*Ana Tomac*

# »BIG DATA«

## Netflix

Trenutno je odprtih že več kot 200 milijonov Netflix računov, uporabnikov le-teh pa še mnogo več. Na Netflixovi prvi strani se v več vrsticah nahajajo priporočila za vsebine, ki jih ponujajo. Med kategorijami priporočil se nahaja tudi »Top Picks for Ana«, ki so, kar verjetno ni presenetljivo, odlično izbrane in se, sledeč po napovednikih, popolnoma ujemajo z mojim okusom za video vsebine.

Za natančno dodelanim modelom, ki na podlagi ogromnega števila podatkov predvidi, katere zvrsti vsebin so za posameznika primerne, stojijo skupine raziskovalcev, ki so razvili številne algoritme za izboljšavo. Personalizacija Netflix izkušnje je bazirana predvsem na izbiri zvrsti vsebin, primernih za vsakega uporabnika, katere vsebine sploh spadajo v posamezno prikazano zvrst ter vrstni red izbranih vsebin.

Za optimizacijo sistema personalizirane izkušnje uporabnika mora Netflix zbirati ogromne količine podatkov; ocene posameznih vsebin se gibljejo v milijardah, na dan pa jih pridobijo še več milijonov. Pomemben vir podatkov so tudi predvajanja vsebin, kjer se beleži ne le posamezna vsebina, pač pa tudi dolžina predvajanja, čas v dnevu in naprava, na kateri se predvaja. Prav tako prejmejo podatke o vsebinah, ki so jih uporabniki uvrstili na svoje čakalne sezname, zbirajo pa se tudi informacije o že predlaganih vsebinah – število klikov na strani, čas odprtosti strani in ali si je oseba pogledala napovednik, ki je na Netflixu omogočen tako, da se z miško postaviš na priporočilo, ki začne samodejno predvajati napovednik. Ne smemo pozabiti tudi na lokacijo, jezik in demografske podatke, dodaten vir informacije je tudi samo iskalno okence, kamor uporabnik zapiše ime serije oziroma filma ter tako preveri, ali ga ima Netflix v svoji bazi vsebin. Vsi zgoraj našteti viri podatkov so popolnoma interni, ki pa jih je potrebno združiti tudi z zunanjimi viri za doseganje kar najboljših rezultatov – primer takšnega vira so kritike in ocene vsebin iz drugih spletnih platform. Za izboljšanje modelov personalizacije priporočil so ključni podatki, pridobljeni iz zunanjih virov. Velikokrat namreč dodajanje takih virov informacij vodi v zanesljivejše modele za predikcijo, pa čeprav so lahko uporabljeni algoritmi manj dovršeni.

Število takšnih podatkov presega milijarde, tako da je to zagotovo primer »big data«, prav tako pa zadošča pravilu t.i. “3 Vs” - volume, velocity in variety. Volume se nanaša na količino ustvarjenih in shranjenih podatkov, ki se dnevno še povečujejo. Pri Netflixu in njegovih več sto milijonov uporabnikov, se tako obseg podatkov vsak dan poveča za milijone, tako da brez dvoma ustreza prvemu kriteriju. To pa se povezuje tudi z velocity - hitrosti ustvarjanja podatkov. Kot sem že omenila, vsakič, ko uporabnik odpre svoj profil na Netflixu, si ogleduje določeno vsebino ali jo oceni, se vse to shranjuje, tako da je pri 200 milijonih računih takih podatkov dnevno ogromno. Zadnji kriterij pa je variety, raznolikost podatkov. Podatki, ki jih zbirajo in shranjujejo, niso zgolj strukturirani, pač pa tudi nestrukturirani, prav tako pa se razlikujejo po svoji obliki, namreč vsa informacija ni zgolj v obliki tekstovnih podatkov, pač pa tudi v slikah, videih, lokaciji, idr.

Na podlagi takšnega števila podatkov je tako možno slediti in odkrivati vzorce, asociacije in trende, ki pripeljejo do pomembnih zaključkov in omogočajo tako velikim organizacijam doseči svoj cilj, ki je v primeru Netflixa zagotoviti uporabniku kar najbolj personalizirano izkušnjo, ki pa vodi v povišano gledanost. V tem in ostalih »big data« primerih velja, da več dostopnih podatkov vodi v boljše rezultate, kar pa je doseženo z optimiziranimi pristopi, metričnimi metodami in eksperimentiranjem.

Zelo pomemben element personalizirane izkušnje, ki so ga implementirali pri Netflixu, je zavedanje uporabnikov, da se na podlagi njihovih ocen in gledanosti prilagaja prikaz priporočil. To jim vliva zaupanje v sistem in na takšen način je promocija podajanja povratnih informacij zelo uspešna, kar pa tudi vodi v zanesljivejše in bolj točne podatke. Glede na to, da je seznam priporočil nekaj, ki koristi uporabniku, mu ni v interesu, da bi načrtno podajal neresnične informacije, zato sem mnenja, da ocene video vsebin spadajo med kvalitetne in točne podatke. Ostali viri informacij zajemajo podatki, ki jih uporabniki ne podajo zavedno, to so čas gledanosti določenega filma ali TV serije, spremljanje napovednikov serij, ki jih je Netflix že priporočil ter iskanja v sami bazi Netflixovih vsebin. Te podatke bi uporabniki le stežka poneverili, tako da bi jih tudi smatrala za točen vir podatkov. Problem pa nastane pri lokaciji – glede na državo, je rahlo prilagojena baza vsebin, ki so na voljo za ogled. Nekateri v ta namen izkoriščajo navidezno privatno omrežje oziroma VPN (virtual private network), da si lahko ogledajo vsebine, ki niso na voljo v njihovi matični državi. Moje mnenje je, da to verjetno vodi v nekoliko izkrivljene podatke, vendar glede na količino uporabnikov, je takih, ki uporabljajo VPN verjetno zgolj zelo majhen delež, kar pa v celoti nima drastičnega vpliva.

Kot sem že omenila, je prednosti, ki se pri uporabi takšnih podatkov pojavijo, veliko. Tako iz strani uporabnikov pa tudi Netflixa kot ponudnika, je videti, da je zbiranje in analiziranje takšnih podatkov v vzajemno korist. Uporabnikom je tako na voljo veliko raznovrstnih priporočil, namreč ravno diverziteta priporočenih vsebin, je po mojem mnenju ključnega pomena za vračanje uporabnikov na spletno platformo Netflix, za kar si le-ta vsekakor prizadeva. To pa naredi uporabniško izkušnjo še bolj preprosto in privlačno, saj uporabnik ne zapravlja nepotrebnega časa za pregledovanje precej velike baze dostopnih vsebin, preden odkrije eno, ki zadošča njegovemu interesu in trenutnemu počutju – Netflix namreč to naredi za uporabnike »sam od sebe«. Vsebine razvršča glede na uporabnikovo povratno informacijo, lokacijo, skupino oseb z enakimi interesi, pa tudi glede na trenutne trende.

Ima pa takšno zbiranje podatkov tudi svoje negativne plati, ena izmed katerih je prav zagotovo milijonska nagrada, ki jo je Netflix ponujal v zameno za rešitev za izboljšanje obstoječega modela za priporočila vsebin. Netflix je namreč objavil podatke, ki so vključevala informacije, prejete od 480.000 uporabnikov Netflixa. Sicer so imena izpustili in vsakega uporabnika označili z identifikacijsko številko. To pa ni preprečilo raziskovalcem, da so že po nekaj tednih po objavi podatkov identificirali več uporabnikov, in sicer prek primerjave anonimnih podatkov z ocenami, objavljenimi na spletni strani IMDB – Internet Movie Database. Razkritja pa niso osebe zgolj identificirala, pač pa so podala tudi informacijo o politični usmeritvi in spolni usmerjenosti. To pa bi lahko predstavljalo velik problem za identificirane posameznike – javno izpostavljanje njihove politične usmeritve in spolne usmerjenosti bi lahko negativno vplivalo na njihov družbeni položaj, kariero in prav tako tudi na njihove družine. Čeprav je bila prvotna ideja o iskanju rešitve s strani Netflixa zastavljena z dobrim namenom, potrudili so se namreč tudi zakriti identiteto uporabnikov, bi bil lahko izid popolnoma nasproten – številne tožbe in milijonske odškodnine. Do tega k sreči ni prišlo, vendar je Netflix vseeno preklical drugo tekmovanje, ki so ga že vnaprej napovedali.

Moje osebno mnenje je, da je z uporabo »big data« možno rešiti številne probleme in optimizirati izkušnjo uporabnikov – ne le v primeru Netflixa, pač pa tudi ostalih podobnih spletnih platform, ki jih predstavljajo predvsem socialna omrežja in t.i. personalizirane reklame. Netflix je preko konstantne izboljšave njihovega sistema priporočil prav zagotovo omogočil prijetnejšo in enostavnejšo uporabniško izkušnjo, vendar pa je treba s tako številnimi podatki ravnati previdno. Glede na število podatkov so »big data« pri tem najnevarnejši, saj lahko z uporabo le-teh pridemo do potrebnih

asociacij, s katerimi je lahko osebo preprosti identificirati. Pri delu s takšnimi podatkovnimi zbirkami bi bilo tako potrebno ozaveščanje o posledicah javnega izpostavljanja vseh podatkov ali zgolj določenih elementov, saj bi v nasprotnem primeru že dobronamerna odločitev lahko vodila do resnih posledic.

**Aha momenti:**

- Presenetilo me je predvsem dejstvo, da lahko preko ZIP kode, rojstnega datuma in spola z 87% možnostjo identificiramo osebo. Menim namreč, da se ljudje na splošno ne zavedamo zadosti, da je kraja identitete resen problem in tako ne pomislimo, preden zgoraj naštete podatke (in tudi še kaj več) objavimo na spletu.
- Da so podatki in njihova analiza, predvsem v primerih »big data«, koristni zgolj pri razumevanju vzorcev in analiziranju posameznih komponent, ne pa tudi pri sklepanju zaključkov na podlagi le-teh. Ključ pri rešitvi takšnih problemov je, da si s pomočjo podatkovnih analiz in velikih količin podatkov razčlenimo posamezne dele, kar bi bilo popolnoma nemogoče brez uporabe računalniških orodij, pri sprejemanju odločitev, pa je potrebno vključiti strokovnjake na dotičnem področju.
- Zanimivo se mi zdi tudi, da si je trenutno skoraj nemogoče predstavljati življenje brez takšnega ali drugačnega sledenja in shranjevanja osebnih informacij preko pametnih telefonov in ur, računalnikov, pa tudi televizije. Gre se o ogromnih količinah informacije, o katerih se niti ne zavedamo, pa vendar nam olajšuje življenje na vsakem koraku.