

# Mesterséges Intelligencia

Az ajánlórendszerek alapjai és főbb  
kihívásai

# A lecke főbb témakörei

- Adatbányászat, data science, big data
- Ajánlórendszerek és információ-visszakeresés
- Főbb kihívások
- Kollaboratív szűrés
  - Memória alapú
  - Modell alapú
  - Hibrid szűrés
  - FunkSVD
- Tartalomalapú szűrés
- Hibrid szűrés

# Az ajánlórendszerek alapjai és főbb kihívásai

Adatbányászat, data science, big data

# Adatbányászat, data science, big data

- A modern technológia révén a korszerű megoldásoknak másodpercenként keletkező több millió, nagyobb szoftveróriások webes szolgáltatásai esetén milliárd rekordos adatmennyiséget kell feldolgozni
- heterogén forrásokból eltérő megbízhatósággal és összetétellel
- komplex összefüggések elemzését végrehajtani, előrejelzéseket, jóslatokat generálni
- A három fogalmat gyakran keverik, ami érthető, mert átfednek

# Adatbányászat

- az adatbázisból történő tudásfeltárás (Knowledge Discovery in Databases, KDD) módszereknél alkalmazott technikák formájában fejlődött
- a nagyméretű adatbázisokban való mintázatokkal és trendekkel kapcsolatos információkinyerés a cél
- leggyakoribb üzleti felhasználása a döntéstámogató rendszerek területén jelenik meg.

# data science

- a részének tekinthető az adatbányászat
  - Gyakran használják az eszközeit
- története az 1960-as évekig vezethető vissza
- A vállalati szférában jellemzően adatalapú szolgáltatások és rendszerek építésére használják
- számos más területhez kapcsolódik
  - például a gépi tanulás
    - data science területén általános felhasználása van
- A terület művelőinek nem elegendő adatfeldolgozási és statisztikai ismeretekkel rendelkeznie
  - kellő háttérre van szükség a gépi tanulás, az adatvizualizáció, a programozás területein is.

# big data

- A big data fogalma alatt többet értünk, mint nagymennyiségű és összetettségű (tipikusan strukturálatlan) adatok (volume) kezelése
- bele értjük annak keletkezésének (esetenként feldolgozásának) sebességét (velocity), valamint teljes környezetet, ami lehetővé teszi ezt
- Aszolgáltatásokhoz gyűjtött adatok köre nagy változatossággal (variety) és megbízhatósággal (veracity) bírhat
- eltérő szemlélettel kezelendők, mint a klasszikus adatbázisok és eltérő üzleti értékkel (value), valamint különböző felhasználhatósági lehetőségekkel (variability) rendelkeznek
- az utóbbi időben a hardver és szoftvertechnológiai megoldások fejlődésével a nagy adatok feldolgozása sem olyan jelentős kihívás
  - mellyel az olyan területek foglalkoznak, mint az általános felhasználású grafikai egységek (General Purpose GPU, GPGPU) és a nagyteljesítményű számítástechnika (High Performance Computing, HPC)
- Ezzel szemben a mesterséges intelligencia területén belül a gépi tanulás és a mély neurális hálók különböző profilozó és adatelemző alkalmazásai, valamint a fuzzy technológiák alkalmazása az adatok és kapcsolatok leírására egyre nagyobb szerepet kapnak.

# Az ajánlórendszerek alapjai és főbb kihívásai

Ajánlórendszerek és információ-  
visszakeresés



# Ajánlórendszerek és információ-visszakeresés

- Az ajánlórendszerek (recommender systems, RS) egészen az információ-visszakeresésig (vagy információ beszerzés, information retrieval, IR) vezethető vissza
- Az IR módszerek rendezett, jellemzően statikus adatbázisokból való hatékony információszerzésre szolgálnak
  - a legnépszerűbb példa a különböző webkereső szolgáltatások, mint a Google és a Bing
  - alkalmazzák még például könyvtárakban és egyéb katalógusos rendszereknél is, mint például a számítógépeken tárolt fájloknál.
- Az ajánlórendszerek az IR módszereken belül az információsűrő (information filtering, IF) megoldások közé sorolhatók
  - melyek célja a felhasználó számára releváns információk kiszűrése nagymennyiségű adatból.

# Ajánlórendszerek

- Természetesen az ajánlórendszerek alkalmazása nem csak az üzleti döntéshozatal támogatására és a webes keresések finomítására korlátozódik
- A világháló és az általa nyújtott szolgáltatások széles köre, mint például az e-kereskedelem, a híradás, vagy on-demand szolgáltatások, továbbá a mobiltechnológia és egyéb IoT eszközök terjedése egyaránt hozzájárult az egyre inkább személyre és helyzetre szabott (kontextust függő) ajánlásokat adó rendszerek fejlődéséhez.

# Ajánlórendszerek

- A korai ajánlórendszerek még csak távoli kapcsolatban sem álltak a mesterséges intelligencia területével
- a megoldások és képességeik nagyon szerények voltak
- maga a központi cél megegyezik: termékek, tartalmak vagy szolgáltatások (items) kiválasztása a felhasználók, vagy fogyasztók (users) részére oly módon, hogy az ajánlott (recommended) termékek a lehető legnagyobb relevanciával (és így a fogyasztás legnagyobb valószínűségével) rendelkezzenek.

# Ajánlórendszerek

- A legkorábbi például webes értékesítő megoldásoknál alkalmazott ajánlási technikák valamilyen egyszerű hierarchikus modellen alapultak
  - a különböző termékeket, termékcsoportokat és az azok közötti kapcsolatokat reprezentálta valamilyen módon, a felhasználókkal kapcsolatosan csak az aktuális (vásárlási) folyamat során gyűjtöttek információt, mint például a webáruházak esetén a megtekintett, vagy a kosárba helyezett termékek
- Egy internetes könyváruház esetében a termékeket (item) a könyvek képviselik, amelyeket különböző szempontok, jellemzők (features, mint például szerzők, témakör stb) szerint lehet hierarchikusan csoportosítani.
- Ha egyszer felépítettük a kívánt hierarchiát, akkor egy vásárló (user) aktuálisan vásárolt (vagy vásárlásra választott, kosárba helyezett) könyv(ek) (item) jellemzői alapján könnyedén megkereshetőek a hierarchikus modellben reprezentálva közel álló további könyvek.

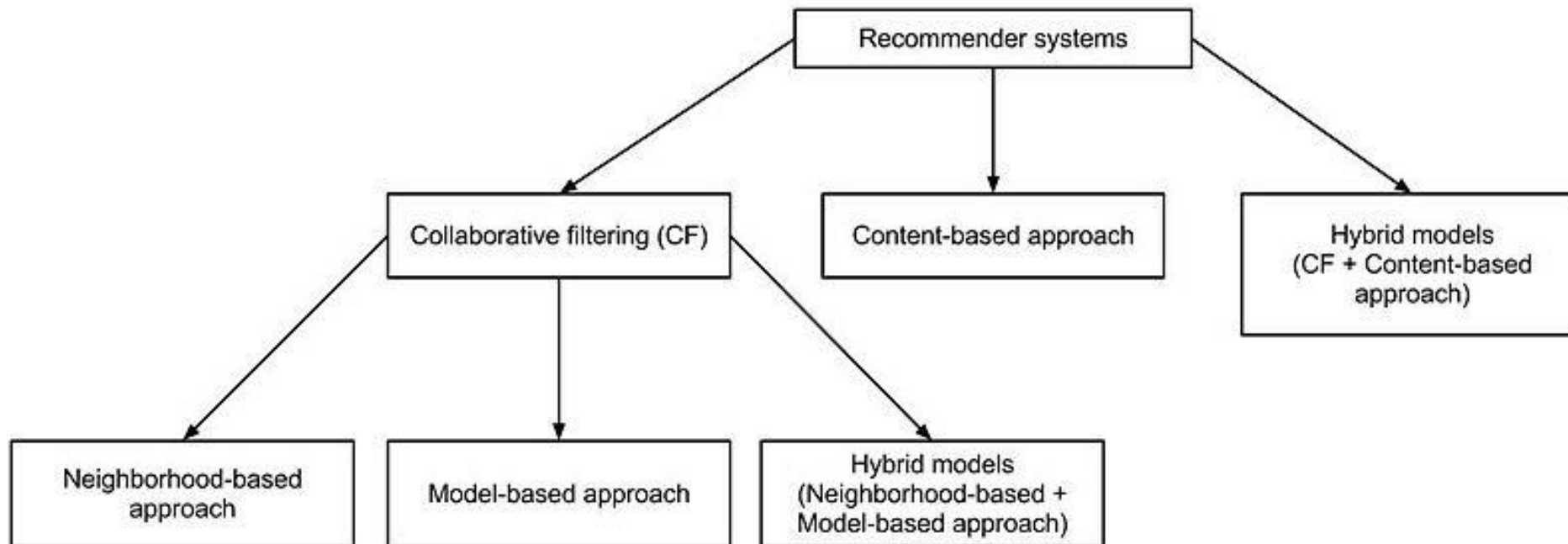
# Ajánlórendszerek

- Egy másik jó szemléltetési lehetőséget adó példa valamilyen elektronikai webáruház ajánlórendszere lehet, ahol a könyvektől eltérően nem biztos, hogy azonos típusú terméket kell ajánlani, hanem ahhoz kapcsolódó kiegészítőket, vagy szolgáltatásokat.
  - Ha egy vásárló konkrét elképzelés alapján választ magának egy TV készüléket, valószínűleg nem választ másikat, így további TV-k ajánlása felesleges lehet
    - kivétel az olyan esetek, ahol például jobb teljesítménnyel, vagy ár-érték aránnyal eltéríthető a fogyasztó választása
    - azonban kapcsolódó termékek ajánlásának sikeressége jelentősen nagyobb lehet
      - Ha valaki TV-t vásárol, akkor nagyobb valószínűséggel nyitott az olyan (kiválasztott termékhez) kapcsolódó termékek vásárlására, mint a kábelek, átalakítók, állványzatok, különböző szolgáltatások, mint a kiterjesztett biztosítások, vagy privát adók.

# Ajánlórendszerek

- A modern ajánlórendszerek már nem statikus fastruktúrákkal és jelentősen bővebb adatkészlettel dolgozhatnak és az egyre pontosabb, személyre szabottabb ajánlások kiválasztása
- Az ajánlórendszereknél alkalmazott megközelítések három fő csoportba sorolhatóak
  - a kollaboratív szűrők (Collaborative Filtering, CF),
  - a tartalom-alapú megközelítések (Content-Based Filtering, CBF),
  - ezek kombinációját használó hibrid modellek
- Számos alternatív csoportosítás, besorolás létezik, melyek esetenként elkülönítik az előbbi három (négy) csoport speciális eseteit is
  - például a knowledge-based model, demográfia, vagy hasznosság alapú szűrések.

# Ajánlórendszerek



# Ajánlórendszerek

- Könnyen belátható, hogy az egyes, specifikus ajánlási feladatoknál felmerülő, termékeket és felhasználókat leíró paraméterek száma és jellege, valamint így a megoldásokhoz szükséges számítási kapacitás az ajánlásokkal szembeni pontossági elvárások függvényében jelentősen megnő
- A mesterséges intelligencia megoldások több szempontból is segítik az ajánlórendszereket.



# Ajánlórendszerek

- Automata adatgyűjtő rendszerek
  - különböző hírek, közösségi portálok bejegyzései, tőzsdei adatok begyűjtésére szolgáló egyszerűbb ágensek
    - Egyszerű példa erre a keresőmotoroknál alkalmazott keresőbotok (crawler bot).
  - A gyűjtött adatokat alapvetően két fő kategóriába sorolhatjuk: implicit és explicit adatokra (visszajelzésekre, értékelésekre, feedback)
    - mindegyike számos eltérő módon keletkezhet.

# Ajánlórendszerek

- Automata adatelemző rendszerek
  - különös jelentősége a különböző multimédiás tartalmak esetében van
    - ahol például egy videó képi és hangi tartalma alapján kell kigyűjteni annak jellemzőit, de akár szöveges tartalmak digitalizálását és annak érzelmi elemzését is elvégezhetik
  - Az ilyen típusú feladatok megvalósítása a modern mesterséges intelligencia megoldások nélkül lehetetlen lenne.

# Ajánlórendszerek

- Modellalkotás
  - alapvetően a termékek és a felhasználók különböző aspektusban való leírására alkalmaznak modelleket
  - A modellek célja, hogy pontos képet tudjanak alkotni a fogyasztói szokásokról, preferenciákról, termékek jellemzőiről, azok kapcsolatairól
    - az ajánlások finomítása érdekében.
  - Nem léteznek univerzális modellek
    - az alkalmazási terület függvényében az egyes felhasználói és termék csoportok körében azonos jellegzetességek mutatkoznak.

# Ajánlórendszerek

- A közösségi média és a fejlett webes technológiák (különös tekintettel a nyomkövetési eljárások), valamint a hatékony (releváns) mesterséges intelligencia megoldások megjelenésével nagyon pontos profilalkotás és a különböző forrásból származó adatok összekapcsolása is lehetővé vált
- olyan szintű információt képesek kinyerni az egyes felhasználókkal kapcsolatosan, melyeknek sokszor még maguk a felhasználók sincsenek tudatában.

# Ajánlórendszerek

- A főbb alkalmazási területek a teljesség igénye nélkül
  - befektetési és egyéb üzleti szolgáltatások
  - kis- és nagykereskedelem
  - utazási és egyéb turisztikai ajánlások
  - médiatartalmak (könyvek, főként on-demand videók, zenék)
  - hétköznapi és luxus ingóságok, ingatlanok
  - Társkeresés
  - egészségügyi szolgáltatások
  - pályaválasztási tanácsadás
  - kompetencia mérésre szolgáló adaptív számonkérési rendszerek
  - lakberendezési, életstílus és divattanácsadás
  - Hírek
  - csoportos tevékenységek
  - emberi erőforrás kezelés.

# Az ajánlórendszerek alapjai és főbb kihívásai

Főbb kihívások

# Főbb kihívások

- Az ajánlórendszerek korlátai, működési elve és túlzott pontossága esetenként épp a céllal szemben komoly ellenszenvet válthat ki
  - Erre példa a Facebookkal kapcsolatosan merült fel
    - egy a magzatát elvesztő nő számára kismamák számára célzott hirdetéseket ajánlott a rendszer, mivel az nem ismerte fel a szituációt, mindezt úgy, hogy közvetlen módon nem, csak a megosztásait, kereséseit elemezve határozták meg azokat a számára.
- Természetesen a legtöbb esetben a modern, tech-óriások által használt algoritmusok a problémakör komplexitásához viszonyítva roppant nagy pontosságot tudnak elérni, az említett eset ritkább, de egyáltalán nem egyedülálló.

# Főbb kihívások

- kulcs kihívása a felhasználói tevékenységek megfelelő értelmezés
  - profilalkotás céljából
- A visszajelzéseknek két kategóriája
  - (1) az explicit visszajelzések
    - például a közvetlen felhasználói értékelések (például pontozások, kedvelések stb.)
    - nagy pontossággal bírnak és pozitív, valamint negatív vélemény abszolút mérésére is képes.
    - begyűjtése nehéz
      - a felhasználónak energiát kell befektetnie az értékelési folyamatba
  - (2) az implicit visszajelzések
    - a tartalmak fogyasztása során gyűjtött adatok, például kattintás
    - Az implicit visszajelzések köre nagyon szerteágazó
    - Előny: könnyű begyűjteni, így nagyobb mennyiségben áll rendelkezésre
    - hátrány, megbízhatatlan, relatív, kizárólag pozitív visszajelzési információk nyerhetők, értelmezésük nehéz
- vannak olyan megoldások, melyek ezt transzformációs feladatként próbálják megközelíteni.



# Főbb kihívások

- A profilalkotás további számos kihívással bír
  - például a felhasználók preferenciáinak időbeli változása, ami vonatkozhat az ízlés változására, vagy egyéb szezonális hatásra
  - Természetesen nem csak hosszú távú időbeni hatások, hanem például az adott napszak, vagy az éppen használt eszköz is, vagyis a kontextus
  - Az egyes alkalmazási területek, valamint a fogyasztókban lezajló kognitív folyamatok alapján begyűjthető tudás hozzájárul a probléma mély ismeretéhez és így a pontosabb ajánlásokhoz
  - Egyes alkalmazások esetén jelentős kihívás a felhasználó megfelelő azonosítása
    - Például egy háztartásban a különböző előfizetéseket, internetes szolgáltatások fiókjait többen is használhatják
    - erre a legjobb példa a televízió
    - Ilyen esetekben a felhasználók profiljának további finomítása, profilhoz tartozó több felhasználó detektálásának lehetősége jelenthet megoldást
      - ennek lehetnek technológiai határai is, míg egy IPTV környezetben kellő visszacsatolással rendelkezhetünk ehhez, addig a hagyományos kábeles szolgáltatások esetén ez már nem mondható el.

# Főbb kihívások

- A hidegindítási (cold-start) probléma a kollaboratív szűrőkön alapuló megoldásokra jellemző jelentős gyakorlati kihívás
  - Az olyan helyzeteket értjük alatta, amikor egy termékről, vagy felhasználóról nem rendelkezünk (kellő, vagy) semmilyen információval, így azokat nem tudjuk csoportosítani.
    - Ez jellemző probléma például új termékek, vagy új felhasználók megjelenésekor, ilyenkor nagyobb hangsúlyt fektetnek a meta-adatokra, részben a hidegindítási probléma megoldása motiválta a tartalomalapú szűrőket.

# Főbb kihívások

- A legtöbb ajánlórendszer számára komoly kihívást jelent a rokonértelműség (synonymy)
  - amikor a kismértékben eltérő leírással rendelkező, vagy azonos, de az adatbázisban külön bejegyzésként szereplő termékeket kell kezelnie
  - A gyakori megoldások közé tartozik a termékek leírásainak részletes elemzése, de ez gyakran költséges, illetve esetenként kiszűrik a nagymértékben hasonló, de valójában különálló termékeket
  - Hibátlan, teljesen automatizált rendszer nem létezik, gyakran igényel emberi közbeavatkozást.

# Főbb kihívások

- Általánosságban elmondható, hogy több mintaadat révén pontosabb termék, vagy felhasználói modell alkotható
  - azonban elméletileg elérhető egy olyan állapot, melyet követően az új mintaadatok már nem tudnak szignifikánsan javítani a modellen
    - ezt hívják telítettségi pontnak (saturation point)
- A hidegindítási fázis és a telítettségi pont közötti időszakot nevezik bemelegítési (warm-up) periódusnak
  - Triviális, hogy minél kevesebb minta révén érhető el a telítettségi pont, annál gyorsabban alkotható meg egy pontos modell, így több kutatás is ezt a problémát vizsgálja.

# Főbb kihívások

- Az olyan felhasználókat, akik visszajelzései kiszámíthatatlanok, egyetlen a rendszerben tárolt felhasználói csoportnak sem feleltethetők meg egyértelműen, szürkebáránynak (gray sheep) nevezzük
  - A kollaboratív szűrésen alapuló megoldások jellemzően nem képesek az ilyen helyzetek kezelésére
- A felhasználók egy másik csoportja a feketebárányok (black sheep), akiknek annyira egyedi ízlésük van, hogy számukra a megfelelő ajánlás generálása szinte lehetetlen.

# Főbb kihívások

- A szélhámosok támadása (Shilling Attack)
  - kollaboratív rendszerekre jellemző, ahol az egyes felhasználók tetszőlegesen értékelhetik a termékeket
    - ezt kihasználva rosszindulatú felhasználók a saját termékeiket számos pozitív értékeléssel látják el, míg a versenytársakét negatívakkal
    - A termékalapú kollaboratív rendszerek kevésbé kitettek az ilyen típusú támadásoknak

# Főbb kihívások

- A skálázhatóság (scalability) gyakori szűk keresztmetszetet jelent az ajánlórendszerek számára.
  - Sok módszer kiválóan működik egyszerűbb vizsgálatok során, azonban valós piaci körülmények között már nem képesek hatékonyan kezelni az adatokat.
  - A legtöbb szolgáltatás esetén (közel) valósidőben kell az adatokat feldolgozni
    - ez többmillió felhasználói bázis és termékpaletta esetén
      - például amivel az Amazon is rendelkezik
    - nem elfogadható olyan algoritmus, aminek a hatékonysága  $O(n)$  vagy azt meghaladó
  - Dimenzióredukciós módszerek segíthetnek csökkenteni a probléma méretét, de jellemzően költséges mátrix faktorizációs megoldásokról van szó.

# Főbb kihívások

- Az utóbbi években különösen nagy figyelmet fordítottak az ajánlások magyarázatára
  - vagyis arra, hogy a rendszer képes legyen indokolni, levezetni, hogy milyen tényezők alapján ajánlottak egy terméket a felhasználó számára
  - Az egyszerűbb megoldások esetén ez könnyedén visszakövethető, azonban a modern, komplex rendszerek nagyon dinamikusak és jellemzően heurisztikusak is, valamilyen fekete-doboz (black box) modellezést alkalmaznak, ahol ez nem lehetséges
  - A magyarázatok hasznossága felhasználói oldalról a releváns ajánlások kiválasztásának szempontjából fontos
    - a fejlesztői, szolgáltatói oldalról a modellek pontosabb megértése miatt.



# Főbb kihívások

- Az ajánlórendszerek talán legnagyobb kihívása a módszer validáció során mutatkozik meg
  - annak ellenére, hogy számos forrásból beszerezhetőek különböző típusú történelmi adatok
    - Az adatok beszerzése inkább csak olyan esetben okoz komolyabb gondot, amikor valami olyan speciális területre kell ajánlórendszert készíteni, melynél nehéz a felhasználói visszacsatolást begyűjteni
  - A probléma forrása, hogy hiába ismert egy felhasználó által megtekintett videók listája és jól jósoljuk meg a listán soron következő elemét, az egyáltalán nem biztos, hogy a felhasználó akkor is azt az adott tartalmat választotta volna, ha alternatív, akár jobb ajánlatokat kap
    - Ennek megfelelően az ilyen adatokra épülő validáció inkább a profilalkotási képességek mérésére alkalmas, nem pedig egyértelműen megítélni az ajánlások jóságát.

# Főbb kihívások

- Természetesen a fentieken kívül számos további kihívással kellhet megküzdeni a gyakorlatban is alkalmazható termékajánló rendszerekkel kapcsolatban, melyek lehetnek
  - technológiai jellegűek, mint például a zaj
  - vagy szándékos szabotázs
  - de előfordulnak a felhasználói profilok alkotásával kapcsolatos jogi kérdések is.

# Az ajánlórendszerek alapjai és főbb kihívásai

Kollaboratív szűrés

# kollaboratív szűrés

- működési koncepcióját úgy lehetne összefoglalni, hogy az ajánlások más, jellemzően valamilyen szempont szerint hasonló, egy csoportba tartozó felhasználókkal kapcsolatos ismeretek alapján kerülnek meghatározásra
  - vagyis abból azzal a feltevéssel működnek, hogy más felhasználók korábbi fogyasztási hasonlósága alapján próbálja megjósolni az ideális terméket
- A kollaboratív szűrésen alapú ajánlórendszereket három fő csoportba sorolhatjuk
  - (1) memóriaalapú (memory-based)
  - (2) modellalapú (model-based),
  - (3) hibrid megoldások.

# kollaboratív szűrés

- egyszerű példán szemléltetve
  - elektronikai cikkek online értékesítése során érdekes lehet az egyes termékekkel való egyéb termékek együttes értékesítése
  - Ekkor a kollaboratív szűrésen alapuló rendszerek a felhasználó kosarában található termékek listájához hasonló korábbi felhasználók kosarait keresi meg
    - az azokban fellelhető egyéb termékek közül kerül kiválasztására az ajánlott termék.

# Az ajánlórendszerek alapjai és főbb kihívásai

Memóriaalapú kollaboratív szűrés

# Memóriaalapú kollaboratív szűrés

- a felhasználóktól gyűjtött termékértékelések (visszajelzések) alapján határozzák meg a termékek és felhasználók közötti hasonlóságot
- A gyakorlati célú rendszerek körében az első megoldások közé tartoztak
  - a szomszédság alapú (neighbor-based),
  - legjobb-N (top-N) ajánlások
- A módszerek nagy előnye, hogy könnyen implementálhatók
- az adatokkal való bővítés könnyen kivitelezhető, nem kell figyelembe vennie a termékek tartalmi jellemzőit
  - nagyon nagy adatbázisok esetén a skálázás kezelhetetlen komplexitáshoz vezethet.
- A hátrányok közé sorolható, hogy az ilyen rendszerek jellemzően explicit adatokra (közvetlen értékelésekre) építenek, az adatok hiánya nagyban befolyásolja az ajánlások jóságát
- új, ismeretlen felhasználók vagy termékek esetében nem tud ajánlást generálni

# Memóriaalapú kollaboratív szűrés

- A módszerek két fő lépésre oszthatók
- (1) minták csoportosítására, vagyis a termékek, vagy felhasználók csoportokba való sorolása
- (2) ajánlási fázis, nincs modellépítés.



# Az ajánlórendszerek alapjai és főbb kihívásai

kNN módszer

# kNN módszer

- gyakran alkalmazzák a  $k$  legközelebbi szomszéd (k-nearest neighbor, kNN) megoldást
- az egyes felhasználók, vagy termékek egy  $n$  dimenziós paraméter tér pontjaiként ábrázolhatók
- Az így kapott ponthalmazt  $k$  darab csoportba (osztályba) sorolható
- a cél, hogy megkeressük egy új adatpont (minta) számára azt a csoportot, amelyikbe a leginkább illeszkedik
- Az alapfeltetelezés az, hogy minél kisebb a (valamilyen módon definiált) távolság két adatpont (jellemzői) között, annál nagyobb a valószínűsége, hogy azonos csoportba tartoznak.

# kNN módszer

- A működési elv az alábbi lépésekre bontva általánosítható:
  - $k$  értékét kiválasztjuk,
  - minden egyes ismert adatpontra (tanító minta) az ismeretlen és az ismert adatpontok közötti távolságmérték meghatározása,
  - az ismert adatpontok sorba rendezése a távolságmérték alapján, majd a legjobb  $k$  darab adatpontot kiválasztjuk,
  - a kiválasztott  $k$  darab ismert adatponthoz rendelt leggyakoribb csoportot kiválasztva visszatérünk azzal, mint az ismeretlen adatpont osztálya.

# kNN módszer

- Az új adatponthoz így meghatározott csoport alapján a hasonló preferenciákkal rendelkező korábbi, ismert adatpontok további (az új felhasználó által még nem véleményezett termékek) értékelései alapján releváns elemek határozhatók meg.

# Az ajánlórendszerek alapjai és főbb kihívásai

Modellalapú kollaboratív szűrés

# Modellalapú kollaboratív szűrés

- az egyes termékekről és/vagy felhasználókról modellt készítünk az alapján, hogy az adott felhasználó milyen jellemzőket keres egy termékben, vagy egy termék milyen tulajdonságokat nyújt egy felhasználó számára
  - Ezek a jellemzők, tulajdonságok (feature) általában nem érhetők el explicit módon, gyakran csak absztraktak, vagy látensek, a felhasználói és termék vektorok leírására szolgálnak
- Az itt használt megoldások jellemzően összetettebbek, gyakran valamilyen dimenzióredukciós technikát alkalmaznak
- megjellennek már mesterséges intelligenciához köthető megoldások is
  - pl. a Markov döntési folyamat, a szinguláris érték felbontás, a különböző klaszterező megoldások, vagy a Bayes valószínűségi hálók, vagy a mesterséges neurális hálózatok
- Lényegében a memória alapú megoldásoknál is használt információt tömörítjük a dimenzió redukció révén, így kisebb adathalmazzal dolgozhatunk.

# Modellalapú kollaboratív szűrés

- előnyei a memóriaalapú megoldásokkal szemben
  - pontosabb és hatékonyabb ajánlásokat generál
  - könnyebben kezeli a hiányos adatokat
  - skálázhatóbb
- Hátránya
  - a modellalkotási folyamat erőforrás-igényes
  - a skálázás mértéke és a hatékonyság között fordított kapcsolat áll fenn
    - kompromisszumokra lehet szükség
    - hasznos információt veszíthetünk a dimenzióredukció miatt
- Az egyes implementációk eltérhetnek az alkalmazott, módszerekben, mint például a különböző klaszterezési megoldásokban
  - vagy a technikák által alkalmazott paraméterek és egyéb ajánláskorrekciós módszerekben

# Modellalapú kollaboratív szűrés

- A klaszterezési technikákat alkalmazó modellalapú kollaboratív szűrők esetén a  $k$ -közép ( $k$ -means), vagy annak fuzzy változatát, a fuzzy  $c$ -közép (fuzzy  $c$ -means) a leggyakoribb példa
- A cél, hogy  $n$  darab megfigyelést, vagy adatpontot  $k$  darab klaszterbe, vagy csoportba soroljunk
- főbb lépések:
  - Klaszterek középpontjainak inicializálása
  - Amíg a konvergencia nem áll fen
  - Minden adatpont (megfigyelést) hozzárendelése a legközelebbi klaszterközépponthez
  - Az azonos klaszterbe sorolt adatpontok alapján új klaszterközéppont meghatározása



# Modellalapú kollaboratív szűrés

- A klaszterközéppontok kezdeti inicializációja véletlenszerűen, míg a negyedik lépésben az azonos klaszterbe tartozó adatpontok valamilyen középértéke alapján történik
- Az algoritmust egyszerűsége miatt kedvelik
- Hátránya
  - hogy a lokális optimum felé viszi a megoldást
  - valamint a kezdeti klaszterközéppontok kiválasztásától nagyban függ a megoldás kimenete
  - a klaszterek számát előre ismernünk kell
- A k-közép algoritmus gyakorlatilag egy nemfelügyelt tanuláshoz is tekinthető
  - feltárja az előttünk rejtett közös jellemzőket.
- A különböző szempontok alapján az eltérő csoportokba besorolt felhasználó számára könnyen meghatározható egy ajánlás az azonos csoportba tartozók zsánerei alapján.

# Az ajánlórendszerek alapjai és főbb kihívásai

Funk SVD

# Funk SVD

- Simon Funk
- A modellalapú megközelítések esetén az egyik leggyakrabban használt megoldások a mátrix faktorizációra épülnek
- lényege az alábbi módon fogalmazható meg
  - a felhasználók és termékekre adott (ismert) értékeléseik egy ritka (hiányos, sparse) mátrixban találhatók (értékelési mátrix)
  - felbontható két másik (felhasználói és termék) mátrix szorzatára
  - Ekkora a felbontás eredményeként kapott két mátrix a felhasználói mátrix és a termék mátrix, melyek mérete kisebb lesz az eredeti értékelési mátrixénál és a hiányzó adatok is eltűnnek
  - A felhasználói és a termék mátrixok ekkor úgynevezett látens jellemzőket (factor, feature) tartalmaznak
  - Természetesen több lehetséges felhasználói és termék mátrix létezhet, ami a ritka értékelési mátrixra illeszkedik

# Funk SVD

- ismert az  $R$  ritka értékelési mátrix (rating matrix)
  - soraiban az egyes felhasználók
  - oszlopaiban az egyes termékek
  - az egyes felhasználók által konkrét termékekre adott értékelése a hozzájuk tartozó cellából olvasható ki
- Ekkor jóslt érzékelési értékeket tartalmazó  $P$  mátrix már nem ritka és előállítható a  $H$  felhasználói, valamint  $W$  termék mátrix szorzataként

$$p_{ui} = \sum_{j \in J} H_{uj} W_{ji}$$

- ahol  $p_{ui}$  az  $u$ . felhasználó jóslt értékelése az  $i$ . termékre,  $J$  pedig a látens faktorok (jellemzők) halmaza

# Funk SVD

- Az algoritmus célja olyan  $H$  és  $W$  mátrixok felkutatása, amelyek eredményül a jósalt és ismert értékek közötti eltérések minimálisak
- Kezdetben e mátrixokat véletlenszerű értékekkel szokás feltölteni
- Minden ismert, értékelési mátrixban szereplő értékre (mintára) meghatározzuk a jósalt értékelést
- A jósalt és a valós értékek közötti eltérést valamilyen  $E$  hiba, vagy költségfüggvénnyel leírjuk
- Az  $H$  és  $W$  mátrixokra vonatkozó legkisebb értékek meghatározása:

$$\arg \min_{H, W} \|R - P\|_F + \alpha \|H\| + \beta \|W\|$$

- $\alpha, \beta$ : tanulási sebességet szabályozó paraméterek
- $\|\cdot\|_F$ : Frobenius norma, vagy annak változata

# Az ajánlórendszerek alapjai és főbb kihívásai

Hibrid kollaboratív szűrők

# Hibrid kollaboratív szűrők

- A legtöbb piaci alkalmazással rendelkező kollaboratív szűrést alkalmazó ajánlórendszer valamilyen hibrid megoldást alkalmaz
- Két fő megközelítés
  - a modell és memória alapú módszereket ötvözése
  - a kollaboratív szűrési eljárásokat tartalom alapú információk erősítésével ötvözve (content-boosted)
- Gyengeségek kiküszöbölése, mint a ritka/hiányos/elvesztett adatok
- Komoly hátránya az efféle megoldásoknak, hogy bonyolultabbak, így implementálásuk is nehezebb

# Az ajánlórendszerek alapjai és főbb kihívásai

Tartalomalapú szűrés



# Tartalomalapú szűrés

- az egyes termékekkel kapcsolatos leírások és egyéb alapadatok állhatnak rendelkezésre a (jellemzően korlátozott) felhasználói profilon túl
- Különösképp olyan helyzetekben alkalmazzák, amikor a termékekről kiterjedt adatokkal rendelkeznek, azonban a felhasználókról egyáltalán nem, vagy korlátozottan
- Az ilyen rendszerek alapvetően a felhasználó korábbi választásai alapján ajánl azokhoz nagyon hasonló termékeket.

# Tartalomalapú szűrés

- Az értékelési mátrixot ( $R$ ) a kollaboratív szűrésen alapuló technikákhoz hasonlóan itt is felbontják a felhasználói és termékvektorokra
- Az egyes termékvektorok tartalmazzák az azokat leíró jellemzőket, amely alapján összehasonlíthatóak
- Az ajánlások a felhasználói ( $U$ ) és termékvektorok ( $I$ ) koszinusz hasonlósága (cosine similarity) alapján kerülnek meghatározásra:

$$\text{similarity}(U, I) = \cos(\theta) = \frac{U \cdot I}{\|U\| \|I\|} = \frac{\sum_{j=1}^n U_j I_j}{\sqrt{\sum_{j=1}^n U_j^2} \sqrt{\sum_{j=1}^n I_j^2}}$$

# Tartalomalapú szűrés

- Előnye

- a hidegindítási problémára megoldást nyújtanak, mivel alacsony szintű, alapadatokra építik az ajánlásokat és a modellek finomítására jellemzően meta-adatokat használnak

- Hátránya

- nem generálnak kellően változatos, vagy új tartalmakat a felhasználók számára
  - alapvetően a korábbi, ismert preferenciákhoz nagyon hasonló termékeket ajánlják, nem segítik a felhasználókat a tartalmak felfedezésében (discovery).

# Az ajánlórendszerek alapjai és főbb kihívásai

Hibrid szűrés

# Hibrid szűrés

- A hibrid szűrésen alapuló megoldások jellemzően valamilyen CF és CBF módszereket alkalmaznak
  - párhuzamosan határozzák meg az ajánlásokat
  - egy külön egység hozza meg a végső ajánlásokat
- Az ilyen megoldások hátránya, hogy jelentős az erőforrásigényük.

# Felhasznált források

- <https://sifter.org/~simon/journal/20061211.html>
- [https://datajobs.com/data-science-repo/Recommender-Systems-\[Netflix\].pdf](https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf)
- <https://recsys.acm.org/>
- <https://www.kdnuggets.com/2017/01/data-science-puzzle-revisited.html>
- <https://www.facebook.com/notes/Facebook-data-science/the-formation-of-love/10152064609253859>
- [https://index.hu/techtud/2018/12/15/techceg\\_hirdetes\\_problema\\_keretlen\\_halva\\_szulett\\_gyermek/](https://index.hu/techtud/2018/12/15/techceg_hirdetes_problema_keretlen_halva_szulett_gyermek/)
- [http://www.uni-miskolc.hu/~matsefi/maliok/cikkek/MAFIOK\\_2013\\_Buza\\_Buzan.pdf](http://www.uni-miskolc.hu/~matsefi/maliok/cikkek/MAFIOK_2013_Buza_Buzan.pdf)
- <http://www.ir.uwaterloo.ca/book/01-introduction.pdf>
- <https://www.oracle.com/big-data/guide/what-is-big-data.html>
- [https://www.sas.com/en\\_us/insights/big-data/what-is-big-data.html](https://www.sas.com/en_us/insights/big-data/what-is-big-data.html)
- <http://infolab.stanford.edu/~ullman/mmds/ch9.pdf>
- [https://www.researchgate.net/figure/Characteristics-of-explicit-and-implicit-feedback\\_tbl1\\_247927239](https://www.researchgate.net/figure/Characteristics-of-explicit-and-implicit-feedback_tbl1_247927239)
- <https://www.hindawi.com/journals/aai/2009/421425/><https://www.uts.edu.au/sites/default/files/desi-publication-recommender%20system%20application%20developments%20a%20survey-accepted%20manuscript.pdf>[http://www.ijircc.e.com/upload/2014/august/5\\_ASurvey.pdf](http://www.ijircc.e.com/upload/2014/august/5_ASurvey.pdf)[https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf\\_ita2017\\_04008.pdf](https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf_ita2017_04008.pdf)
- <https://www.sciencedirect.com/science/article/pii/S0957417418300903>
- <https://arxiv.org/pdf/1803.00146.pdf> L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining Collaborative Filtering Recommendations", Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, New York, USA, 2000, pp. 241–250. P. Resnick and H. R. Varian, Recommender systems, Communications of the ACM, vol. 40 (3), 1997, pp. 56–58. G. Jawaheer, M. Szomszor, and P. Kostkova, "Comparison of implicit and explicit feedback from an online music recommendation service", Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec '10), ACM, USA, 2010, pp. 47–51. X. Amatriain, J. Pujol, and N. Oliver, I like it... I like it not: Evaluating User Ratings Noise in Recommender Systems, User Modeling, Adaptation, and Personalization, Springer, Berlin, 2009, pp. 247–258. Y. Hu, Y. Koren, and C. Volinsky, "Collaborative Filtering for Implicit Feedback Datasets", Data Mining, ICDM '08, Eighth IEEE International Conference, Pisa, 2008, pp. 263–272.
- H. Sobhanam, and A. K. Mariappan, "Addressing cold start problem in recommender systems using association rules and clustering technique", Computer Communication and Informatics (ICCCI), 2013 International Conference, Coimbatore, 2013, pp. 1–5. R. Van Meteren, and M. Van Someren, "Using content-based filtering for recommendation", Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop, 2000, pp. 47–56. D. Parra, A. Karatzoglou, X. Amatriain, I. Yavuz: Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping, Proceedings of the CARS-2011, USA, 5 pages, 2011. [http://navatintarev.com/papers/Nava%20Tintarev%20PhD%20Thesis%20\(2010\).pdf](http://navatintarev.com/papers/Nava%20Tintarev%20PhD%20Thesis%20(2010).pdf)<https://upcommons.upc.edu/bitstream/handle/2099.1/22602/102384.pdf>[https://www.researchgate.net/publication/220173171\\_A\\_Survey\\_of\\_Collaborative\\_Filtering\\_Techniques](https://www.researchgate.net/publication/220173171_A_Survey_of_Collaborative_Filtering_Techniques)[https://www.jl.mit.edu/sites/default/files/page/doc/2018-05/22\\_1\\_6\\_Gadepally.pdf](https://www.jl.mit.edu/sites/default/files/page/doc/2018-05/22_1_6_Gadepally.pdf)
- [http://users.cs.northwestern.edu/~pardo/courses/eecs349/readings/Recommender\\_Systems\\_An\\_Introduction\\_Chpt2.pdf](http://users.cs.northwestern.edu/~pardo/courses/eecs349/readings/Recommender_Systems_An_Introduction_Chpt2.pdf)<http://facweb.cs.depaul.edu/mobasher/classes/ect584/Papers/ContentBasedRS.pdf><https://www.semanticscholar.org/paper/A-hybrid-recommender-system-using-multi-layer-Sanandaj-Alizadeh/13671855cb404f373beb5952a84d52182552661><http://cites.eerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.9607&rep=rep1&type=pdf>