajánlás
* Letöltés: a tartalom elfogyasztásának folyamata (pl. video-stream, kattintás)
* …

# Az adatvagyon

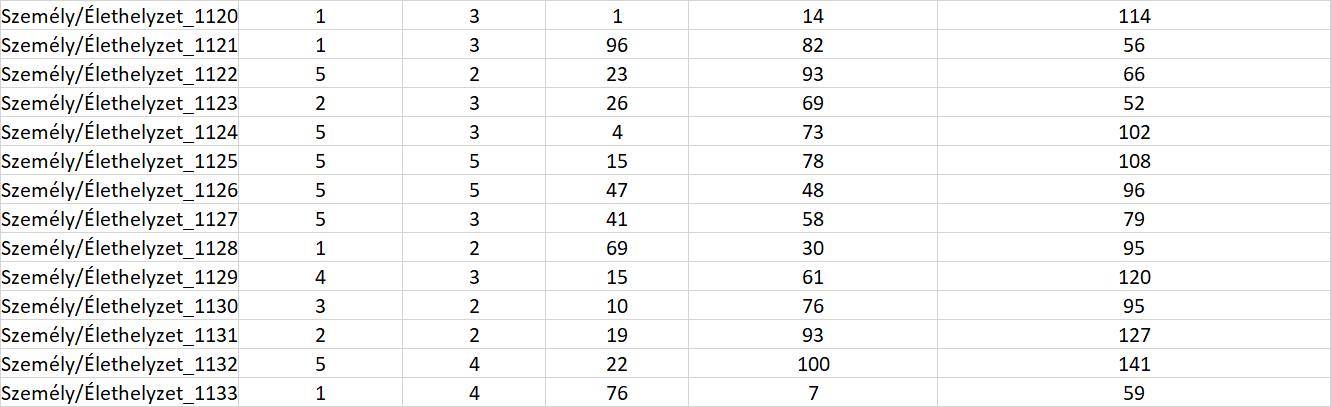
Mivel egy ajánlórendszer finomhangolásának tervezési fázisában valódi log-ok nem állhatnak rendelkezésre, így triviálisan véletlen számokkal kell bizonyítani, hogy egy adott adatáramlás/elemzéssorozat valóban képes racionális eredményekre vezetni.

Az adatvagyon tervezése egyben a loggolás és a log-előfeldolgozás (vagyis a mérési pontok és az adatelőkészítés) tervezése is. Jelen esetben gondolat-kísérletként induljunk ki az alábbi, tranzakciós szintű deklarációkból:

* minden egyes ajánlási „parancs” kapcsán a célszemély id-ja, az ajánlandó tartalom id-ja, az ajánlandó tartalom sorszáma a következő megnyitás utáni képernyőn, s az ajánlási sablon, módszer, elv id-ja tárolásra kerül az ajánlás keletkezésének timestamp adatával együtt
* speciális becslésként minden ajánlható tartalomhoz hozzá lehet rendelni egy normaidőt, mely ennek tényleg „elfogyasztásához” szükséges, s minden ennél rövidebb idő csak „belekóstolásnak” minősül (pl. videó-tartalmak esetén a videó hossza)
* speciális naplózást igényel az ugyanazon tartalmat érintő többszöri „belekóstolások” értelmezni tudása, vagyis annak levezetése, vajon ’n’ nekifutásból végül is a teljes tartalom megtekintésre került-e vélelmezhetően
* mivel nem lehet tudni, hogy egy adott felhasználó az utolsó, minden ajánlási pozíciót tartalommal lefedni képes ajánláshoz képest mikor lép be, így az ajánlások keletkezésének timestamp-je nem releváns adat a finomhangoláshoz
* ellenben az ajánlási pozíciókra vonatkozó adatok akkor is frissülhetnek, ha egy korábbi ajánlás belépés hiányában nem lett lehívva (s erről stratégiai döntés hozható, ahol természetesen az is egy ajánlás-módszertani kezelés, hogy a frissítés, vagy a kivárás vezet-e jobb ajánlás-akceptálási adatokhoz – de ez a megközelítés egy önálló cikket igényel a jövőben)
* naplózható emellett a felhasználói magatartás is, vagyis az, hogy mikor milyen ajánlási pozícióra/tartalomra kattintott a felhasználó elsőként, ill. ezt követően a tudatos vagy technikai kiléptetésig,
* a letöltési/megkattintási időpecsétek között eltelt idő alapján számolható az adott tartalommal való foglalkozás potenciális ideje, ahol a becsült normaidőhöz képest hosszabb idők nem jelentenek pl. többlet-fogyasztást - elemzői stratégiai döntés keretében fixálva ezen paramétert,
* mely természetesen szintén módosítható – hiszen a többletidő lehet elmélyült/elgondolkodó, azaz videót megállító, jegyzetelő, kommentáló, stb. idő éppúgy, mint a lefutó, eleve talán ténylegesen meg sem tekintett tartalom utáni felhasználói „szöszmötölés” ideje,
* ahol ismét csak ezen cikk fókuszán túlmutat a felhasználói magatartás reprodukálhatóságára törekvés a log-ban, s az így szerzett többletadatok alapján a ténylegesen tartalomfogyasztással töltött idő egyre pontosabb becslését lehetővé tevő módszerek kialakításának kihívása
* a nyers log-okból számítható az is, vajon az adott felhasználó általános (eleve norma-alapú) hírfogyasztási szokásaihoz (ennek átlagához) képest több, vagy kevesebb időt töltött-e el az adott hírrel
* a belépés és kilépés között eltelt (értelemszerűen az adott alkalmazás aktív jellegét feltételező, vagyis az aktív ablakban az ajánlást tartalmazó alkalmazást valóban használó) időben letöltött összes tartalomhoz tartozó ajánlás-módszertani id-k közül a legnagyobb arányt képviselő módszer kódja is meghatározható
* ahol holtverseny(ek) esetén több, az 1. ábrán látható rekord jön létre más-más id-val a harmadik oszlopban



…

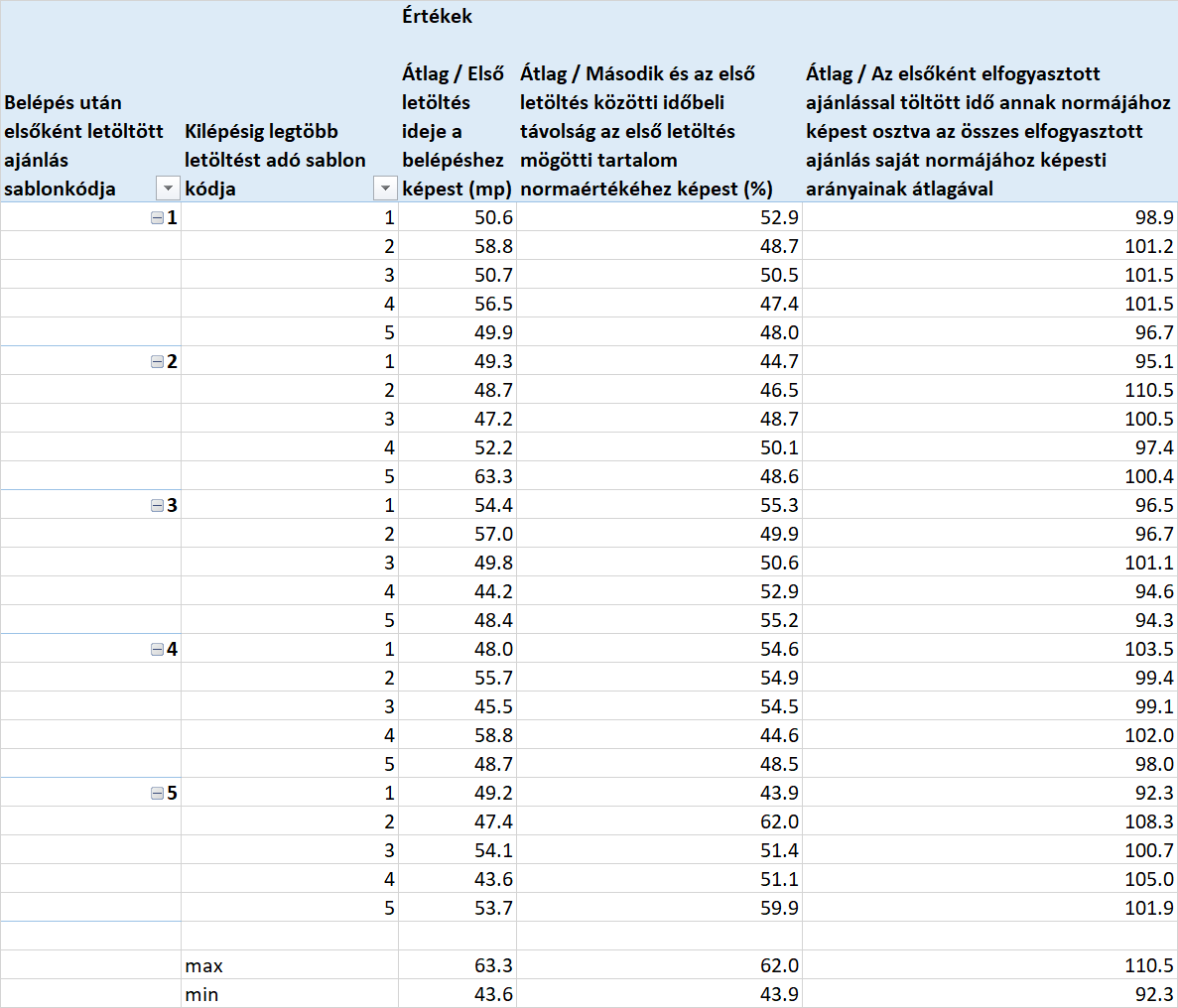


1. ábra: A nyers log-ok tovább-feldolgozásának eredménye (forrás: saját számítások)

Az 1. ábra alapján látható még, hogy

* az ajánlási sablonok kódjának nincs iránya, hiszen ezek önállóan, vagy kombinációban éppen azok az objektumképző jelenségek, melyek kapcsán az így kialakított objektumok anti-diszkriminatív versenye matematikailag értelmet nyer egy többdimenziós és esetegesen többkörös értékelési rendszerben
* az első tartalom minél gyorsabb lehívása értelemszerűen egy adott személy egy adott belépését követő (kilépéséig tartó) élethelyzetében annak a jele, hogy az ajánlott tartalom ténylegesen felkeltette az érdeklődését, akár már előre tudva másoktól (esetleg korábbi, megzavart, letöltéshez nem vezető belépéséből), hogy a tartalom létezik
* ahol a hír leírásának nem tartalmi, hanem pl. szövegbányászati meta-adatai, ill. az ajánlási pozíciók sorrendisége, stb. is majd egy önálló megközelítésben és így egy új cikkben kerülhetnek csak figyelembe vételre (pl. leírás hossza karakterszámban kifejezve a címre, a lead-re, a kép egérrel előhívható szövegére, a beszédes URL-re, stb. vonatkozóan)
* adott tartalom normaidejéhez képest eltöltött idő (vö. 0-100% között értelmezve – a túlfogyasztást 100%-ra korrigálva) minél hosszabb, annál inkább lett a tartalom tényleg végig befogadva
* s végül az 1. ábra jobb felső sarkában látható irány=0 kapcsán elmondható: minél nagyobb az adott személy átlagos, norma-alapú hírfogyasztási aránya, annál inkább figyel relatíve arra a tartalomra

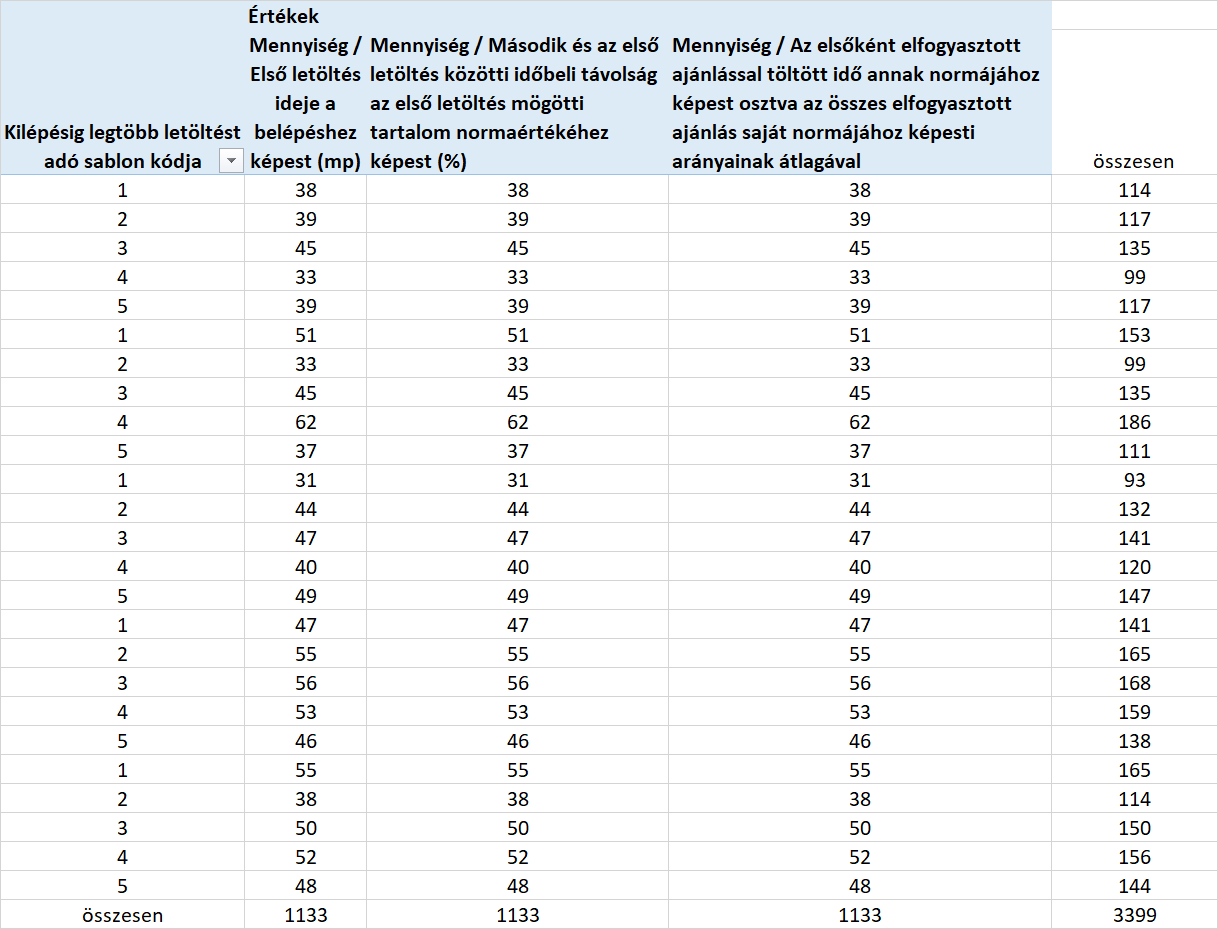
Az 1. ábra még nem az a nézet, melyre tanulási, elemzési folyamatokat lehet alapozni. Egy új nézet (OAM = objektum-attribútum mátrix) új objektumot vár el, s ez legyen az elsőként letöltött hír sablonkódjának és a csoportos letöltéseken belül győztes sablon kódjának kombinációja (jelen példában 1-től 5-ig, értelemszerűen mindkét, azonos értelmezési tartományt jelentő kódpozíció esetén, vagyis max 5\*5=25 objektumot elvárva az 1. ábra által jelzett 1133 darab rekord adataiból levezetve:



1. ábra: OAM1, vagyis a 25 objektumos tanulási minta átlagos értékekkel minden pozíció esetén (forrás: saját számítások)

A 2. ábra és a 3. ábra alapján látható, hogy

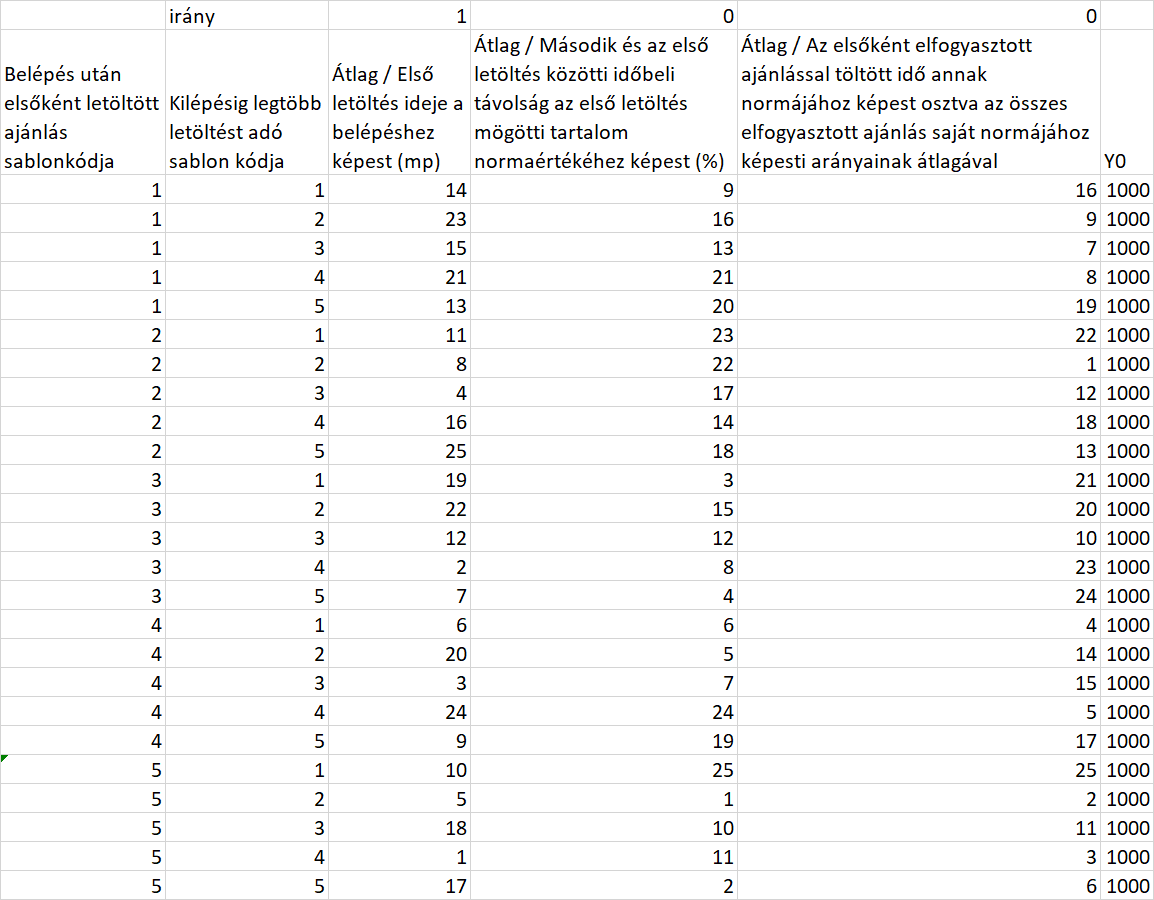
* a minden egyes attribútum (oszlop) minden adatot (1133) felhasznál
* mind a 25 objektum-variáns létezik
* egy-egy mező mögött több tucat rekord áll (az 1. ábra struktúráját alapul véve)
* az egyes kódpár-objektumok mögött eltérő esetszám áll
* az első hír leggyorsabban 43.6, míg leglassabban átlagosan 63.3 mp alatt került aktiválásra
* az első hír elfogyasztásának aránya 43.9-62% között mozog átlagosan a norma-időhöz képest
* az adott személyek norma-időhöz képesti átlagos hírfogyasztásának mértéke az első hírre 92.3-110.5% között mozog



1. ábra: OAM1 – darabnézet (forrás: saját számítások)

A 4. ábra értelmében:

* minden iránnyal rendelkező oszlopban a 2. ábra átlagos adatainak oszloponkénti sorrendje látható
* az Y0 oszlop az anti-diszkriminatív modell konstansa
* az elemzés célja (COCO Y0: <http://miau.gau.hu/myx-free/coco/index.html>) annak levezetése, vajon lehet-e minden objektum másként egyforma
* ehhez lépcsős függvény kerül online optimalizálás keretében levezetésre, mely megadja, vajon az egyes kódpárok értéke mennyi a többi kódpárhoz kapcsolódó hírfogyasztási szokások versengő erőterében
* ahol, ha minden kódpár becsült értéke maga a modell-konstans (1000), akkor az egyes tartalomajánlási módszerek (ezek kombinatív párosításai) között nincs olyan eltérés, mely a hírajánlási rendszer bármely paraméterének változtatását várná el
* ha nem lesz igaz, hogy minden objektum másként lehet egyforma, akkor kényszerűen (a mérlegszerűség okán) lesznek jobb és gyengébb kombinációk, mely a kombinációt alkotó kódok mentén önállóan is értelmet nyernek
* s lehetővé teszik immár az 1-től 5-ig terjedő kódok végső versenyét (kódpárok nélkül)



1. ábra: A hasonlóságelemzés tanulási mintája (forrás: saját számítások)

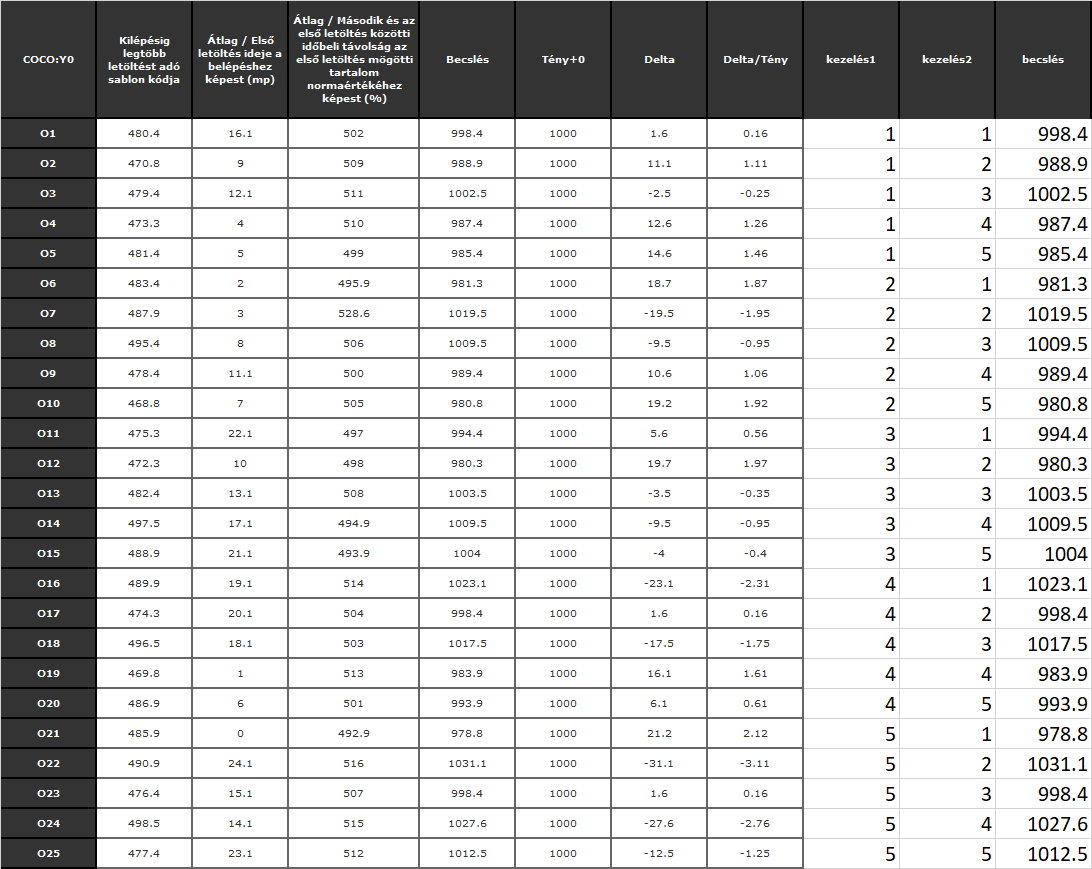
# A modell

Az 5. ábra értelmében:

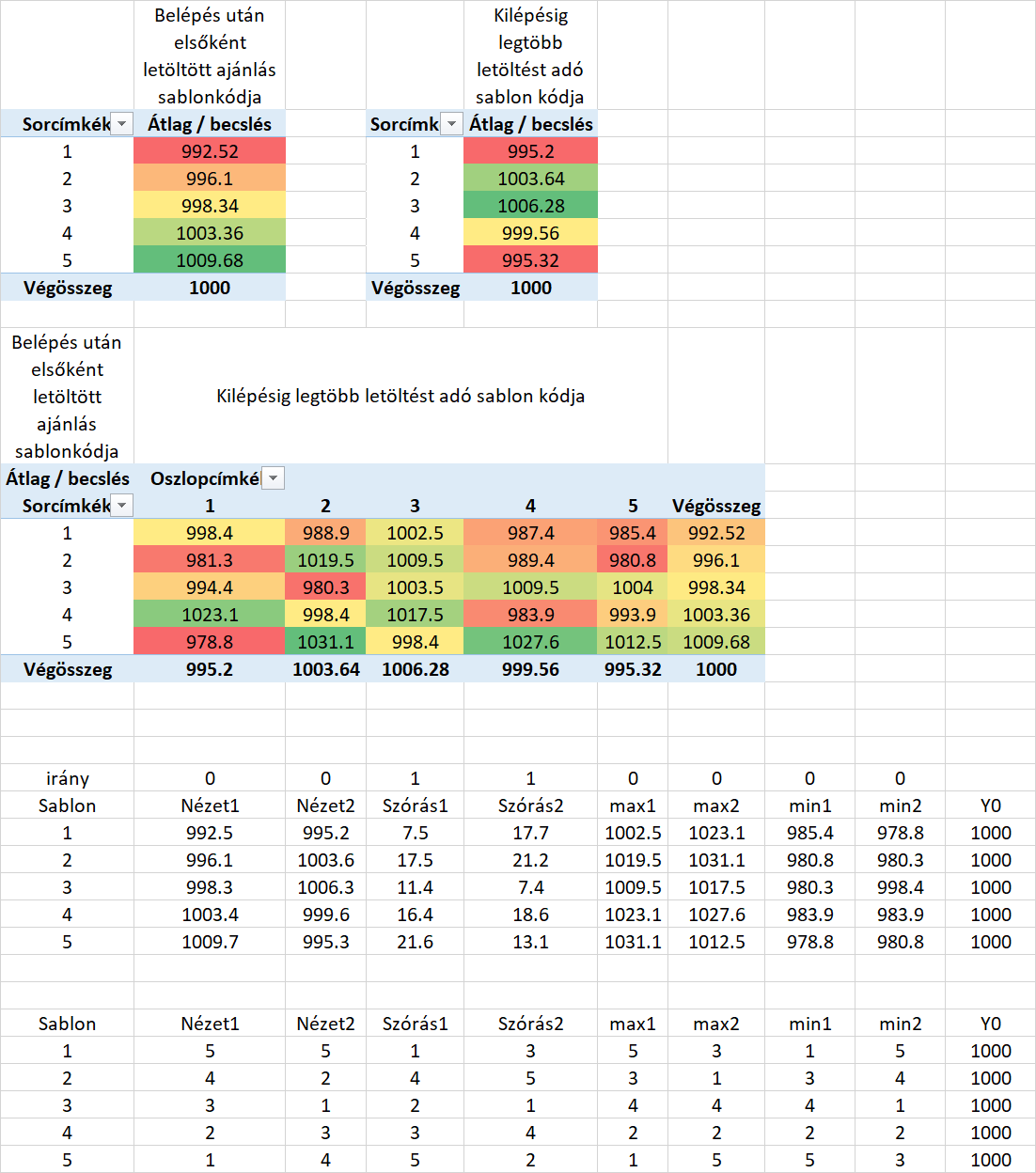
* a 25 kódpár (O1—O25) nem tekinthető azonos értékűnek
* (jelen esetben a függvény-szimmetriára alapozó konzisztencia-számítástól való eltekintés egyszerűsíti a modellezési folyamat értelmezését a komplexitás-redukció miatt)
* ha a kódokat oszloponként önálló hatásmechanizmusnak tekintjük, akkor a kezelés1 és kezelés2 oszlopok, ill. a becslés oszlop kiemelése elvezet a kódok (immár párképzés nélküli) értékeléséhez (vö. 6. ábra)

A 6. ábra értelmezése:

* a felső két (feltételes formázással, vagyis színkódokkal támogatott adat-vizualizációs) riport (1D) a kódok kezelésenkénti átlagos becslési értékeit, vagyis a kódok egymáshoz képesti sikerességét mutatja ki
* az ideális állapot az lenne, hogy az első hír és az összes hírből a leggyakoribb módszer-kód azonos lenne (vagyis létre sem jönnének a nem azonos kódpárt tartalmazó objektumok, ami a további elemzések attribútumainak létére is kihatna)
* vagyis a 2D-ben színes második riportegységben zöldes és pirosas sorok/oszlopok alakulnának ki a véletlen számokra alapozó tarka mintázat helyett
* az 1D és a 2D riportok alapján alakul ki az alsó két fekete-fehér nézet (OAM2), ahol a cél olyan új attribútumok definiálása és ezek irányának meghatározása, melyek alapján egy ÚJ anti-diszkriminációs vizsgálat feltárja, hogy a korlátozott véletlen számtömegben van-e egyenszilárdsági repedés, vagyis van-e legjobb módszer
* ahol az alsó fekete-fehér táblázat a felső sorszámok alapján standardizált állapota
* Nézet1 és Nézet2 a kezelések szerinti átlagokat ismétli meg és rangsorolja oszloponként, elvárva, hogy a nagyobb becslés vezessen majd jobb módszertanhoz
* a Szórás1 és Szórás2 a 2D-s sorok és oszlopok szórásai, ahol elvárás, hogy a kisebb szórások kötődjenek a jobb módszertanhoz
* a Max1 és Max2, ill. a Min1 és Min2 a 2D-s nézet sorainak és oszlopainak szélső értékeit mutatják be, azt feltételezve, hogy annál jobb egy módszer, minél nagyobbak a becslések szélső értékei



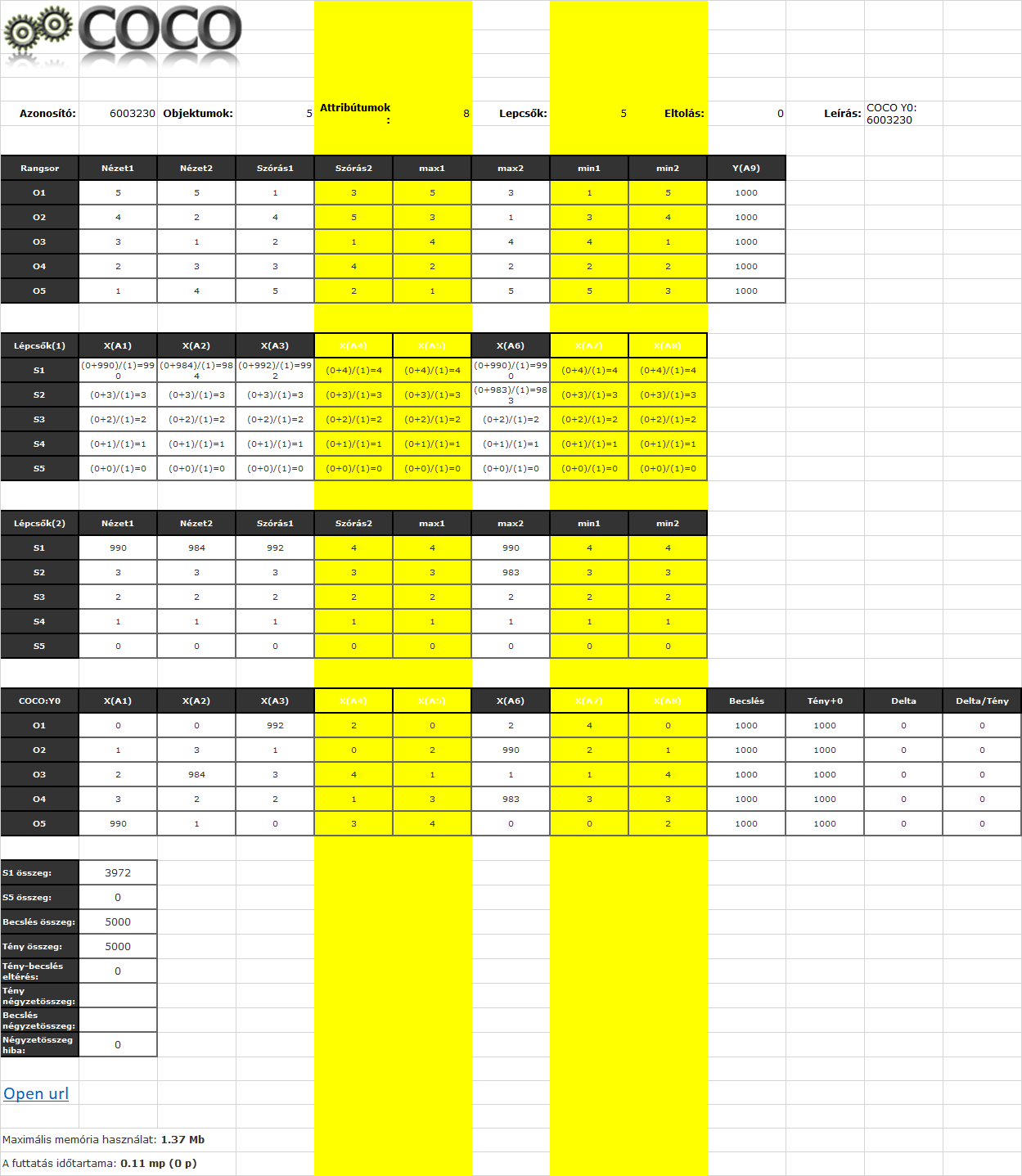
1. ábra: Az anti-diszkriminatív kódpár-összehasonlítás eredményei (forrás: saját számítások)



1. ábra: Az első modellből fakadó második anti-diszkriminációs modell (forrás: saját számítások)

A 7. ábra:

* az 5 kód 8 attribútum alapján levezetni engedi a minden kód (módszer) másként egyforma állapotot
* azonban a sárgával kiemelt attribútumok információtartalma még nem került feldolgozásra
* így egy másik modellezési lépés is szükséges (vö. 8. ábra)

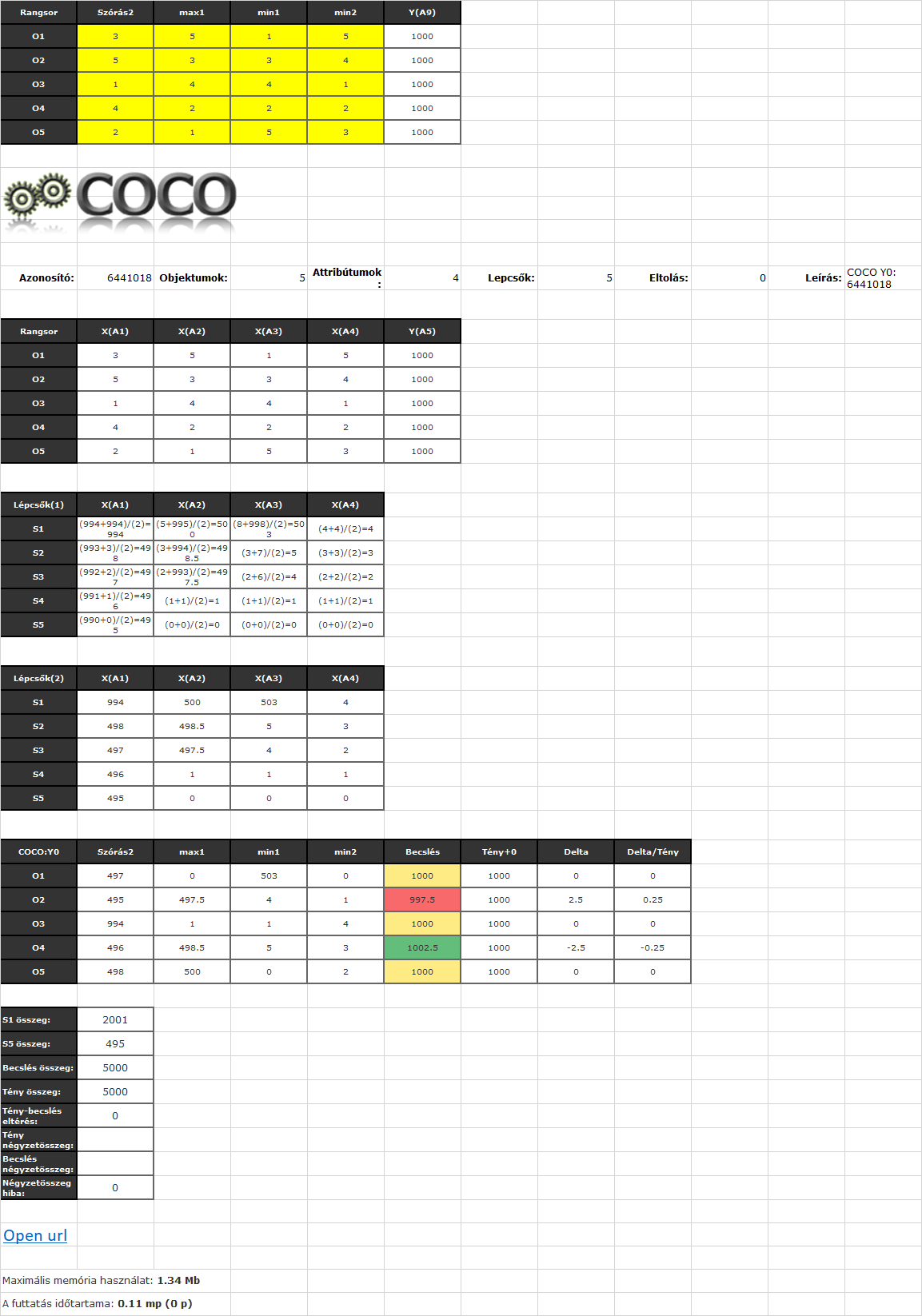


1. ábra: A kódok versenyének első állomása (forrás: saját számítások)

# Eredmények

A 8. ábra maga az eredmény, vagyis látható, hogy

* a legjobb módszer a 4. (zöld) és
* a legrosszabb ajánlási módszer a 2. (piros)
* s a többi módszer normaszerűen viselkedik (becslés = 1000 pont, sárga)
* (a függvényszimmetria-vizsgálatokról az egyszerűség kedvéért itt is eltekintünk)



1. ábra: A legjobb módszer és a leggyengébb módszer levezetése (forrás: saját számítások)

# Konklúziók

A fenti soklépcsős, sok objektumos, sok attribútumos eljárás példa arra,

* hogyan lehet az irányok stratégiai/szakértői szintű felvállalása esetén
* a nyers adatokból
* további emberi beavatkozások nélkül
* automatikusan
* olyan döntési paramétert levezetni, mely alapján
  + a preferálandó ajánlási sablon/módszer által feltárt tartalmak előkelőbb
  + míg a versenyt elvesztő módszer által feltárt tartalmakat hátrányosabb pozícióba lehet sorolni
  + több cikluson keresztül követve a fentebb leírt vezérlést

A fenti elv minden rétege önmagában is kielégíti a KNUTH-i elvet, vagyis: tudás/tudomány az, ami forráskódba átírható, minden más emberi (szakértői?) aktivitás művészet!

Amennyiben egy ilyen ún. finom hangoló hírsorrend-vezérlés beépítésre kerül egy ajánló rendszerbe, attól sajnos még a hírfogyasztás maga nem feltétlenül kell, hogy javuljon, hiszen ez a megoldás önmagában is lehet egy Sokall-effektus (vö. nem tipikus OAM-képzés, ahol a kódpárok speciális kölcsönhatásokat írnak le egy-egy objektumnak tekintett esetben), vagy egy RND-alapú látszatmegoldás.

Ennek a gondolati szálnak a kinyitása átvezet a komplex rendszerek monitorzásának kényszerei felé, ahol elsőként a rendszerállapotok minősítő értékelése a feladat: pl. itt és most ideálisnak vélhető lenne az az állapot, ahol egy felhasználó, egy session alatt minden hírt a felsorolás sorrendjében és teljes mértékben elfogyaszt. Minden eltérés az ideáltól szub-optimális hírfogyasztás, melynek oka a szub-optimális hírajánlás maga….