Perszonalizált tartalomajánló szolgáltatás IPTV és OTT rendszerek számára

ZIBRICZKY DÁVID
ImpressTV
david.zibriczky@impresstv.com

Kulcsszavak: ajánlórendszer, IPTV, OTT, adatbányászat, gépi tanulás

A Netflix Prize óta ugrásszerű kereslet figyelhető meg az IPTV és OTT piacon az ajánlórendszerek által nyújtott üzleti lehetőségek iránt. Az egyre növekvő lineáris és nemlineáris tartalom kínálat személyre szabott pozícionálása, valamint a tartalomfogyasztási adatok feldolgozása mind adatbányászati, mind technológiai oldalról kihívást jelent.
A szolgáltatók továbbá a heterogén médiatartalom-források, valamint a különböző megjelenítő felületek elterjedése miatt üzleti sikerességük megtartása érdekében olyan platformfüggetlen megoldásokat keresnek, melyek egységes módon képesek kezelni a kontextusfüggő ajánlási problémákat. Jelen tanulmány a CRISP-DM módszertan mentén ismerteti az IPTV és OTT környezetben alkalmazott ajánlórendszer megoldásokat, kitérve az aktuális főbb kutatási irányokra.

1. Bevezetés

Az utóbbi tíz évben a médiatartalom-fogyasztási trendek szignifikáns változást mutattak a digitális fejlődés hatására, az internetes szolgáltatások bővülésével több időt töltünk videó tartalmak fogyasztásával, mint valaha. A legmeghatározóbb szereplőkké vált Netflix és YouTube tartalmi és fogyasztói bázisában rohamos növekedést lehetett megfigyelni, így piaci előnyük megtartásának érdekében a TV-szolgáltatók igyekeztek termékpalettájukat egyaránt növelni újabb csatornák és előfizetési csomagok bevezetésével. Az IPTV-rendszerek elterjedésével és a „set-top-box"-ok (STB) megjelenésével új funkciókat vezettek be, mint például személyes videórögzítő (PVR), időeltolásos tévézés, elérhetővé váltak további nemlineáris tartalmak, mint például a videotéka filmjei, vagy a már korábban sugárzott műsorok archívuma.

Az „over-the-top" (OTT) szolgáltatások elterjedésével ezen tartalmak már nem csak televízión, de bármely más megjelenítő felületen is elérhetők, ezzel szélesítve a tartalomfogyasztások változatosságát. Az elérhető tartalmak kibővülése ugyan nagyobb kínálatot eredményez a végfelhasználóknak, mégis egyre nehezebben kezelhetővé válik még a műsorújság, megfelelő menüstruktúra és kereső funkciók alkalmazásával is. A TV-szolgáltatók emiatt olyan platformfüggetlen megoldásokat keresnek, melyek támogatást nyújtanak a felhasználóknak a megfelelő tartalmak megtalálásában, növelve ezzel a felhasználói élményt és piaci penetrációjukat.

Ezen probléma orvoslását hivatott szolgálni az ajánlórendszerek [1] bevezetése, amelyek adatbányászati algoritmusok segítségével különböző felületeken személyre szabott ajánlásokat nyújtanak a felhasználónak, ezzel elősegítve a megfelelő tartalmak megtalálását. Egyrészt a TV-szolgáltató által elérhető adatokat, másrészt külső információforrásokat alkalmazzák a tartalmak modellezésére és a felhasználók adaptív profilozására. Rendszer szinten külön funkcionális egységként működnek a háttérben, melyek az ajánláskérések során rendezik az elérhető tartalmakat, amiket ezután az eszközök felületén jelenít meg a szolgáltató. A személyre szabás eredményeképpen nő a felhasználói élmény, ami közvetetten az üzleti sikerességi mutatókat is növeli.

A tanulmány az ajánlórendszerek IPTV és OTT rendszerekben történő alkalmazását a CRISP-DM módszertan alapján mutatja be. A CRISP-DM [2] egy robusztus, széleskörűen alkalmazott módszertan adatbányászati projektek feladatainak leírására, ami hat fő fázisból áll:

(1) üzleti modell megértése, célok megfogalmazása,
(2) az adatok megértése,
(3) az adatok előkészítése,
(4) modellezés,
(5) kiértékelés és
(6) telepítés és üzemeltetés.

Ezen vezérfonal mentén haladva a 2. szakaszban az ajánlórendszerrel kapcsolatos üzleti elvárásokat tárgyaljuk, majd összefoglaljuk a tartalomfogyasztással és metaadatokkal kapcsolatos adatelemzési és feldolgozási kérdéseket. A 4. szakaszban bemutatjuk az ajánlórendszer területén leggyakrabban alkalmazott modellezési módszereket, melyre vonatkozó kiértékelési és optimalizálási megfontolásokat az 5. szakaszban vitatjuk. Ezt követően tömören kitérünk az ajánlórendszer, mint éles szolgáltatás legfontosabb üzemeltetési kérdéseire, végül az utolsó szakaszban áttekintjük az aktuális kutatási irányokat, mellyel a tudományos világ foglalkozik az ajánlórendszerek területén.

2. Az ajánlórendszer és az üzleti célok

Az ajánlórendszer egy olyan információszűrő és döntéstámogató szolgáltatás, mely az adott kontextusban adatbányászati algoritmusok segítségével a fogyasztói

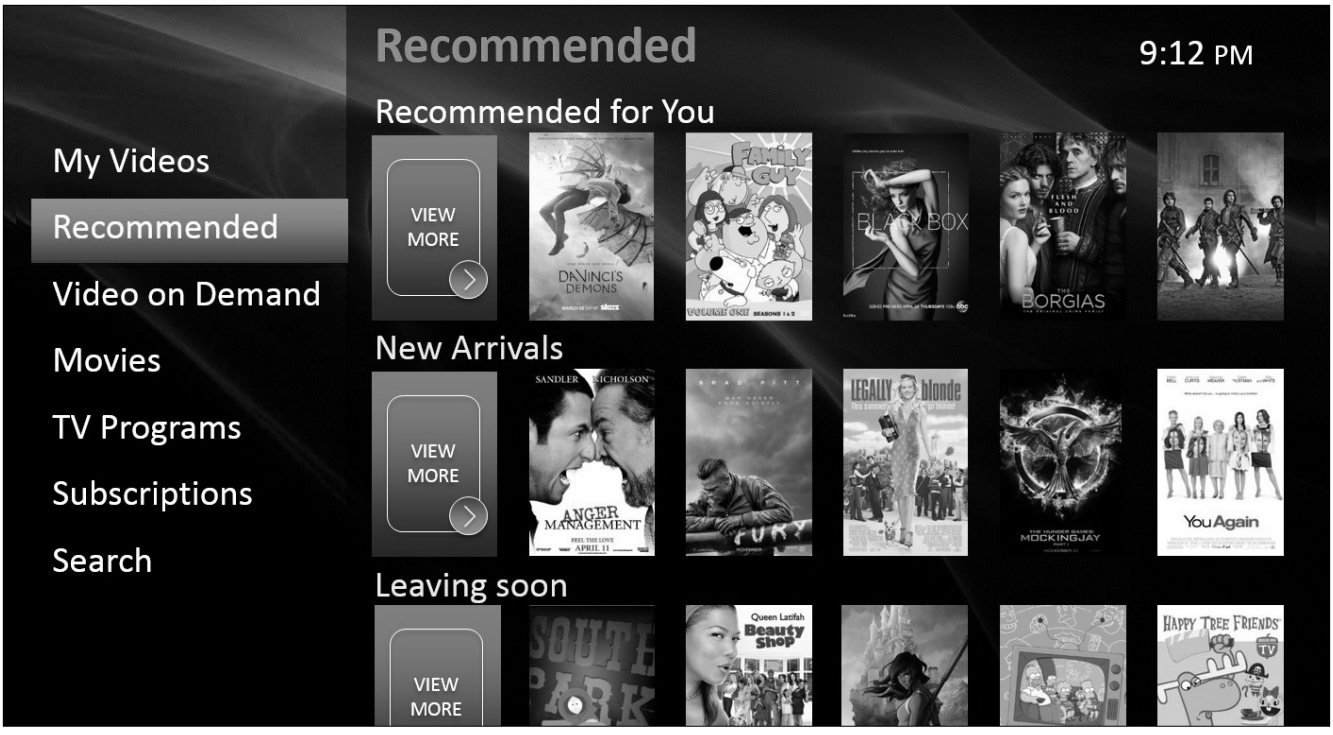
preferencia szerint személyre szabott termékajánlást nyújt. Részletezve a definíciót, a megoldás célja az, hogy az elérhető tartalmak sokaságát megszűrje és olyan listát kínáljon a végfelhasználóknak, mely nagy valószínűséggel érdekes lesz neki. A hagyományos keresési módszereket hivatott felváltani, mely elősegíti az elérhető tartalmak felfedezését, ezzel megkönnyítve a végfelhasználók választási döntéseit, melyet az *1. ábra* szemléltet. Egy ajánlórendszer külön, független modulként funkcionál, ami megfelelő interfészeken kommunikálva egyrészt gyűjti az információt, másrészt kiszolgálja az ajánláskéréseket. Az információ feldolgozása és a személyre szabott ajánlási listák előállítása adatbányászati probléma, melyeket különböző megközelítésekkel oldanak meg, figyelembe véve a kontextust leíró paramétereket (például idő, hely, eszköz típusa).

A végfelhasználók szemszögéből a szolgáltatás felé irányuló implicit elvárás egyrészt az, hogy a felhasználó minél hamarabb megtalálja a preferált tartalmakat, elrejtve előle a számára irreleváns lehetőségeket, másrészt változatos, friss és érdekes listákat mutasson, amire a felhasználó esetleg nem is gondolna először. A felhasználói élmény növelésével emiatt érdekesebbnek találja a TV-szolgáltató által elérhető tartalmakat és valószínűbben fog visszatérni, vagy többet fogyasztani. Bár az ajánlórendszer közvetlenül a felhasználói élmény növelésére irányul, végső soron üzleti érdekeket hivatott szolgálni. Üzleti szempontból az ajánlórendszer elsődleges célja a sikerességi mutatók növelése, a fogyasztási statisztikák nyomon követése, valamint támogatás nyújtása promóció és szegmentálás esetén. A lineáris TV fogyasztások esetén üzleti cél lehet például további előfizetési csomagok értékesítése, továbbá a jelenlegi ügyfélbázis megtartása az előfizetett csatornákon elérhető tartalmak felé történő érdeklődés megtartásával. Másik fontos cél a fizetős „video on demand” (VoD) tartalmakat fogyasztó – alapvetően csekély – felhasználóbázis bővítése, valamint a vásárlások összértékének növelése. Webes OTT felületek esetén további üzleti cél a hirdetések megnézésének és az azokra történő kattintások számának növelése, melyet a felhasználói élményből adódó tartalomfogyasztás növelésével érhetnek el. Ezen kívül említést érdemel még a kampányok során megcélzott felhasználói csoportok megtalálása, melyben az ajánlórendszer a fogyasztási mintázatok alapján nyújt támogatást.

A szolgáltatással kapcsolatos végfelhasználói és üzleti érdekek egyaránt hasonlóak, és különbözőek is lehetnek. Egyrészt, egy ingyenes videómegosztó oldalon például a felhasználói élmény növekedése több tartalom fogyasztásában mutatkozik, ami az üzletnek is előnyös, mivel magasabb lesz a hirdetésekre történő átkattintás száma is. Másrészt viszont egy VoD-szolgáltatás esetén bár az üzleti igény a bevétel növelése, a végfelhasználók nem feltétlenül többet szeretnének költeni, hanem saját preferenciájukat szeretnék kielégíteni tartalom csomagok vásárlásával. Emiatt az ajánlórendszer tervezőknek egyaránt szem előtt kell tartani mind a felhasználói, mind az üzleti igényeket.

A tartalomajánlások széles skáláját különböztethetjük meg. A legelterjedtebb a személyre szabott ajánlási lista, valamint hasonló tartalmak ajánlása, továbbá megemlíthető még a többsoros, zsáner preferencia szerinti rendezés, a közösségi hálók integrációjával hasonló ízlésű felhasználók, vagy csoportok ajánlása, e-mailben történő kampányok folytatása, célzott hirdetések, vagy éppen az ajánlások szöveges formában történő

1. ábra Személyre szabott ajánlási felület IPTV rendszerben



magyarázata. A jelenlegi trendek alapján világosan láthatjuk, hogy az ajánlórendszerek adta lehetőségeket a TV-szolgáltatók igyekeznek minél több formában kihasználni.

3. Az adatok megértése és feldolgozása

Az IPTV rendszerek elterjedésével, valamint a funkciók kibővülésével nagy mennyiségű nyomon követhető adat keletkezik, melyben lévő információtartalom kiaknázása jelentős üzleti értékkel bírhat. Alapvetően kétféle adattípust különböztetünk meg, a metaadatokat, illetve a fogyasztási adatokat. A szolgáltatók nyilvántartanak egy termékeket, tartalmakat leíró metaadatbázist. Ezen adatbázis olyan adatokat tartalmaz, (1) melyek a tartalom leírására szolgál (például cím, zsáner, színész lista, rendező), (2) technikai paramétereket ír le (például minőség, csatorna, sugárzási időpont), illetve (3) üzletileg fontos információ (például ár, előfizetői csoportok, licenc). A felhasználóról rendelkezésre álló metaadatok jellemzően a nem, kor és lakhely, továbbá esetenként a felhasználók a regisztrációkor kitölthetnek egy kérdőívet, melyben megadhatják a tartalmakra vonatkozó preferenciájukat is (például kulcsszavak, zsánerek, értékelési tartományok). A tartalmakat leíró metaadatból jellemzően több áll rendelkezésre, sőt külső források segítségével bővítetők is.

Az adattípusok másik csoportja a fogyasztási adatok (interakciók), melyek a tartalmak és a felhasználók között létesítenek kapcsolatot. Megkülönböztetünk ún. „explicit” visszajelzést, ami a felhasználó preferenciájának egyértelmű visszajelzése (például értékelés), illetve „implicit” visszajelzést, mely az interakciót leírja ugyan, de nem egyértelmű információtartalommal bír annak preferencia értékéről (például csatornaváltás, filmkölcsönzés, adatlap-megtekintés). Míg az explicit visszajelzés jellemzően tisztább információforrás, de kevés van belőle, addig a zajosabb implicit visszajelzésekből nagyságrendekkel több áll rendelkezés. Jelentőséggel bír az események kontextusa, mely olyan paraméter-együttes, ami az interakció bekövetkezése során leírták a rendszert. Explicit módon ide sorolható az idő, a napszak, a hét napja, ünnepnap van-e, a felhasználói készülék típusa, a böngésző típusa, időjárási tényezők, implicit módon pedig a felhasználó kedve, illetve, hogy kik ülnek a készülék előtt. A következőkben áttekintjük az IPTV-rendszerekben legjellemzőbb, lineáris- és nemlineáris tartalomfogyasztáshoz kapcsolódó specifikus problémákat, illetve az adatbővítési megközelítéseket.

3.1. Lineáris TV

A lineáris TV fogyasztások esetén a legjellemzőbb típus a tradicionális csatornák közötti váltogatás („channel zapping”). Ezen interakciók interpretálása nehéz feladat, mivel a felhasználó nem fejezi ki explicit módon a preferenciáját. A gyakori csatornakapcsolási interakció értelmezhető zajként, de értelmezhető negatív visszajelzésként is az adott műsorra vonatkozóan. Az adatok értelmezésének másik jellemző technikai nehézsége, hogy a felhasználó bekapcsolva hagyja a tévét a háttérben, vagy kikapcsolja ugyan, de a STB továbbra bekapcsolva marad, tovább generálva a nem releváns adatokat. Egy felhasználó akár ezer interakciót is generálhat havonta, így nagyobb felhasználóbázis esetén ezen adatok feldolgozása és tárolása technológiai kihívást jelenthet, illetve a gépi tanulási metódusok futtatása skálázhatósági megfontolásokat igényelnek. Az interakciós adatok jellemzően csatornákra vonatkoznak, a felhasználói preferencia modellezést viszont a műsorok alapján szeretnénk végezni. Emiatt szükséges egy idő alapú csatorna-műsor feloldás is a modellezés és ajánlás során.

A tévénézési szokások elemzése alapján megfigyelhető a műsorok időbeli preferenciája, például reggel híreket, délután sorozatokat nézünk. Általános nehézséget okoz az, hogy nem tudjuk eldönteni, ki ül a televízió előtt, így problémás a jellemzően többfős háztartás televíziózási preferenciáit megkülönböztetni. A lineáris TV sajátossága, hogy egy adott időpillanatban viszonylag kevés (csatornánként csak egy) tartalom érhető el, időben ezek azonban folyamatosan változnak. Előfordulhat, hogy a felhasználó preferenciáját az ajánló algoritmus megfelelően azonosította, de nem sugároznak számára releváns tartalmakat. További nehézséget okoz a hangulat detektálása az aktuális időpillanatban, illetve a megjelenítő eszköztől függő preferencia kezelése.

Az ajánlórendszerek jellemző problémája az ún. hidegindítási probléma („cold-start problem”), mely az olyan tartalmak, vagy felhasználók modellezési nehézségét jelenti, akire nem, vagy csak nagyon kevés fogyasztási adat áll rendelkezésünkre. Ekkor az őket leíró metaadatokra kell támaszkodnunk, ám ezek sok esetben hiányosak, vagy kevésbé informatívak. A lineáris TV sajátossága, hogy lényegében minden tartalom új, mivel még nem került lejátszásra. Bár az ismétlések és sorozatok esetén a probléma megoldható metaadat alapú csoportosítással, az egészestés filmek esetén továbbra is fennáll a nehézség. Felhasználói oldalról is jelentkezhet hidegindítási probléma, elsősorban jogi akadályok esetén, amikor a felhasználó nem egyezik bele abba, hogy harmadik fél felhasználja a fogyasztási történetet.

3.2. Nemlineáris TV

A Video on Demand (VoD) tartalmak esetén üzleti modelltől függően fizetés alapú fogyasztás történik, amely jellegében eltér a lineáris TV-től. A felhasználók jobban megfontolják, hogy mire költenek, így az adat tisztább, viszont kevesebb fogyasztási történetet is generálnak. Az adatok nagyságrendjét csökkenti az is, hogy a teljes TV előfizetői kör csak egy része fogyaszt ilyen típusú tartalmakat. Számottevő a felhasználói hidegindítási probléma a VoD tartalmak esetén, mivel azok szignifikáns része nem fogyaszt ilyen termékeket. A probléma kézenfekvő megoldása a lineáris TV fogyasztási preferenciáinak alkalmazása VoD tartalmak ajánlására, melyek kereszt-ajánlási módszernek hívnak. A lineáris tar-

talmak között azonban számos olyan található, melyből kevésbé tudunk következetni a VoD preferenciára, például hírműsorok alapján nehéz megbecsülni, hogy melyik egészestés film tetszene a felhasználónak, így ezen műsorok relevanciáját alul kell súlyozni a gépi tanulás során. Egyre elterjedtebb tartalom típus a lineáris TV tartalmainak archívuma („catch up" tartalmak), melyeket bizonyos ideig újranézhetik a felhasználók. Mivel a lineáris fogyasztás esetén ezen tartalmakra jó esetben már érkezett információ, nincs már jelen a tartalom hidegindítási probléma. Másrészt ezeket a tartalmakat menürendszerből érheti el a felhasználó, kevésbé zajos, így jobb minőségű adat keletkezik ezen fázisban, hasonlóan a VoD tartalomfogyasztáshoz.

3.3. Adatbővítés

Az utóbbi években az ajánlórendszer versenyszférában egyre elterjedtebbé vált a külső adatforrások alkalmazása az ajánlások minőségének javításának érdekében. Legjellemzőbb külső információforrások a metaadat-szolgáltatók, illetve a közösségi hálók. A metaadat-szolgáltatók (mint például a Gracenote, DBpedia vagy IMDb) leíró adatokat tartanak nyilván médiatartalmakról. Megfelelő kapcsolódási pontokon (például cím, sugárzási időpont, csatorna) a TV-szolgáltatók által elérhető tartalmak adatai tovább bővíthetők. Mivel a tévés tartalmak halmaza jól körülhatárolható, magas lefedettség érhető el a metaadat-szolgáltatók által nyilvántartott adatokkal (a gyakorlatban kivétel ez alól a sportközvetítések és hírműsorok). Ennek ellenére adatbányászati probléma a hibás, többértelmű és a hiányzó adatok kezelése, valamint technológiai kihívás a külső források adatainak folyamatos letöltése és a centralizált adatbázis karbantartása.

A közösségi hálókon (például Facebook, Twitter vagy Google+) jelentős mennyiségű információ érhető el a médiatartalmak iránti preferenciáról. Egyrészt kollektív népszerűségi (szezonális) trend mérhető egy adott filmről vagy műsorról (például mennyi és milyen hangvételű posztokat írnak róluk), másrészt egyéni szinten is nyomon követhető, ki milyen tartalmakat kedvel, illetve mely felhasználókat követ.

Egyrészt ezen adatokra illesztett adatbányászati megoldások javíthatják a TV szolgáltatónak nyújtott ajánlások pontosságát (elsősorban hidegindítási probléma javításával és szezonális trendek detektálásával), másrészt viszont a közösségiháló-alapú személyre szabott ajánlás nehézsége, hogy a TV/OTT felhasználók jellemzően csekély arányban rendelkeznek közösségi profillal.

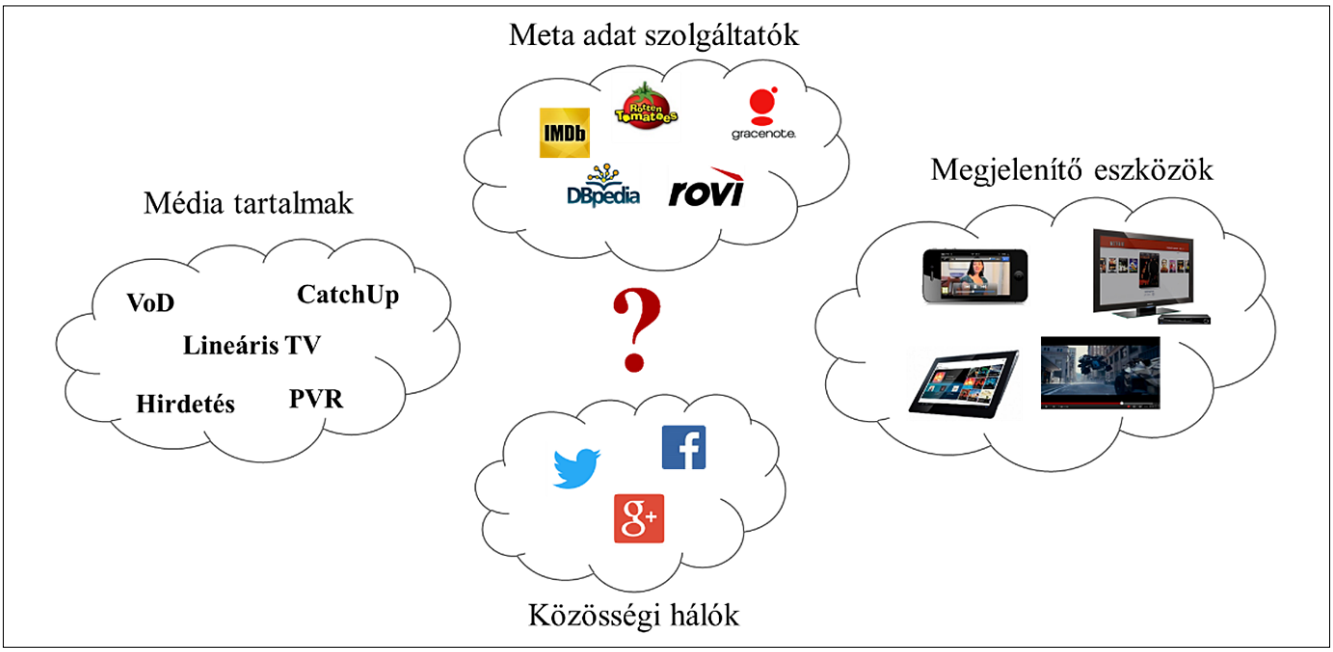
4. Modellezés

Az ajánlási probléma az ajánlórendszerek népszerűsítését eredményező Netflix Prize [3] idején a filmek értékelésének legpontosabb becslését jelentette. Mivel a hangulati faktor és a népszerűségi hatás jelentős szereppel bír abban a tekintetben, hogy a végfelhasználó mit szeretne nézni, az értékelés alapú célfüggvény nem bizonyult üzletileg túl sikeresnek, így az igények fejlődésével ezek átalakultak kontextus függő sorrendezés és felületoptimalizálási problémává. További elvárás az adaptív, újszerű, változatos és minden tartalmat lefedő algoritmusok alkalmazása. Jelen trendnek megfelelő modellezési probléma a különböző tartalomtípusok eszközfüggő modellezése és ajánlása külső heterogén adatforrások bevonásával, melyet a *2. ábra* szemléltet.

Az ajánlórendszerrel szemben támasztott üzleti elvárások kielégítésére alkalmazott módszereket öt különböző csoportba oszthatjuk:

(1) szerkesztői ajánlások;
(2) népszerűség alapú ajánlások;
(3) tartalom-alapú szűrés;
(4) kollaboratív szűrés;
(5) hibrid- és kombináló módszerek.

2. ábra Heterogén ajánlási probléma



4.1. Szerkesztői ajánlások

A szerkesztői ajánlások kézzel definiált ajánlási listák, melyek a legegyszerűbb ajánlási formák. Segítségükkel egy marketinges egyértelműen meghatározhatja, mit szeretne látni az ajánlódobozokban. A módszer előnye, hogy gyakorlatilag nincs szüksége adatra, egyszerű és meghatározott célt szolgál, mivel emberi beavatkozással áll elő. Hátránya, hogy nem személyre szabott (legfeljebb célcsoportra) és folyamatosan karban kell tartani.

4.2. Népszerűség alapú ajánlás

A termékfogyasztási mintázatokban megfigyelhető egy népszerűségi hatás. Ez alatt azt értjük, hogy a végfelhasználók hajlamosabbak népszerű termékek vásárlására, gyakran a saját preferenciájuk ellenében is. Az ajánló algoritmusnak figyelembe kell vennie ezt a hatást, ahhoz, hogy a legpontosabban el tudja találni a felhasználói fogyasztási preferenciákat. Másrészt a népszerűségi faktor modellezése gyakran alkalmazott módszer új felhasználóknak adott ajánlás során. Az újonnan érkező végfelhasználókról kezdetben nem tudunk semmi, így a saját preferenciájára vonatkozólag csak közelítéseket tudunk tenni a tömeg preferenciájának alkalmazásával, melyre legkézenfekvőbb ajánlási módszer a népszerű termékek ajánlása. Szofisztikázható az ajánlás felhasználói metaadatok alkalmazásával, melynek során csak az adott csoporton belüli népszerűséget mérjük. A módszer előnye, hogy közelítést tud nyújtani a felhasználói hidegindítási problémára, illetve bizonyos esetekben, ahol erős a népszerűségi hatás, jól működik. A módszer gyengesége, hogy nem képes személyre szabott ajánlások adására, mivel nem használja egyéni szinten a felhasználói fogyasztási történetet, még akkor sem, ha az rendelkezésre állhat.

4.3. Tartalom alapú szűrés

A tartalom alapú szűrés [4] („content-based filtering”, CBF) elve szerint két tartalom akkor hasonló, illetve egy felhasználói preferenciára (például a 80%-ban vígjátékot 20%-ban pedig drámát néz) egy tartalom akkor illeszkedik, ha az ajánlásban résztvevő termék leíró meta adatai szignifikáns fedésben vannak egymással. A „szűrés” kifejezés arra vonatkozik, hogy az ajánlás során a metaadatok mentén kiszűrjük azon elemeket, melyek nem relevánsak az adott preferenciához, azaz nincsenek megegyező adataik. A tartalomra vonatkozó metaadatokon kívül alkalmazható a felhasználókra vonatkozó információ is, ezzel pontosítva az ajánló profilozását.

A metaadatok – elsősorban tartalmi leírások – értelmezésében alkalmaznak ún. természetes nyelvfeldolgozó eszközöket is, melyek egyrészt képesek kifejezések kinyerésére, illetve bonyolultabb szemantikai összefüggéseket feltárására, elősegítve a tartalom alapú szűrés pontosságát. A CBF módszer leggyakrabban használt algoritmusai a metaadat-egyezési arány és a koszinuszos hasonlóság alapú metódusok. Előnye, hogy megoldja a tartalmak hidegindítási problémáját, az ajánlások explicite megmagyarázhatók, valamint nagy lefedettséget mutatnak a katalógus terén. Hátránya viszont, hogy támaszkodik a metaadatok minőségére, valamint nem képes azok között átjárni.

4.4. Kollaboratív szűrés

A végfelhasználók preferenciáit az általuk megadott adatok mellett azok interakcióiból lehet tovább finomítani. A felhasználói interakciók segítenek a felhasználói szokások megértésében, illetve a preferencia modell finomításában. Ezen információ alapján nem csak a felhasználói preferencia érthető meg pontosabban, hanem viselkedésmintázatok felismerése. A kollaboratív szűrés [5] („collaborative filtering”, CF) a felhasználói bázis fogyasztási szokásaiban kinyert információt alkalmazza, mely szerint hasonló felhasználók hasonló jövőbeli tartalmak/termékek iránt érdeklődnek. A CF módszer szerint két felhasználó hasonló, ha sok azonos tartalmat fogyasztottak, illetve két műsor hasonló, ha sok felhasználó látta mindkettőt.

A „szűrés” kifejezés ebben az esetben olyan tartalmak kiszűrését sugallja, melyeket a hasonló felhasználók sem fogyasztottak, így azok valószínűleg nem relevánsak. Leggyakrabban alkalmazott algoritmusai a legközelebbi szomszéd módszerek [5], a mátrix faktorizáció [6] és az asszociációs szabályok [7]. A CF módszer előnye, hogy nem feltételezi a metaadatok meglétét, csak a látens fogyasztási mintázatokat az interakciós adatsorban. Képes olyan preferenciákat feltárni, melyet metaadatokkal kevésbé pontosan lehet modellezni. Hátránya viszont az, hogy a hidegindítási problémára nem tud megoldást adni, hiszen szükséges számára az interakciós történet megléte, illetve az ajánlások közvetlenül nehezen magyarázhatók.

4.5. Hibrid- és kombináló módszerek

A hibrid szűrés („hybrid filtering”, HF) ötvözi a CBF és CF előnyös tulajdonságait [8]. Egyidejűleg próbálja megoldani a hidegindítási problémát a tartalom és felhasználók leíró metaadatai segítségével, valamint kinyerni az interakciós adatokban rejlő fogyasztási mintázatokat. A módszer a kombinálás mellett nemcsak a gyengeségek erősítését célozza meg, de képes összefüggéseket feltárni két szó között, valamint hiányos metaadatokra javaslatot tenni és inkonzisztens címkézését detektálni (például, ha egy vígjáték akciónak van címkézve, de olyanok nézik, akik jellemzően vígjátékot szeretnek, akkor a módszer detektálja, hogy a címke nincs összhangban a tartalomra vonatkozó preferenciával). Napjainkban a hibrid modellezés a legelterjedtebb forma, leggyakrabban alkalmazott módszerek a hibrid faktorizációs modellek [9], valamint a CF és CBF algoritmusok kimeneteinek lineáris, vagy személyes preferencia szerinti kombinációja. Nem szerves része a hibrid szűrésnek, de kombinálási módszer még a marketingesek által definiált kimeneti logika, mely több ajánlási ágból választ tartalmakat, például 10 ajánlott tartalom között szerepeljen pontosan 4 lineáris és 6 VoD tartalom.

5. Kiértékelés

Az ajánlórendszerek optimalizálási folyamatában fontos szerepet játszik a mérési módszer és a célfüggvények helyes megválasztása. A kiértékelési módszerek két alapvető fajtáját különböztetjük meg: (1) offline, vagy megfigyelési adatsoron történő kiértékelés; illetve (2) online, vagy élesített szolgáltatás által mért teljesítmény. Ennek alapján egy kétlépcsős optimalizálási módszert alkalmazunk.

5.1. Offline mérés

Az offline kiértékelés egy statikus megfigyelési adatsoron történő mérési módszer, melyet teljesen függetlenül végeznek a valós rendszertől. Első lépésben, ezen mérés során az algoritmusok paraméterhangolását és kombinálási súlyok beállítását végzik. Az adatsor két részre történő felosztása eredményeképpen előáll egy tanító adatsor, melyen az optimalizálást végezzük, illetve egy teszt-adatsor, melyen méréseket végzünk. Ahhoz, hogy a valós rendszerhez legközelebbi szimulációt végezzük, az adatsort időpont szerinti vágással célszerű felosztani.

Az ajánlórendszerek területén a pontosság kiértékelésére leggyakrabban alkalmazott mérőszámok [1] explicit adatsoron a RMSE („root mean squared error”), implicit adatsoron a recall, precision és nDCG („normalized discounted cumulative gain”). A pontosságon kívül érdemes szem előtt tartani az ajánlási metódusok diverzitását entrópia méréssel, tartalom lefedettségét („coverage”), illetve termékjellemzők szerinti előfordulási arányt az ajánlási listákban (például népszerű, vagy friss elemek aránya). Mivel az offline mérés során az algoritmusokat egy független adatsoron értékeljük ki, nem lehet pontosan következtetni arra, milyen hatással lesznek a fogyasztásra, így az offline optimalizálás során előállított algoritmus nem feltétlenül lesz optimális az éles környezetben is. Ennek ellenére, az így beállított algoritmus jó kezdő konfigurációja lehet az éles környezetben történő optimalizálásnak.

5.2. Online mérés

Második lépésben az online mérés során közvetlenül az ajánlórendszer hatásait mérjük, melynek optimalizálási módszere az ún. „A/B tesztelés”. Ennek során a felhasználói bázist két- vagy több diszjunkt halmazra osztunk, melyeket egyidejűleg, különböző algoritmusokkal szolgálunk ki. A módszer referencia algoritmusa az „A” jelű algoritmus, melyhez képest jobb eredményt szeretnénk elérni. Annak eldöntésére, hogy egy adott mérési időszak alatt „B” jelű algoritmus jobban teljesített-e a referenciánál, statisztikai próbákkal döntjük el. Ha jobb eredményt érünk el, a legjobb algoritmust választjuk referencia algoritmusnak, és újrakezdjük a mérést. Lineáris TV-fogyasztás esetén a leggyakrabban alkalmazott mérőszámok (1) a televíziózás idejének hossza; (2) annak aránya, hogy a nézők végignézik a műsort; (3) illetve, hogy a műsorok időtartamának átlagosan hány százalékát nézik végig. VoD fogyasztás esetén üzletileg a legfontosabb mérőszámok a forgalom értéke és darabszáma, valamint a konverziós ráta. OTT megoldások esetén a felületből adódóan érdemes még az átkattintási arányt („click through rate”), illetve az odalátogatottságot mérni („page impression”).

6. Telepítés és üzemeltetés

Mint ahogyan a korábban említettük, az ajánlórendszer funkcionálhat beépített, vagy külön modulként is az IPTV szolgáltató rendszerében. Az ajánlórendszerek kétféle telepítési formája elterjedt, függően attól, ki üzemelteti az ajánlórendszer szervereit. Egyrészt üzemelteteti maga a szolgáltató a saját („on site”), illetve történhet külső, jellemzően az ajánlószoftvert gyártó cég környezetében („software as a service”). Mindkét esetben fontos tényező az IPTV szolgáltató és az ajánlórendszer adatbázisa közti szinkronizálás gyakorisága (mely akár egy nap is lehet), az ajánló algoritmusok tanítási ideje és gyakorisága; a szolgáltatás válaszideje (melynek ipari sztenderdje 100 ms), illetve a rendszer rendelkezésre állása (melyek ipari sztenderdje IPTV rendszertől függően 3 és 4 „kilences” között alakul). Technológiai oldalról említést érdemel az egyszerveres és az elosztott rendszerű megoldások közötti választás, mind az adatbázis, a kiszolgálás, mind az algoritmus futtatás terén.

Míg a kiszolgálás esetén a beérkező ajánláskérések terheléselosztása több szerver között egyszerűbb, az elosztott adatbázisok transzparens kezelése már nehezebb feladat, továbbá az algoritmus tanítások elosztott párhuzamosítása bonyolult probléma, mivel egyrészt párhuzamosítható algoritmusok esetén működhet csak hatékonyan, másrészt fontos tényező a szálak közti kommunikációs időtöbblet minimalizálása, mely aktuális kutatási irány az ajánlórendszerek területén.

7. Aktuális kutatási irányok

Az ajánlórendszer szolgáltató cégek esetében megfigyelhető trend a külső heterogén adatforrások központi integrációja, melynek bevonásával bonyolultabb algoritmusokra van szükség, amelyek hiányos adatokkal is tudnak dolgozni, képesek detektálni az inkonzisztenciákat, valamint összekötni az azonos entitásokra érkező információt. Aktívan vizsgált terület a keresztajánlási módszerek hidegindítási problémákra történő alkalmazása, faktorizációs automaták használata hibrid szűrési problémákra és hiányos információ kezelésére, a többrétegű neurális hálók alkalmazása („deep learning”), továbbá automatikus meta adat címkék generálása (például „romantikus tini vígjáték”), mely segítségével részletesebb preferencia kategóriák és beszédesebb magyarázatok állíthatók elő a felhasználóknak.

Nehéz gyakorlati probléma annak detektálása, hogy ki ül a televízió előtt, illetve milyen hangulatban van éppen, így aktuálisan kutatott téma olyan algoritmusok tervezése, mely képesek több preferenciát egyidejűleg kezelni, illetve a fogyasztási preferenciában történő változásokat detektálni. Érdekes kutatási terület az ajánl-

lási stratégiák alkalmazása, amely spekulatív ajánlásokkal próbálja a lehető legtöbb információt megszerezni a felhasználó preferenciáiról.

8. Összefoglalás

Az ajánlórendszerek iránt jelentős kereslet figyelhető meg a TV-piacon. A digitális fejlődés hatására nagy mennyiségű adat vált elérhetővé, melyek alkalmasak mind a felhasználói élmény, mind az üzleti sikeresség növelésére. A TV-szolgáltatók által elérhető implicit visszajelzések értelmezése nehéz feladat, tovább a nagy mennyiségű belső és külső adatforrások centralizációja mind technológiai, mind algoritmikus kihívást jelent. A Netflix Prize óta számos módszer látott napvilágot, mely különböző problémákat hivatott megoldani. Ezen módszerek kombinálásával és kétlépcsős optimalizálásával az ajánlórendszerek egyedileg igazíthatók az üzleti igényekhez.

A sikerességi mutatókon kívül fontos szempont a rendszer adaptivitásának, skálázhatóságának és gyors válaszidejének biztosítása, mely különböző technológiai megfontolásokat igényel. Az ajánlórendszerek területén továbbra is számos kutatási irány vázolható fel, melyet mind az akadémiai-, mind az ipari szféra érdeklődéssel vizsgál.

A szerzőről

ZIBRICZKY DÁVID okleveles mérnök-informatikus, közgazdász doktorjelölt. 2010 óta foglalkozik ajánlórendszerekkel főállásban, mely során számos ügyfélprojekt teljes életciklusát követte nyomon, jelentős tapasztalatot szerezve adatelemzés, rendszerfejlesztés valamint algoritmus-optimalizálás és kutatás terén. 2014-ben részese volt egy TV-s üzletági akvizíciónak, jelenleg az ImpressTV adatbányászati és kutatási részlegének vezetője. Az ajánlórendszerek terén számos cikk társszerzője, a területhez kapcsolódó konferenciákon bíráló, valamint egyetemi hallgatók külső konzulense. Az informatika mellett a közgazdasági tudományokkal is aktívan foglalkozik. Befektetési elemzőként szerzett tapasztalatot, TDK/OTDK első díjas, BME rektori különdíjas, jelenleg doktori (PhD) fokozatszerzés legutolsó szakaszában jár.

Irodalom

[1] Kantor, P. B., Rokach, L., Ricci, F., Shapira, B., Recommender systems handbook. Springer, 2011.

[2] Shearer, Colin, „The CRISP-DM model: the new blueprint for data mining." Journal of data warehousing 5.4 (2000): 13–22.

[3] Bennett, James, Stan Lanning, „The netflix prize." Proceedings of KDD cup and workshop, Vol. 2007.

[4] Pazzani, Michael J., Daniel Billsus, „Content-based recommendation systems." The adaptive web. Springer Berlin Heidelberg, 2007. pp.325–341.

[5] Sarwar, Badrul, et al. „Item-based collaborative filtering recommendation algorithms." Proc. of the 10th Int. Conference on World Wide Web. ACM, 2001.

[6] Koren, Yehuda, Robert Bell, Chris Volinsky, „Matrix factorization techniques for recommender systems." Computer 8 (2009): 30–37.

[7] Lin, Weiyang, Sergio A. Alvarez, Carolina Ruiz, „Efficient adaptive-support association rule mining for recommender systems." Data mining and knowledge discovery 6.1 (2002): 83–105.

[8] M. Prem, R. J. Mooney, Ramadass Nagarajan, „Content-boosted collaborative filtering for improved recommendations." AAAI/IAAI, 2002.

[9] Barragáns-Martínez, Ana Belén, et al. „A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition." Information Sciences, 180.22 (2010): 4290–4311.