UNIVERZITET U BEOGRADU

ELEKTROTEHNIČKI FAKULTET

Marko Ž. Krstić

PERSONALIZOVANI PROGRAMSKI VODIČI ZA DIGITALNU TELEVIZIJU

doktorska disertacija

Beograd, 2018.

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTRICAL ENGINEERING

Marko Ž. Krstić

PERSONALIZED PROGRAM GUIDES FOR DIGITAL TELEVISION

Doctoral Dissertation

Belgrade, 2018

Mentor:

dr Milan Bjelica, vanredni profesor,

Univerzitet u Beogradu, Elektrotehnički fakultet

Članovi komisije:

dr Milan Bjelica, vanredni profesor,

Univerzitet u Beogradu, Elektrotehnički fakultet

dr Mirjana Simić Pejović, vanredni profesor,

Univerzitet u Beogradu, Elektrotehnički fakultet

dr Vladimir Čeperić, docent,

Univerzitet u Zagrebu, Fakultet elektrotehnike i računarstva

Datum odbrane:

ZAHVALNICA

Najpre bih želeo da se zahvalim mom mentoru, dr Milanu Bjelici na svemu što me je naučio o pisanju naučnih radova, na razumevanju i podršci koju mi je pružio tokom doktorskih studija.

Zahvalio bih se i mojim talentovanim kolegama i prijateljima, Aleksandri Stefanović, Zorani Nedić, dr Mariji Tasić i Aleksandri Malinić koji su svojim primedbama i diskusijama pomogli da moje istraživanje bude predstavljeno na lepši i bolji način.

Ipak, najveću zahvalnost dugujem mojoj porodici, koja mi je bila bezgranična podrška tokom ovog perioda, ali i tokom celog mog života:

- mojoj majci koja mi je još u detinjstvu usadila ljubav prema nauci i istraživanju,
- mom ocu na svim životnim savetima koji mi je dao i
- na kraju mojoj sestri koja je zajedno prošla sa mnom prošla kroz sve lepe ali i burne periode tokom ovih studija.

Ovaj, ali i svi moji budući uspesi posvećeni su njima.

Naslov teze: Personalizovani programski vodiči za digitalnu televiziju

Rezime – Razvoj digitalne televizije je doveo do značajnog porasta broja TV sadržaja dostupnih korisnicima, ali je otežao izbor onog koji je od interesa. Sve do pojave personalizovanih programskih vodiča sposobnih da nauče korisnička interesovanja i preporuče odgovarajuće sadržaje nije postojalo rešenje koje je na adekvatan način razmatralo ovaj problem. Ranija rešenja, kao što su štampani i elektronski vodiči, su pretežno samo pretvarala problem viška informacija u drugi oblik. Napredak tehnologije i društva postavlja sve veće zahteve pred personalizovane programske vodiče za digitalnu televiziju, što zahteva njihovo pažljivo planiranje i projektovanje. Vodiči moraju da budu u mogućnosti da modeliraju različite načine donošenja odluka pojedinačnih korisnika, da rade u realnom vremenu na mobilnim uređajima s ograničenim hardverskim resursima, da vode računa o karakteristikama prikupljenih podataka, da uzimaju u obzir kontekst u kome se pristupa TV sadržaju i da štite privatnost svih korisnika, jer neki od njih nisu svesni mogućih opasnosti. Pažljivim izborom arhitekture i algoritma učenja, lokalno implementiran vodič baziran na neuralnim mrežama može da ispuni sve ove zahteve. S obzirom na to da korisnici znatno češće pružaju informacije o sadržajima koji im se dopadaju nego o onim koji im se ne dopadaju, u ekstremnim slučajevima se dešava to da su prikupljene samo pozitivne interakcije. Da bi se taj problem prevazišao, predložen je sistem s dva režima rada. U prvom režimu sistem uči i pruža preporuke samo na osnovu TV sadržaja koje korisnik voli, dok u drugom izjednačava uticaj sadržaja koje korisnik voli i onih koje ne voli na proces pružanja preporuka. Povećan uticaj pozitivnih interakcija dovodi do degradacije predikcije sadržaja koje posmatrač ne želi da gleda, te će se, usled greške u klasifikaciji, neželjeni sadržaji često pojavljivati u listi preporuka i na taj način smanjiti zadovoljstvo korisnika. Korišćenjem serije simulacija pokazali smo da je postignuto trajanje treniranja neuralne mreže kratko, čak i na uređajima s ograničenim hardverskim resursima. Zaključak je da je predloženi vodič veoma pogodan za implementaciju na mobilnim uređajima od kojih se očekuje da u budućnosti postanu dominantan način pristupa TV sadržajima.

Ključne reči: digitalna televizija, preporučivači, neuralne mreže, zaštita privatnosti, disbalans klasa

Naučna oblast: tehničke nauke, elektrotehnika

Uža naučna oblast: telekomunikacije

UDK broj: 621.3

Dissertation title: Personalized program guides for digital television

Summary – The development of digital television significantly increased the quantity of media contents available to the users, but made it difficult to make the right choice. Before the invention of the personalized program guides capable of learning user preferences and recommending adequate contents, there were no means of properly addressing this problem. Former solutions, such as printed or electronic program guides, mostly converted the problem of having to deal with too much information into another form. The advancements in both technology and society put higher demands to the personalized program guides for digital TV, which require careful planning and design processes. Guides must be able to model various individual decision making approaches, work in real-time on mobile devices with limited hardware resources, take into account the characteristics of the collected data, take into consideration the program accessing context and protect the privacy of all users, since some of them are not aware of the possible risks. By carefully choosing the architecture and learning algorithms, a locally implemented guide based on neural networks can fulfil all the aforementioned requirements. Due to the fact that the users provide information about the content they like much more often than about the one they dislike, only positive interactions are collected in extreme cases. In order to overcome that situation, a system having two operating modes is proposed. The first mode enables the system to learn and give recommendations based on preferred TV contents, while the second equalizes the influence of the liked and disliked contents on the recommending process. The increased influence of positive interactions degrades the unwanted content prediction process, resulting in classification error, appearance of unwanted content in the recommendation list and user dissatisfaction. By applying a series of simulations, we showed the accomplished neural network training time to be short, even in cases of devices with limited hardware resources. It can be concluded that the proposed guide is very convenient for implementation on mobile devices which are expected to become a dominant way of accessing media contents in the future.

Keywords: digital television, recommender systems, neural networks, privacy protection, class imbalance

Scientific area: technical sciences, electrical engineering

Scientific subarea: telecommunication

UDK code: 621.3

SADRŽAJ

1. Uvod ... 1
1.1 Postavka problema ... 1
1.2 Predmet i cilj istraživanja ... 2
1.3 Polazne hipoteze ... 4
1.4 Metode istraživanja ... 4
1.5 Struktura disertacije ... 4
2. Preporučivači ... 6
2.1 Filtriranje sadržaja ... 6
2.2 Kolaborativno filtriranje ... 7
2.3 Hibridni sistemi ... 10
2.4 Dominantni pravci istraživanja kod preporučivača ... 11
2.4.1 Mere performansi sistema ... 13
2.4.1.1 Tačnost pružanja preporuka ... 15
2.4.1.1.1 Tačnost pružanja preporuka kod predikcije ocena ... 15
2.4.1.1.2 Tačnost pružanja preporuka kod klasifikacije objekata ... 16
2.4.1.1.3 Tačnost pružanja preporuka kod rangiranja sadržaja ... 20
2.4.1.2 Ciljevi provajdera servisa ... 23
2.4.1.3 Pokrivenost objekata ... 24
2.4.1.4 Pouzdanost preporuka ... 24
2.4.1.5 Poverenje korisnika ... 25
2.4.1.6 Novina preporuka ... 25
2.4.1.7 Neočekivanost preporuka ... 25
2.4.1.8 Raznovrsnost preporuka ... 26
2.4.1.9 Rizik usled prihvatanja preporuke ... 26
2.4.1.10 Robusnost sistema ... 26
2.4.1.11 Zaštita privatnosti korisnika ... 26
2.4.1.12 Adaptivnost sistema ... 27
2.4.1.13 Skalabilnost sistema ... 27
2.5 Pružanje preporuka grupi korisnika ... 28
2.6 Rezime ... 29
3. Zaštita privatnosti korisnika ... 30
3.1 Podaci o ličnosti korisnika ... 30

3.2 Zakonski aspekti obrade podataka o ličnosti 32
3.2.1 Načela obrade podataka 34
3.3 Rizici za narušavanje privatnosti 36
3.4 Moguća rešenja za zaštitu privatnosti 39
3.4.1 Sistemska rešenja 39
3.4.2 Algoritamska rešenja 41
3.4.3 Zakonska i regulatorna rešenja 44
3.5 Rezime 45
4. Uticaj kontekstualnih informacija na performanse sistema 46
4.1 Definicija i podela kontekstualnih informacija 46
4.2 Izbor kontekstualnih informacija 46
4.3 Načini za korišćenje kontekstualnih informacija 49
4.4 Rezime 54
5. Primena neuralnih mreža u preporučivačima 56
5.1 Arhitektura neuralne mreže 56
5.2 Algoritmi učenja neuralne mreže 59
5.2.1 RP algoritam 60
5.2.2 SCG algoritam 62
5.2.3 LM algoritam 63
5.2.4 ELM algoritam 64
5.3 Sposobnost generalizacije neuralne mreže 66
5.4 Disbalans klasa 68
5.4.1 Mehanizam delovanja disbalansa klasa 69
5.4.2 Metode za borbu protiv problema disbalansa klasa 69
5.5 Rezime 71
6. Pregled aktuelnog stanja u oblasti 73
6.1 Rezime 79
7. Opis predloženog sistema 80
7.1 Zaštita privatnosti 83
7.2 Mogućnost rada sistema na uređajima s ograničenim resursima 85
7.3. Izbor i način korišćenja kontekstualnih informacija 88
7.4 Problem disbalansa klasa 95
7.5 Izbor mera performansi 105

7.6. Konačni predlog personalizovanog programskog vodiča 108
8. Analiza performansi 110
8.1 Rezime 113
9. Zaključak 115
Literatura 117
Biografija autora 129
Izjava o autorstvu 131

1. Uvod

1.1 Postavka problema

Gledanje televizije svakako je jedna od omiljenih ljudskih aktivnosti. Prelaskom s analognog na digitalno emitovanje programa, bez obzira koji se način pristupa usluzi koristio, količina i raznovrsnost sadržaja koji se nudi korisnicima značajno su povećane.

Za potrebe korisnika koji žele da besplatno koriste ovaj servis, u Republici Srbiji izgrađena je digitalna mreža sa tri multipleksa od kojih svaki pruža mogućnost emitovanja od 14 do 16 radio kanala [1]. Iako broj programa kojima korisnik potencijalno na ovaj način može da pristupi zavisi od željenog kvaliteta slike i zvuka, i teritorije na kojoj se korisnik nalazi, on je svakako višestruko veći od mogućeg broja programa analogne televizije. Razlog za to je povećana efikasnost digitalne televizije koja omogućava emitovanje većeg broja programa preko jednog radio kanala dok za analognu televiziju to nije slučaj.

Druga opcija koju korisnici mogu izabrati je pristup digitalnoj televiziju posredstvom pružalaca usluga distribucije medijskih sadržaja. Iako broj ponuđenih programa varira od izabranog pružaoca usluga i izabranog paketa programa, mogu se doneti uopšteni zaključci. Kod osnovnog paketa korisnici u Republici Srbiji mogu očekivati oko 100 programa, dok za veće pakete ovaj broj ima vrednost od skoro 150 programa [2].

S daljim razvojem tehnologije, trend povećanja količine i raznovrsnosti sadržaja koji se nudi korisnicima postaje još izraženiji, pa tako korisnici HBB TV (Hybrid *Broadcast Broadband Television*) [3] pored pristupa televizijskim, istovremeno mogu pristupati i multimedijalnim sadržajima s Interneta. Količina multimedijalnih sadržaja kojima se može pristupiti preko Interneta je ogromna, što podaci za Youtube, popularni sajt za deljenje i gledanje multimedijalnih sadržaja, jasno pokazuju. Prema podacima ažuriranim 24.01.2017. godine preko 5 milijardi video klipova je dostupno preko ovog sajta, i svakog minuta korisnici u proseku okače još 300 časova multimedijalnih sadržaja [4].

Jasno je da dostupnost sadržaja više ne predstavlja problem, kao što je to bio slučaj u ranijim danima razvoja televizije, ali se korisnici digitalne televizije suočavaju s novim problemom. Pronalaženje željenog sadržaja u ovakvom okruženju nije lak zadatak, jer količina dostupnih sadržaja prevazilazi mogućnost korisnika da ih sam pretraži u razumnome vremenu. Iz ovog

razloga, korisnici ne mogu efikasno da iskoriste prednosti digitalne televizije, pa se u većini slučajeva zadržavaju samo na nekolicini omiljenih kanala - čak i po cenu da propuste zanimljiviji sadržaj koji se u isto vreme nudi na nekom od, za njih, nepoznatih kanala.

Iako su pružaoci usluga distribucije medijskih sadržaja kroz štampane i elektronske programske vodiče pokušali da reše ovaj problem, ovakvim rešenjima problem pretraživanja sadržaja je samo prebačen u drugi oblik, pa sve do pojave personalizovanih programskih vodiča koji uče interesovanja svojih korisnika i potom im preporučuju sadržaje za koje procene da odgovaraju tim interesovanjima, nije bilo značajnijeg napretka u rešavanju ovog problema [5].

Kako problem viška informacija nije specifičan samo za digitalnu televiziju, već je prisutan kod mnogih okruženja (Internet, mobilne aplikacije) i u različitim aspektima ljudskog života (izbor garderobe, putovanja itd), personalizovani programski vodiči spadaju u širu klasu sistema, takozvanih preporučivača [6].

1.2 Predmet i cilj istraživanja

Imajući u vidu veliki broj raznovrsnih oblasti primene preporučivača, te da ih je sve veoma teško obuhvatiti u doktorskoj disertaciji, kao i činjenicu da gledanje televizije igra značajnu ulogu u životu pojedinca odlučili smo da predmet našeg istraživanja ograničimo na personalizovane programske vodiče za digitalnu televiziju. Konkretno, posmatraćemo mogućnost implementacije ovih sistema pomoću neuralnih mreža.

Iako postoji veliki broj algoritama za pružanje preporuka, veštačke neuralne mreže smo izabrali zbog njihove široke primene u rešavanju različitih problema [7], te očekujemo da se pomoću njih mogu modelirati različiti načini razmišljanja korisnika prilikom donošenja odluke o izboru TV sadržaja koji će gledati. Opravdanost ovog izbora potrvđuje i činjenica da se i drugi autori odlučuju za ovakav pristup [8].

Ipak, projektovanje personalizovanog programskog vodiča baziranog na neuralnim mrežama nije lak zadatak, te je neophodno uzeti u obzir različite aspekte kako iz oblasti digitalne televizije i preporučivača, tako iz oblasti mašinskog učenja.

Shodno tome, cilj našeg istraživanja je projektovanje efikasnog vodiča koji

- može da radi u okruženjima s ograničenim hadverskim resursima,

- uzima u obzir problem disbalansa klasa,
- koristi kontekstualne informacije i
- brine o zaštiti privatnosti korisnika.

Kako se od mobilnih uređaja očekuje da postanu dominantan način pristupa TV sadržajima, personalizovani programski vodiči moraju uzeti u obzir njihove ograničene hardverske resurse [9]. Iako se velika većina vodiča u literaturi i dalje bazira na klijent – server arhitekturi, lokalnom implementacijom sistema može se postići da se gledaoci, koji usled zabrinutosti za svoju privatnost ne žele da dozvole sakupljanje njihovih interesovanja u mrežnom centru pružaoca usluge distribucije medijskih sadržaja, odluče za njihovo korišćenje.

Štaviše briga o zaštiti privatnosti, danas, više nije samo pitanje za koje su zainteresovani samo pojedinci. Evropska Unija donela je Uredbu o zaštiti privatnosti – GDPR (*General Data Protection Regulation*) kojom se uređuje ova oblast i promoviše princip projektovanja sistema po kome zaštita privatnosti mora da biti uzeta u obzir već u ovoj fazi [10]. Prilikom projektovanja našeg vodiča primenićemo ovaj princip .

Nasuprot ovih aspekata koji se tiču preporučivača i digitalne televizije, problem disbalansa klasa karakterističan je za sisteme koji koriste tehnike mašinskog učenja – u koje spadaju i neuralne mreže [11]. Javlja se usled nejednakog broja podataka u pojedinačnim klasama, što prouzrukuje nejednak uticaj posmatranih na sam proces učenja neuralne mreže i lošu predikciju klasa predstavljenih s malo podataka. Kod personalizovanih programskih vodiča, ovaj fenomen se javlja usled tendencije korisnika da znatno češće pružaju informacije o TV sadržajima koji im se sviđaju nego o TV sadržajima koji im se ne sviđaju [12]. Ukoliko se ne primene odgovarajuća rešenja za borbu protiv ovog problema, može doći do značajne degradacije performansi.

Na kraju ne treba zaboraviti i na uticaj kontekstualnih informacija [13]. U zavisnosti od toga u kom kontekstu se pristupa TV sadržaju, ponašanje korisnika može značajno varirati. Korisnik će tako neretko birati različite sadržaje u situacijima kada ih gleda sa svojim partnerom i kada ih gleda s prijateljima. Imajući u vidu veliki broj faktora kojima se kontekst može definisati, kao i načina za prikupljanje ovih informacija, neophodno je istražiti koji su odgovarajući za posmatranu primenu.

1.3 Polazne hipoteze

Istraživanje se zasniva na sledećim polaznim hipotezama:

- Gledaoci televizije od ponuđenih TV sadržaja biraju one sadržaje koji odgovaraju njihovim interesovanjima.
- Interesovanja gledalaca su sporo promenljiva.
- Gledaoci imaju izraženu tendenciju da češće pružaju informacije o TV sadržajima koji im se sviđaju nego o TV sadržajima koji im se ne sviđaju.
- Personalizovani programski vodič za digitalnu televiziju se mora implementirati lokalno na korisničkom uređaju kako bi sistem koristili i gledaoci koji usled zabrinutosti za svoju privatnost ne žele da dozvole sakupljanje njihovih interesovanja u mrežnom centru servis provajdera.

1.4 Metode istraživanja

U okviru ovog istraživanja biće korišćene sledeće metode:

- U cilju upoznavanja sa trenutnim stanjem u oblasti personalizovanih programskih vodiča, sa algoritmima učenja neuralnih mreža i sa problemom disbalansa klasa biće izvršena analiza naučno stručne literature i objavljenih radova.
- Predloženi programski vodiči bazirani na neuralnoj mreži treniranoj različitim algoritmima biće realizovani u programskom paketu MATLAB. Računarskim simulacijama vršiće se treniranje i testiranje performansi sistema u cilju određivanja optimalnih parametara sistema (broj skrivenih čvorova mreže, parametar regularizacije, itd). Ostvarene performanse sistema biće analizirane i ispitana uspešnost predikcije sadržaja koje korisnik voli da gleda i sadržaja koje korisnik ne voli da gleda.

1.5 Struktura disertacije

Disertacija je organizovana u devet poglavlja.

U prvom poglavlju je dat uvod doktorske disertacije. U drugom će biti predstavljene osnovne tehnike pružanja preporuka i aktuelni pravci istraživanja kod preporučivača. U trećem i

četvrtom razmotrićemo zaštititu privatnosti korisnika i uticaj kontekstualnih informacija na performanse sistema, respektivno, dok će u petom biti opisana primena neuralnih mreža u preporučivačima. Pregled aktuelnog stanja u oblasti, prikaz postojećih rezultata, njihovo poređenje i diskusija dati su u šestom poglavlju. Na osnovu teorijskih istraživanja u prethodnim poglavljima, u sedmom će se kroz seriju eksperimenata definisati konačan predlog personalizovanog programskog vodiča za digitalnu televiziju, a u osmom analizirati i diskutovati postignute performanse. Deveto poglavlje predstavlja zaključak doktorske disertacije s predlozima mogućih pravaca istraživanja.

Svi termini kojima su u disertaciji označeni korisnici izraženi su u gramatičkom muškom rodu, ali se bez impliciranja rodne pripadnosti, podjednako odnose i na muški i na ženski rod lica.

2. Preporučivači

Pod preporučivačima (*recommender systems*) podrazumevaju se softverski agenti koji pomažu korisnicima da izaberu objekte od interesa u okruženjima u kojima broj raspoloživih objekata nadmašuje sposobnost korisnika da ih pojedinačno pregledaju u razumnome vremenu [14].

Problem pružanja preporuka se može posmatrati kao problem estimiranja ocene objekta koji korisnik još uvek nije video ili koristio, na osnovu podataka o već ocenjenim objektima[6]. U zavisnosti od toga koji se pristup koristi za estimiranje ocene, sistemi za pružanje preporuka, a samim tim i personalizovani programski vodiči, mogu se grubo podeliti na sisteme na bazi filtriranja sadržaja, kolaborativne i hibridne sisteme [15].

2.1 Filtriranje sadržaja

Sistemi na bazi filtriranja sadržaja preporučuju one objekte koji su slični objektima koji su se u prošlosti svideli korisniku. Kod ovog pristupa dostupne objekte najpre treba predstaviti odgovarajućim vektorom odlika, kojim su opisane karakteristike objekata, a zatim se za proračun sličnosti između profila korisnika i objekata koriste heurističke metode ili se odnos između objekata za svakog od korisnika modelira tehnikama mašinskog učenja. U slučaju kada se koriste heurističke metode, profil korisnika se dobija usrednjavanjem vektora odlika objekata koji su se u prošlosti svideli korisniku po pojedinačnim koordinatama. Iako se za proračun sličnosti između objekta i profila korisnika mogu koristiti različite heuristike najčešće se koristi kosinusna sličnost

$$sim(\mathbf{x},\mathbf{y}) = \cos(\mathbf{x},\mathbf{y}) = \frac{\sum_{d=1}^{D} x_d y_d}{\sqrt{\sum_{d=1}^{D} x_d^2}\sqrt{\sum_{d=1}^{D} y_d^2}} \qquad 2.1$$

pomoću koje se računa kosinus ugla između vektora kojim je predstavljen profil korisnika i vektora odlika svakog od dostupnih sadržaja [15]. Sa x_d, odnosno y_d predstavljene su vrednosti d-te koordinate vektora **x**, odnosno vektora **y**, dok je sa D označen broj dimenzija vektora. Što je vrednost kosinusne sličnosti veća to objekat više odgovara korisničkim interesovanjima.

Za realiziciju sistema na bazi filtriranja sadržaja kod kojih se odnos između objekata za svakog od korisnika modelira na raspolaganju je veliki broj tehnika mašinskog učenja.

Reprezentativan primer tehnike koja se može koristiti u ovu svrhu su neuralne mreže [16]. U daljem tekstu doktorske disertacije posebno poglavlje biće posvećeno neuralnim mrežama.

Kao mana sistema na bazi filtriranja sadržaja u literaturi se navodi problem formiranja vektora odlika, preporučivanje previše sličnih objekata, i problem pružanja preporuka novim korisnicima [15].

Kako bi se sistem na bazi filtriranja sadržaja koristio u produkciji, gde se neretko objekti koji su u ponudi dinamički menjaju neophodno je automatski kreirati vektor odlika za svaki od njih. U ranijim fazama razvoja sistema za pružanje preporuka ovo nije bilo moguće uraditi za sve vrste objekata, od kojih se kao primer najčešće navode multimedijalni sadržaji. Ipak, sa napretkom nauke razvijeni su novi metodi za automatsko izdvajanje odlika multimedijalnih sadržaja [17], pa se ovo ne može više smatrati aktivnim problemom.

Nasuprot ovoga, zbog pristupa koji ovi sistemi koriste prilikom pronalaženja objekata koje će preporučiti, oni često mogu preporučiti objekte koji su previše slični objektima koji su se korisniku svideli u prošlosti. U zavisnosti od primene sistema ovo u nekim situacijama može biti poželjno, a u nekim ne. U situacijama kada to nije poželjno, definisanjem pravila na osnovu kojih će sistem od objekata za koje je procenio da će se svideti korisniku formirati listu preporuka, moguće je ublažiti ovaj problem.

Da bi sistem na bazi filtriranja sadžaja mogao da nauči korisnička interesovanja i krene sa pružanjem pouzdanih preporuka, neophodno je da korisnik prvo oceni dovoljan broj objekata. Ovo se u terminologiji sistema za pružanje preporuka naziva problemom novog korisnika ili problemom hladnog starta sistema. Jedan od uobičajenih načina na koji se on rešava je korišćenjem stereotipa, ali po našem mišljenju oni mogu biti nepouzdani jer su neretko kreirani za drugo područje u odnosu na ono za koje se sistem koristi. Naglašavamo i da će korisnici sa specifičnim interesovanjima, biti veoma nezadovoljni ovim sistemom i verovatno već na početku prestati s njegovim korišćenjem.

2.2 Kolaborativno filtriranje

S druge strane, koloborativni sistemi imaju drugačiji pristup prilikom određivanja objekata od kojih očekuju da će se svideti korisnicima. Za razliku od sistema na bazi filtriranja sadržaja, kod ovih sistema se ne koriste podaci o karakteristikama objekata već samo podaci o tome kako su korisnici ocenili dostupne objekte. Kao i kod prethodnih i ovde se mogu koristiti

heurističke metode ili metode mašinskog učenja. U slučajevima kada se koriste heurističke metode koloborativni sistemi mogu preporučivati one objekte [18]

1. koji su se svideli korisnicima sa interesovanjima sličnim kao kod posmatranog korisnika, ili
2. koji su slični objektima koji su se svideli posmatranom korisniku.

Iako drugi način određivanja objekata od interesa izgleda isto kao kod sistema na bazi filtriranja velika razlika je u tome koji se podaci koriste. Kod koloborativnih sistema se koriste podaci ostalih korisnika, dok se kod potonjih sistema koriste samo podaci posmatranog korisnika. Ovaj način određivanja objekta od interesa je uveden kod kolaborativnih sistema kako bi se poboljšala skalabilnost sistema kod kojih je broj objekata za nekoliko veličina manji nego broj korisnika.

Za proračun sličnosti kod kolaborativnih sistema mogu se koristiti različite heuristike, ali je kao i kod sistema na bazi filtriranja sadržaja jedna od najčešće korišćenih kosinusna sličnost definisana izrazom (2.1).

Nakon proračuna sličnosti određuju se najsličniji korisnici ili objekati koji će se koristiti prilikom estimiranja ocene objekata. Ovi korisnici ili objekti nazivaju se susedima, a njihov broj se smatra parametrom sistema kojeg treba podesiti prilikom projektovanja i korišćenja sistema.

U slučaju kada se kao susedi koriste korisnici, estimirana ocena objekta $r_{c,s}$ se može dobiti korišćenjem izraza:

$$\hat{r}_{u,o} = \overline{r}_u + \frac{\sum_{n \in S_u} sim(u,n) \cdot (r_{n,o} - \overline{r}_n)}{\sum_{n \in S_u} |sim(u,n)|}, \qquad 2.2$$

gde je S_u skup najsličnijih korisnika sa posmatranim korisnikom u, $r_{n,o}$ ocena koju je korisnik n iz skupa najsličnijih korisnika dodelio neocenjenom sadržaju, $sim(u,n)$ prethodno definisana mera sličnosti, a $\overline{r}_u$ i $\overline{r}_n$ prosečna ocena korisnika u i n, respektivno [15]. Korišćenjem ovog izraza prilikom estimiranja ocena uzima se u obzir činjenica da različiti korisnici koriste dostupnu skalu ocena na različite načine.

Nasuprot ovome, ukoliko se kao preporučivač koristi koloborativni sistem koji pronalazi najsličnije objekte, estimirana ocena objekta se računa na osnovu izraza

$$r_{c,o} = \frac{\sum_{i \in S_o} sim(c,o) \cdot \overline{r_o}}{\sum_{i \in S_o} |sim(c,o)|}, \qquad 2.3$$

gde je S_o skup objekata koji su najsličniji sa posmatranim objektom c, a $\overline{r_o}$ prosečna ocena objekta o iz skupa S_o dobijena kada se u obzir uzmu korisnici koji su ocenili objekte c i o [18].

Što se tiče sistema koji koriste tehnike mašinskog učenja jedina razlika u odnosu na sisteme na bazi filtriranja sadržaja je u podacima koji se koriste prilikom estimiranja ocena.

Kao mana kolaborativnih sistema u literaturi se navodi problem pružanja preporuka novim korisnicima, problem preporučivanja novih objekata i problem malog procenta ocenjenih objekata (*sparsity*) [15].

Kako bi preporučivač mogao da pronađe slične korisnike i estimira ocenu za neocenjene objekte, novi korisnik mora da oceni dovoljan broj objekata. Ovo se kao i kod sistema na bazi filtriranja sadržaja naziva problemom novog korisnika.

Pored toga što novi korisnici moraju da ocene dovoljan broj objekata, kako bi koloborativni sistem mogao da preporuči nove objekte i oni moraju da budu ocenjeni dovoljan broj puta. Ovo se naziva problemom preporučivanja novih objekata.

Problemi prilikom njihovog korišćenja mogu da se jave čak iako je pojedinačni korisnik ocenio dovoljan broj objekata i ako su novi objekti ocenjeni dovoljan broj puta. Usled malog procenta ocenjenih objekata od strane korisnika, oni koji odgovaraju specifičnim interesovanjima ne mogu biti preporučeni jer nije moguće pronaći dovoljan broj sličnih korisnika koji su koristili ovaj objekat ili objekata koji su slični sa ovim objektom. Dodatno, kolaborativni sistem ne može da pruži preporuke ni korisnicima sa specifičnim interesovanjima jer za njih ne može pronaći slične korisnike, pa ovo može veoma negativno uticati na njihova iskustva s ovom vrstom sistema.

U novije vreme nastala je još jedna klasa sistema koji se prema korišćenoj klasifikaciji preporučivača svrstavaju u kolaborativne sisteme [19]. Ovi sistemi koriste različite tehnike faktorizacije matrica kako bi proračunali latentne faktore, $\mathbf{q}_o$ - koji karakterišu osobine

posmatranog objekata *o*, odnosno latentne faktore, $\mathbf{p}_u$ - koji opisuju koliko je koja osobina objekta važna za posmatranog korisnika *u*. Estimirana ocena $\hat{r}_{u,o}$ objekta *o* za korisnika *u* računa se pravolinijski prema jednačini

$$\hat{r}_{u,o} = \mathbf{q}_o^{\mathrm{T}} \mathbf{p}_u,$$

2.4

množenjem latentnih faktora korisnika i objekata.

Ipak, zbog malog procenta ocenjenih objekata nije moguće direktno primeniti popularne metode faktorizacije matrica, kao što je SVD (*Singular Value Decomposition*) [20], već se problem pronalaženja latentnih faktora posmatra kao optimizacioni problem čiji je cilj da minimizira sledeću grešku

$$\min_{\mathbf{q}^*,\mathbf{p}^*} \sum_{(u,o)\in G} (r_{u,o} - \mathbf{q}_o^{\mathrm{T}} \mathbf{p}_u)^2 + \frac{1}{R}(\|\mathbf{q}_o\|^2 + \|\mathbf{p}_u\|^2), \qquad 2.5$$

gde su $\mathbf{q}^*, \mathbf{p}^*$ optimalne vrednosti faktora, $(u,o) \in G$ skup parova korisnik-objekat za koje postoji ocena, $r_{u,o}$ stvarna vrednost ocene sadržaja *o* za korisnika *u*, R konstanta kojom se kontroliše sposobnost generalizacije sistema i || || euklidska norma vektora.

Korišćenjem sistema na bazi faktorizacije matrica moguće je prevazići problem malog procenta ocenjenih objekata [21].

2.3 Hibridni sistemi

Mane preporučivača na bazi filtriranja sadržaja, odnosno kolaborativnog filtriranja moguće je ublažiti korišćenjem hibridnih sistema koji kombinuju ova dva pristupa. Postoje četiri načina na koje je moguće kombinovati ove sisteme [15]:

1. kombinovanjem rezultata dobijenih na izlazu pojedinačnog sistema na bazi filtriranja sadržaja i kolaborativnog sistema;
2. dodavanjem nekih od karakteristika pristupa sistema na bazi filtriranja sadržaja u pristup koji koriste kolaborativni sistemi;
3. dodavanjem nekih od karakteristika pristupa kolaborativnih sistema u pristup koji koriste sistemi na bazi filtriranja sadržaja;

4. korišćenjem jednog sistema koji ima sve karakteristike sistema na bazi filtriranja sadržaja i kolaborativnog sistema.

Kod hibridnih sistema koji kombinuju rezultate pojedinačnih sistema, najpre sistem na bazi filtriranja sadržaja i kolaborativni sistem estimiraju ocenu objekta, a zatim se konačna ocena formira kao linearna kombinacija pojedinačnih ocena ili izborom pojedinačne ocene sistema za koji se procenilo da u datom trenutku ima bolje performanse. Na ovaj način se brže prevazilaze problemi novog korisnika i novog objekta, i ublažava problem malog procenta ocenjenih sadržaja.

Hibridni sistemi nastali dodavanjem nekih od karakteristika pristupa sistema na bazi filtriranja sadržaja u pristup korišćen kod kolaborativnih sistema, tipično čuvaju profile za svakog od korisnika i koriste ocene sistema na bazi filtriranja sadržaja u slučajevima kada za objekat nemaju dovoljno prikupljenih ocena korisnika [22]. Na ovaj način se ublažava problem malog procenta ocenjenih sadržaja. U preostalim situacijama koristi se kolaborativni deo sistema.

Nasuprot ovome, kod hibridnih sistema nastalih dodavanjem nekih od karakteristika pristupa kolaborativnih sistema u pristup korišćen kod sistema na bazi filtriranja sadržaja, najčešće se neka od tehnika smanjivanja dimenzija vektorskog prostora primenjuje nad grupom korisničkih profila, a zatim u vektorskom prostoru smanjenih dimenzija računa sličnost između pojedinačnih profila [23]. Na ovaj način se kreira vektor odlika koji bolje opisuje dostupne objekte, što ima za posledicu bolje performanse sistema na bazi filtriranja sadržaja [15].

Na kraju, hibridni sistemi koji kombinuju sve karakteristike sistema na bazi filtriranja sadržaja i kolaborativnog sistema se uglavnom baziraju na tehnikama mašinskog učenja, te se za ilustritivan primer može uzeti sistem baziran na neuralnoj mreži koji pored vektora odličja objekata kao ulaze koristi informacije o tome kako su preostali korisnici u sistemu ocenili objekte [24].

2.4 Dominantni pravci istraživanja kod preporučivača

Ipak, u cilju projektovanja efikasnog programskog vodiča za digitalnu televiziju nije dovoljno razmotriti samo pristupe koji se mogu koristiti za pružanje preporuka. Pored toga neophodno je istražiti i dodatne aspekte kao što su načini za bolje opisivanje korisnika i objekata, korišćenje kontekstualnih informacija, način prikupljanja ocena, fleksibilnost sistema i izbor

mera performansi sistema - koji mogu značajno uticati na zadovoljstvo korisnika sistemom [15].

Dominantan pravac istraživanja načina za bolje opisivanje korisnika i objekata bazira se na korišćenju semantike. U literaturi su definisane semantičke ontologije i odgovarajuće mere sličnosti koje ih mogu koristiti tako da neki autori ove sisteme čak smatraju i posebnom vrstom sistema za pružanje preporuka [5].

S druge strane, pored korisničkih interesovanja i karakteristika dostupnih objekata, jasno je da na izbor objekta utiče i kontekst u kome se on bira. Tako na primer izbor televizijskog sadržaja kojeg će korisnik gledati u mnogome može zavisiti od društva s kojim korisnik planira da ga gleda. Zato ne čudi činjenica da je značajan broj istraživanja posvećen definiciji i izboru kontekstualnih informacija, kao i načinima na koje se ove informacije mogu prikupiti i koristiti u sistemu. Kao i kod semantičkih sistema, i preporučivače koji koriste kontekst pojedini autori smatraju posebnom vrstom sistema.

U zavisnosti od oblasti primene različiti načini prikupljanja korisničkih ocena mogu biti pogodni. Prikupljanje ocena, pre svega, mora biti takvo da previše ne ometa uobičajen način korišćenja usluge, ali i da bude dovoljno informativno tako da sistem može da nauči korisnička interesovanja. Iako se eksplicitnim načinom prikupljanja ocena dobijaju podaci koji pouzdano i tačno opisuju korisnička interesovanja, ovaj način prikupljanja može značajno promeniti način uobičajenog korišćenja servisa. Nasuprot ovome, implicitno prikupljanje, bez eksplicitne akcije korisnika, ne ometa način korišćenja usluge ali zato može biti nepouzdano. Na primer, ukoliko se kao implicitna ocena koristi procenat vremena trajanja televizijskog sadržaja koji je korisnik odgledao, to može dovesti do netačnih zaključaka. Sistem na ovaj način može detektovati da je korisnik ujutru odgledao cele vesti, i zaključiti da posmatrani voli da gleda vesti, iako je možda korisnik samo uključio televizor i bio zaokupljen drugom akcijom kao što je spremanje za posao i nije uopšte gledao vesti. Jasno je da način prikupljanja ocena mora predstavljati kompromis između ova dva pristupa. Iako se kod usluga koje se pružaju putem Interneta može dozvoliti češće korišćenje eksplicitnog načina prikupljanja ocena, u slučaju digitalne televizije, gde korisnici nisu navikli na ovaj način interakcije, potrebno je minimizirati broj situacija u kojima korisnik mora eksplicitno da reaguje. Pored toga moguće je ocenjivati i različite aspekte objekata [15], kao što je u slučaju filmova, njihov žanr, režiser, glavni glumci, ali i to može negativno uticati na uobičajen način

gledanja digitalne televizije. Što se tiče preporučivača koji koriste kontekstualne informacije potrebno je razmotriti i način na koji se one prikupljaju.

Korisnici sistema za pružanje preporuka najčešće nemaju mogućnost podešavanja opcija kao što su broj preporuka, koji će se tačno korisnici uzimati u obzir, koje karakteristike objekata će biti presudne prilikom formiranja liste preporuka. Ovaj problem je opisan u [25], i za potrebe njegovog rešavanja razvijen poseban jezik pomoću kojeg bi korisnik mogao da podešava opcije sistema. Ipak, u slučaju personalizovanih programskih vodiča za digitalnu televiziju ovaj pravac istraživanja nije zaživeo [5]. Mogući razlog za to je što pored povećanja zadovoljstva korisnika, ovi sistemi mogu imati i druge ciljeve koji su direktno povezani sa poslovanjem nudioca usluge, kao što je povećanje profita ili reklamiranje pojedinih televizijskih sadržaja, pa pružaoci usluge distribucije medijskog sadržaja ne žele da dozvole korisnicima mogućnost ovakvih podešavanja. Izabrani cilj preporučivača direktno utiče na izbor mera performansi i ova zavisnost biće detaljnije objašnjena u tekstu koji sledi.

2.4.1 Mere performansi sistema

Izbor mera performansi je značajan korak u projektovanju preporučivača jer korišćenje neodgovarajućih može dovesti do toga da se sistem koji za posmatrani scenario nije najbolji proglasi optimalnim. Na njihov izbor utiču ciljevi koje sistem treba da ispuni, karakteristike prikupljenih podataka, način prikaza dobijenih rezultata, kao i očekivanja korisnika i provajdera servisa [26].

Najčešći ciljevi koje provajder servisa želi da postigne su poboljšanje zadovoljstva korisnika i povećanje profita. Ipak, iako su oba cilja međusobno povezana, u zavisnosti od njihovih prioriteta lista preporuka biće formirana na različite načine. Ukoliko je glavni cilj da se brzo uveća profit onda će se u listi preporuka pronaći oni objekti za koje je procenjeno da će se svideti korisniku, a koji istovremeno donose najveću zaradu. Nasuprot tome, ukoliko je glavni cilj da se zadrže korisnici tako što će se povećati njihovo zadovoljstvo servisom, lista preporuka biće formirana od objekata za koje je procenjeno da će im se najviše svideti. Kako su različiti sistemi optimalni za različite scenarije, izabrane mere performansi moraju biti u skladu s ciljevima provajdera servisa.

U zavisnosti od karakteristika prikupljenih podataka o korisničkim interesovanjima i načina prikaza dobijenih rezultata, problem pružanja preporuka se svodi na:

1. problem predikcije ocena objekata,

2. problem klasifikacije objekata, ili
3. problem rangiranja objekata.

Ukoliko korisnik ocenjuje ponuđene objekte ocenama na skali koja nije binarna (sviđa/ne sviđa) i dobijena lista preporuka sadrži informaciju o estimiranoj oceni za svaki od objekata, problem pružanja preporuka se smatra problemom predikcije ocena. S druge strane, ukoliko sistem prikazuje listu preporuka u kojoj su objekti poređani po tačno definisanom redosledu - od onih koji bi se najviše svideli korisniku ka onima koji se manje dopadaju korisniku, onda se ovakav pristup smatra rangiranjem sadržaja. Na kraju, u slučajevima kada korisnik koristi binarnu ili čak unarnu skalu (pruža informacije samo o objektima koji mu se sviđaju) i kada su dobijeni rezultati prikazani u veoma kratkom listom bez definisanog redosleda, problem pružanja preporuka se posmatra kao klasifikacija objekata na one koje bi se svideli korisniku i one koje ne bi.

Kako problem pružanja preporuka može biti definisan na različite načine, za merenje njihove tačnosti u različitim sistemima moraju se koristiti različite mere performansi. Pored toga, usled tendencije korisnika da u pojedinim primenama preporučivača češće pružaju informacije o objektima koji im se sviđaju nego o onima koji im se ne sviđaju, sve posmatrane klase ili ocene neće biti predstavljene podjednakim brojem interakcija - pa ukoliko se pri izboru mera performansi ova činjenica ne uzme u obzir sistem s veoma lošom predikcijom pojedinih klasa ili ocena može proglasiti optimalnim [27]. Ovakvo ponašanje je tipično za gledaoce digitalne televizije, te smo ovom aspektu izbora mera posvetili pažnje u daljem istraživanju.

Na kraju pored ciljeva za koje je preporučivač projektovan i tačnosti pružanja preporuka, treba spomenuti i sledeće osobine koje korisnici i provajder servisa očekuju od ovih sistema [28]:

- Pokrivenost objekata,
- Pouzdanost preporuka,
- Poverenje korisnika,
- Novina preporuka,
- Neočekivanost preporuka,
- Raznovrsnost preporuka,
- Rizik usled prihvatanja preporuke,
- Robusnost sistema,

- Zaštita privatnosti korisnika,
- Adaptivnost sistema i
- Skalabilnost sistema.

U daljem tekstu doktorske disertacije razmotrićemo i njih.

2.4.1.1 Tačnost pružanja preporuka

Tačnost pružanja preporuka je najčešće razmatrana osobina preporučivača, jer veliki broj istraživanja polazi od pretpostavke da što su preporuke tačnije proračunate, to će korisnik biti zadovoljniji. U zavisnosti od toga da li je problem pružanja preporuka definisan kao problem predikcije ocena, klasifikacije ili rangiranja objekata, pogodno je koristiti različite mere performansi.

2.4.1.1.1 Tačnost pružanja preporuka kod predikcije ocena

Kod sistema koji problem pružanja preporuka posmatraju kao problem predikcije ocena, tačnost pružanja preporuka se procenjuje na osnovu tačnosti estimiranih ocena.

Jedna od najčešće korišćenih mera performansi je koren srednje kvadratne greške (*Root Mean Squared Error - RMSE*) definisan sa:

$$RMSE = \sqrt{\frac{1}{|G|}\sum_{(u,o)\in G}\left(\hat{r}_{u,o} - r_{u,o}\right)^2}, \qquad 2.6$$

gde je $|G|$ kardinalni broj skupa formiranog od svih dostupnih parova korisnik-sadržaj.

Pored nje, često se kao mera performansi koristi i srednja apsolutna greška (*Mean Absolute Error – MAE*) definisana sa

$$MAE = \frac{1}{|G|}\sum_{(u,o)\in G}\left|\hat{r}_{u,o} - r_{u,o}\right|. \qquad 2.7$$

Sistem koji je izabran kao najbolji na osnovu korena srednjekvadratne greške, u opštem slučaju, karakteriše se većim broj malih grešaka, dok će onaj koji je izabran na osnovu srednje apsolutne greške ređe grešiti, ali će te greške biti znatno veće.

Ukoliko se na osnovu prikupljenih podataka ustanovi da se pojedini objekti, odnosno korisnici, češće pojavljuju u skupu, kako bi se izbegao njihov dominantan uticaj na procenu performansi sistema preporučuje se da se *RMSE* i *MAE* metrike najpre proračunaju za svaki objekat - odnosno korisnika, a zatim tako dobijeni rezultati usrednje [29].

Najveća mana ovih metrika je to što uzimaju u obzir samo vrednost grešaka pri predikciji ocena, a ne i uticaj grešaka na korisnikov izbor objekata. Kako kod sistema na bazi predikcije ocena sam korisnik formira prag na osnovu kojeg procenjuje da li će koristiti ponuđeni objekat ili ne, ukoliko se usled greške estimirana ocena ne nalazi u istom delu opsega kao i stvarna, to može dovesti do pogrešne preporuke. Ovo se donekle može prevazići definisanjem razlika između pojedinačnih ocena tako da, ukoliko estimirana ocena prelazi prethodno definisani prag, greška bude veća nego u slučaju kada ostaje sa iste strane praga kao i stvarna ocena.

2.4.1.1.2 Tačnost pružanja preporuka kod klasifikacije objekata

Nasuprot ovome, kod preporučivača čiji je cilj klasifikacija objekata na one se korisniku sviđaju i na one koje mu se ne sviđaju, pod tačnošću pružanja preporuka podrazumeva se sposobnost sistema da formira listu preporuka od objekata koji se korisniku stvarno sviđaju. Kako sistem može tačno predvideti ili pogrešiti prilikom predikcije da li objekat treba ili ne treba da se nađe u listi preporuka, pogodno je dobijene rezultate razmotriti pomoću matrice konfuzije (tabela 2.1):

Tabela 2.1. Matrica konfuzije

Lista preporuka	**Stvarna klasa**	
	sviđa	ne sviđa
da	SS	PS
ne	PN	SN

gde je *SS* broj objekata koji se stvarno sviđaju korisniku, *SN* broj objekata koji se stvarno ne sviđaju korisniku, *PS* broj objekata koji su greškom proglašeni da se sviđaju korisniku i *PN* broj objekata koji su greškom proglašeni da se ne sviđaju korisniku. Od ovih vrednosti moguće je formirati različite mere performansi.

Veoma često se za procenu performansi koriste metrike koje dolaze iz oblasti koja se bavi problemom pretraživanja informacija. Tipični primeru su preciznost, odziv i F-mera.

Pod preciznošću podrazumevamo procenat objekata iz liste preporuka koji pripada klasi objekata koji se strvarno sviđaju korisniku. Može se izračunati na sledeći način:

$$Preciznost = \frac{SS}{SS + PS}. \qquad 2.8$$

Odziv predstavlja procenat objekata koji sviđaju korisniku, a koji je prikazan u listi preporuka. Računa uz pomoć sledećeg izraza:

$$Odziv = \frac{SS}{SS + PN}. \tag{2.9}$$

Preciznost i odziv se uobičajeno posmatraju zajedno, jer se u slučaju kada se posmatra samo jedna od metrika podešavanjem dužine liste preporuka može uticati na predstavu o performansama sistema. Na primer, ukoliko se u listi preporuka prikažu svi dostupni objekti posmatrani sistem će imati maksimalni odziv, ali i lošu preciznost, dok će sistem koji preporučuje mali broj objekata imati odličnu preciznost, ali i loš odziv.

Ipak, u slučajevima kada postoji problem disbalansa klasa, bilo koja metrika kod koje uticaj pojedinačnih klasa na performanse nije jasno definisan može dovesti do izbora sistema koji će biti veoma dobar u klasifikaciji klase s većim brojem odbiraka, ali istovremeno veoma loš u klasifikaciji klase s manjim brojem odbiraka [16]. Kako preciznost ne zavisi samo od procenta pogrešno klasifikovanih objekata koji se korisniku ne sviđaju, već i od procenta tačno klasifikovanih objekata koji se sviđaju korisnika, disbalans klasa ima uticaj na ovu metriku, te je nije pogodno koristiti u ovim slučajevima.

Što se tiče F-mere, ona predstavlja kombinaciju preciznosti i odziva. Formalno je definisana sa:

$$F - mera = \frac{(1+\phi)^2 \cdot Odziv \cdot Preciznost}{\phi^2 \cdot Odziv + Preciznost}, \tag{2.10}$$

gde je ϕ koeficijent pomoću kojeg se kontroliše relativni uticaj preciznosti i odziva na F-meru. Kako preciznost zavisi od raspodele podataka po klasama, samim tim ova raspodela ima uticaja i na F-meru.

Pored mera performansi iz oblasti pretraživača informacija, često se koriste i mere performansi iz oblasti mašinskog učenja. Jedna od najčešće korišćenih je tačnost klasifikacije

$$Tačnost = \frac{SS + SN}{SS + SN + PS + PN}. \tag{2.11}$$

Ona je veoma podložna uticaju disbalansa klasa. Imajući u vidu tendenciju korisnika da češće pružaju informacije o objektima koji im se sviđaju nego o onima koji im se ne sviđaju, ukoliko na primer 99% podataka pripada klasi objekata koji se sviđaju korisniku, sistem koji

klasifikuje sve ponuđene sadržaje u ovu klasu imao bi tačnost klasifikacije od 99% i bio smatran odličnim sistemom. Kao posledica izbora ovog sistema, svi objekti koji se korisniku ne sviđaju našli bi se u listi preporuka što bi dovelo do značajnog nezadovoljstva korisnika sistema.

Iz ovog razloga kao moguća alternativa tačnosti klasifikacije, u ovim slučajevima, predlaže se korišćenje G-mean metrike [30]. Ova mera performansi definisana je kao geometrijska sredina tačnosti klasifikacije klase sadržaja koji se korisniku sviđaju, *Ts,* i tačnosti klasifikacije klase sadržaja koji se korisniku ne sviđaju, *Tn*

$$G\text{-}mean = \sqrt{Ts \cdot Tn} = \sqrt{\frac{SS}{SS + PN} \cdot \frac{SN}{SN + PS}} . \qquad 2.12$$

Visoke vrednosti G-mean metrike dobijaju se samo u slučajevima kada je sistem dobar u predikciji obeju klasa, dok u slučaju kada sistem sve podatke iz jedne klase dodeli drugoj, pogrešnoj klasi, ova metrika ima vrednost nula.

Naglašavamo da na izbor mera performansi, pored disbalansa klasa, utiče i činjenica o tome koja klasa je primarni cilj predikcije. Iako, koliko je nama poznato, primena metrika koje uzimaju u obzir ovu činjenicu nije razmatrana u oblasti preporučivača, one se mogu pronaći u oblastima, kao što je na primer bioinformatika, u kojima je problem disbalansa klasa daleko bolje istražen [31]. U ovim slučajevima predlaže se korišćenje AG-mean (*Adjusted G-mean*) metrike. Formula prilagođena našoj primeni data je sa

$$AG\text{-}mean = \begin{cases} \dfrac{G\text{-}mean + Ts \cdot Ns}{1 + Ns}, & \text{za Ts>0} \\ 0, & \text{za Ts=0} \end{cases}, \qquad 2.13$$

gde je N_s procenat podataka koji pripadaju klasi objekata koji se korisniku sviđaju. Od sistema izabranog korišćenjem ove metrike očekuje se da poboljša predikciju klase objekata koji se korisniku sviđaju u odnosu na performanse koje se postižu sistemom izabranim korišćenjem *G-mean* metrike, ali i da ne dozvoli preveliki uticaj problema disbalansa klasa. Primenu *AG-mean* metrike ispitaćemo detaljnije prilikom implementacije našeg personalizovan programskog vodiča za digitalnu televiziju.

Kako korišćenje samo skalarnih metrika nije dovoljno da bi se stekao uvid u prave performanse sistema preporučuje se korišćenje i grafičkih metoda kao što je ROC (*Receiver Operating Characteristics*) grafik [32].

Svaki klasifikator na ovom grafiku predstavljen je tačkom u ravni, sa *x* i *y* koordinatama koje odgovaraju FP_rate i TP_rate vrednostima, respektivno. One su definisane sledećim izrazima

$$TP\_rate = \frac{TP}{mes\{\mathrm{N}_p\}}, \tag{2.14}$$

$$FP\_rate = \frac{FP}{mes\{\mathrm{N}_n\}}, \tag{2.15}$$

gde su TP broj objekata koji su ispravno dodeljeni klasi sa manjom količinom prikupljenih podataka (tzv. pozitivnoj klasi), a FP broj objekata koji su joj greškom dodeljeni, dok u stvari pripadaju klasi sa većom količinom podataka (tzv. negativnoj klasi). Vrednostima $mes\{\mathrm{N}_p\}$ i $mes\{\mathrm{N}_n\}$ u imeniocu izraza (2.14) i (2.15), predstavljeni su brojevi korisničkih interakcija koji pripadaju pozitivnoj i negativnoj klasi, respektivno. U slučaju personalizovanih programskih vodiča za digitalnu televiziju, koji su predmet istraživanja ove doktorske disertacije, podrazumevaćemo da je klasa sadržaja koje korisnik ne želi da gleda pozitivna, a klasa sadržaja koje žele da gledaju negativna.

Kako su na *x*-osi prikazane performanse sistema za negativnu klasu, a na *y*-osi za pozitivnu, na ROC grafiku moguće je lako proceniti performanse za pojedinačne klase. Mala vrednost za FP_rate odgovara klasifikatoru koji je dobar u klasifikaciji objekata koji pripadaju negativnoj klasi, dok mala vrednost za TP_rate odogovara onome koji je loš u klasifikaciji objekata iz pozitivne klase, i obrnuto. Prema tome, idealni klasifikator bi odgovarao tački na grafiku sa FP_rate vrednošću jednakoj nuli i TP_rate vrednošću jednakoj jedinici, dok će ove vrednosti kod klasifikatora koji dostupnim objektima dodeljuje klase na slučajan način (bacanjem novčića) biti međusobno jednake.

Iako u slučajevima kada je broj podataka u negativnoj klasi značajno veći od broja podataka u pozitivnoj, čak i veće promene u pogrešnoj klasifikaciji objekta koji pripadaju negativnoj klasi mogu proći neprimećeno, pogodno je koristiti ROC grafik ukoliko se performanse različitih sistema porede nad istim podacima [32].

2.4.1.1.3 Tačnost pružanja preporuka kod rangiranja sadržaja

Kod sistema koji rangiraju dostupne objekte po tome koliko bi se svideli korisniku, tačnost pružanja preporuka se dobija na osnovu razlika predložene liste u odnosu na referentnu listu ili procene korisnosti liste.

U slučajevima kada ne postoje pouzdane informacije o idealnom redosledu dostupnih objekata u referentnoj listi, kao na primer kada se lista formira tako da se svi objekti kojima je korisnik pristupao nalaze na vrhu, a oni kojima nije na dnu liste, pogodno je koristiti metriku koja ne smatra da je sistem lošiji ukoliko se raspored objekata, za koje ne znamo tačan poredak, razlikuje od rasporeda u referentnoj listi. Primer takve metrike je NDPM (*Normalized Distance-based Measure*) [33], definisane sa

$$NDPM = \frac{C^{-} + 0.5 \cdot C^{u0}}{C^{u}}, \qquad 2.16$$

gde je C^{-} broj parova objekata iz liste koje je sistem poređao pogrešanim redosledom, C^{u0} broj parova objekata za koje je sistem estimirao da će imati isti rang, a za koje u referentnoj listi postoji tačno definisan redosled i C^{u} broj parova objekata koji u referentnoj listi imaju tačno definisan redosled. Što je vrednost NDPM metrike manja, sistem će predlagati listu preporuka u kojoj je redosled objekata za koje postoji definisan poredak sličniji sa redosledom u referentnoj listi.

Nasuprot ovome, u slučajevima kada u referentnoj listi postoje informacije o idealnom rangiranju za sve dostupne objekte, mera performansi mora uzeti u obzir činjenicu da ukoliko je korisnik dodelio objektima isti rang onda on to očekuje i od sistema. Mere performansi koje se koriste za proračun korelacije između rangiranih lista, kao što je *Kendall* τ [34] posebno su pogodne za primenu u ovim slučajevima. Koeficijent korelacije *Kendall* τ se računa na osnovu izraza:

$$\tau = \frac{C^{+} - C^{-}}{\sqrt{C^{u}} \cdot \sqrt{C^{s}}}, \qquad 2.17$$

gde je C^{+} broj parova sadržaja iz liste koje je sistem poređao pravim redosledom, a C^{s} broj parova objekata za koje je sistem procenio da nemaju isti rang.

Tačnost pružanja preporuka moguće je proceniti i bez formiranja referentne liste - samo na osnovu njene korisnosti, predstavljene ponderisanom sumom korisnosti pojedinačnih

objekata, pri čemu su objektima na vrhu liste dodeljeni veći težinski faktori. Kod ovakvog pristupa polazi se od pretpostavke da što se objekat nalazi niže, to je manja verovatnoća da će ga korisnik videti, pa je lista korisnija ukoliko su objekti što tačnije poređani po opadujućoj vrednosti interesovanja korisnika.

U slučajevima kada je lista preporuka male dužine, predlaže se korišćenje metrika kod kojih korisnost objekata u listi ka nižim pozicijama veoma brzo opada, kao što je R-score metrika [35]. Vrednost ove metrike za pojedinačnog korisnika računa se pomoću sledećeg izraza

$$R_u = \sum_u \sum_j \frac{\max(r_{uo_j} - d, 0)}{2^{\frac{j-1}{\alpha-1}}}, \qquad 2.18$$

gde je r_{uo_j} korisnička ocena objekta o_j koji se nalazi na *j*-toj poziciji u listi, *d* ocena kojom korisnik ocenjuje objekat za koji mu je svejedno da li bi ga koristio ili ne, α parametar kojim se kontroliše koliko brzo korisnost objekata opada ka nižim pozicijama u listi. Pravilnim izborom parametra α se može postići da R-score metrika smatra da objekti koje sistem nije prikazao korisnicima ne doprinose korisnosti liste.

Nakon proračuna R-score metrike za pojedinačne korisnike, R-score metrika na nivou celog sistema računa se pomoću izraza

$$R = 100 \cdot \frac{\sum_u R_u}{\sum_u R_u^*}, \qquad 2.19$$

gde je R_u^* vrednost R-score metrike za pojedinačnog korisnika *u* koja se dobija u slučaju da je sistem poređao objekte u listi preporuka tačno po korisničkim preferencijama.

S druge strane, u slučajevima kada sistem prikazuje korisnicima dužu listu preporuka pogodno je koristiti metrike kod kojih korisnost objekata ka nižim pozicijama u listi sporije opada. Često korišćena metrika kod ovakvih sistema je NDCG (*Normalized Discounted Cumulative Gain*) metrika [36], normalizovana verzija DCG (*Discounted Cumulative Gain*) metrike. DCG metrika se računa na osnovu izraza:

$$DCG = \frac{1}{U} \sum_{u=1}^{U} \sum_{j=1}^{J} \chi_{uo_j} \cdot d_j \qquad 2.20$$

gde je χ_{uo_j} vrednost funkcije korisnosti objekta o_j koji se za korisnika *u* nalazi na *j*-toj poziciji u listi preporuka, d_j smanjenje korisnosti usled toga što se objekat nalazi na *j*-toj poziciji u listi, *N* ukupan broj korisnika u sistemu, a *J* dužina liste preporuka. Uobičajeno se kao funkcija korisnosti uzima korisnička ocena objekta, a smanjenje korisnosti d_j računa pomoću izraza:

$$d_j = \frac{1}{\log(1+j)}. \qquad 2.21$$

NDCG metrika se zatim računa pomoću izraza:

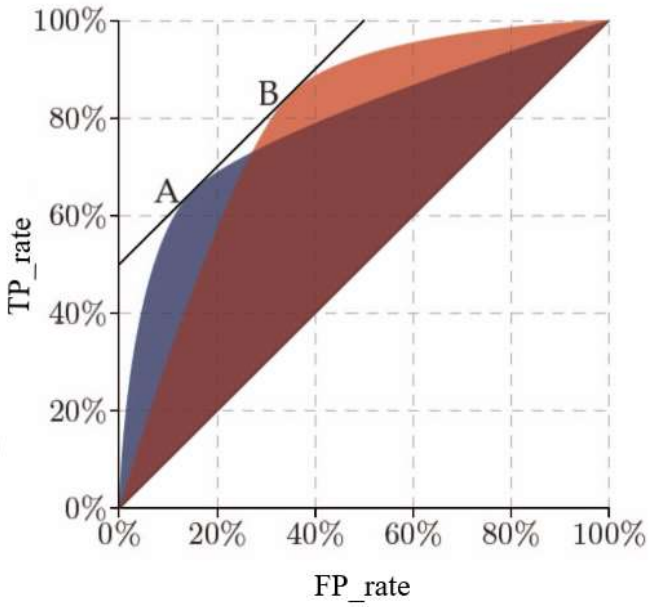
$$NDCG = \frac{DCG}{DCG*} \qquad 2.22$$

gde je $DCG*$ vrednost DCG metrike koja se dobija kada su objekti u listi idealno poređani u skladu s korisničkim interesovanjima.

Kao i kod sistema koji problem pružanja preporuka posmatraju kao klasifikacioni problem, i kod sistema koji rangiraju dostupne objekte pogodno je prazmotriti performanse pomoću grafičkih metoda. Ilustrativni primeri ovih metoda su ROC i Preciznost-Odziv krive [32].

ROC kriva se formira na osnovu ROC grafika, tako što se najpre za svaku moguću dužinu liste preporuka sistem predstavi odgovarajućom tačkom, a zatim sve tačke spoje. Što se dobijena kriva oštrije savija veća je verovatnoća da će se objekti koji pripadaju pozitivnoj klasi naći na vrhu liste, dok što ravnije kriva prilazi liniji koja spaja tačke (0,1) i (1,1) to je veća verovatnoća da će se objekti koji pripadaju negativnoj klasi naći na dnu liste. Kako bi detaljnije objasnili ovaj metod na slici 2.1. prikazali smo performanse dva sistema uz pomoć ROC krive.

Koristeći goreopisanu metodologiju, jasno je da će sistem A objekte koji pripadaju pozitivnoj klasi češće postavljati na vrh liste nego što će sistem B, dok će se u slučajevima kada se posmatraju objekti negativne klase koji su na dnu, njihove uloge obrnuti.



Slika 2.1: Poređenje performansi sistema na ROC krivoj [32].

Nasuprot ovome, na preciznost-odziv krivoj predstavljena je zavisnost preciznosti od odziva posmatranog sistema. Iako i kod nje postoji uticaj disbalansa klasa, preporučuje se da se ona koristi kod primena kod kojih je uspešna predikcija pozitivne klase glavni cilj sistema, kao na primer kod pretraživača podataka [32]. Ipak, u primeni razmatranoj u doktorskoj disertaciji to nije slučaj, jer je osnovni cilj personalizovanih programskih vodiča za digitalnu televiziju uspešna predikcija negativne klase – sadržaja koje bi korisnik voleo da gleda, uz zadovoljavajuću predikciju pozitivne klase – sadržaja koje korisnik ne bi želeo da gleda. Iako sam provajder servisa može postaviti dodatne ciljeve koje želi da postigne, prilikom merenja performansi vodiča mora se voditi računa o osnovnom cilju koji korisnici očekuju.

2.4.1.2 Ciljevi provajdera servisa

Cilj provajdera servisa, kao što smo već napomenuli, može biti direktno povećanje profita ili sprečavanje odlazaka korisnika kod drugih [26], ali i bilo koji proizvoljno definisani cilj koji se može iskazati odgovarajućom funkcijom korisnosti [28]. Ukoliko je u fazi projektovanja sistema ova funkcija jasno definisana, njena vrednost se uzima kao dominantna mera performansi i prilikom procene uspešnosti sistema značajnija je čak i od mera kojima je definisana tačnost pružanja preporuka. Ispunjenost ciljeva koji uzimaju u obzir samo pogodnosti koje provajder servisa dobija je lakše izmeriti, dok je kod ciljeva koji su fokusirani na zadovoljstvo korisnika to znatno teže uraditi jer zadovoljstvo korisnika nije jasno definisano i teško ga je modelirati [37].

2.4.1.3 Pokrivenost objekata

Što se tiče pokrivenosti objekata, ona se može posmatrati na više načina [28]. Uobičajeno se pod ovim terminom podrazumeva procenat dostupnih objekata koje sistem može da preporuči. Ograničenja mogu nastati bilo zbog samog preporučivača (sistemi na bazi filtriranja sadržaja preporučuju samo objekte slične objektima koji su se u prošlosti svideli korisniku), bilo zbog karakteristika prikupljenih podataka. U slučajevima kada se zahteva detaljnija analiza pokrivenosti objekata, pogodno je posmatrati procenat nepopularnih objekata koje sistem može da preporuči, jer se za popularne smatra da je korisnik već upoznat sa njima, te mu nije potreban sistem koji će mu ih preporučivati.

S druge strane, pod pokrivenošću objekata, može se podrazumevati i procenat korisnika kojima sistem može pružiti preporuku. Ukoliko korisnik ne oceni dovoljan broj objekata kod sistema na bazi kolaborativnog filtriranja, ili ukoliko ne pruži dovoljno informacija o tome šta mu se sviđa, a šta ne, preporučivač neće moći da ih preporuči konkretnom korisniku. Tada se, u cilju povećanja procenta korisnika za koje se mogu formirati preporuke, moraju koristiti sistemi koji samo na osnovu informacija o objektima koji se sviđaju korisniku uče njegova interesovanja. Na primeru našeg personalizovanog programskom vodiča za digitalnu televiziju, u daljem tekstu doktorske disertacije objasnićemo detaljnije i ovu vrstu sistema.

Pored načina na koji se pokrivenost objekata definiše, potrebno je uzeti u obzir i trenutak u kome se ona računa. Preporuka je da se prilikom određivanja performansi sistema, pokrivenost objekata računa i za situaciju kada sistem tek uči korisnička interesovanja, takozvano stanje hladnog starta. Kako u većini slučajeva korisnik formira svoje mišljenje o sistemu kada je on u stanju hladnog starta, u svim našim istraživanjima posmatrali smo performanse sistema i pod ovim uslovima.

2.4.1.4 Pouzdanost preporuka

Pod pouzdanošću preporuka podrazumeva se mera performansi kojom se može proceniti poverenje sistema u preporuke koje pruža [38]. Najčešće se kao pouzdanost preporuka koristi interval poverenja koji pored estimirane ocene objekta predstavlja izlaz iz sistema. Interval poverenja se može ili prikazati korisniku, pa da sam korisnik na osnovu njegove vrednosti izabere da li će verovati u tu preporuku, ili sam sistem na osnovu vrednosti ovog intervala može odlučiti da li će prikazati objekat u listi preporuka [39].

2.4.1.5 Poverenje korisnika

Iako je poverenje sistema u svoje preporuke veoma važno, još je važnije poverenje koje korisnici imaju u njih. Jedan od načina za sticanje ovog poverenja je da sistem preporuči korisniku nekoliko objekata za koje je siguran da su se svideli korisniku i da voli da ih koristi [28]. Ova osobina sistema može biti izuzetno značajna u periodu hladnog starta kada korisnik formira mišljenje o sistemu. Ipak, kako poverenje korisnika nije lako izmeriti, najtačnije vrednosti ove mere performansi možemo dobiti samo korišćenjem upitnika [40].

2.4.1.6 Novina preporuka

Pod novinom preporuka podrazumeva se da li je korisnik već znao za postojanje predloženog objekta ili ne [41]. Iako se predlaganje samo novih preporuka u idealnom slučaju može postići tako što se objekti kojima je korisnik pristupao ne prikazuju u listi preporuka, u praksi to nije toliko jednostavno, jer korisnici često ne pružaju informacije o svim objektima kojima su pristupali [28]. Na primer, u slučaju personalizovanog programskog vodiča kod kojeg se informacije o sadržajima koje korisnik voli da gleda prikupljaju implicitno, a o sadržajima koje ne voli da gleda eksplicitno [42], sistem će imati informacije o svim sadržajima koje korisnik voli da gleda, ali ne i o svim sadržajima koje ne voli da gleda, jer sam korisnik odlučuje kada će dostaviti ovu informaciju, a kada ne.

Kod sistema koji u obzir uzimaju i novinu preporuka, pored objekata kojima je korisnik pristupao najčešće se kriju i popularni objekti, ili se mera tačnosti preporuka modifikuje tako da veći težinski faktor bude dodeljen objektu koji nije toliko popularan [43].

2.4.1.7 Neočekivanost preporuka

Što se tiče neočekivanosti preporuka, prilikom određivanja načina na koji će se formirati lista treba posebno voditi računa o njoj jer ova osobina može imati velikog uticaja na to da li će korisnik nastaviti da koristi preporučivač. Baš kao i premali, i preveliki broj neočekivanih preporuka treba izbegavati jer korisnicima vremenom ovakvo ponašanje sistema može postati iritantno.

Neočekivanost preporuka sistema može se povećati, bez negativnog uticaja na zadovoljstvo korisnika, ukoliko se lista preporuka formira od objekata koji najviše odstupaju od dominantnih interesovanja korisnika, a za koje je sistem ipak procenio da bi se korisniku svideli [44].

2.4.1.8 Raznovrsnost preporuka

Osobinom raznovrsnosti preporuka određeno je koliko se objekti iz liste preporuka međusobno razlikuju. Ona se može proceniti na osnovu srednje vrednosti sličnosti između pojedinačnih objekata koji su preporučeni. Zbog svoje prirode, njeno poboljšanje dolazi po cenu smanjenja tačnosti preporuka [44]. Nivo raznovrsnosti preporuka koji korisnik očekuje zavisi od tipa ličnosti korisnika [45]. Podešavanje raznovrsnosti se obično obavlja nakon estimiranja ocena objekata ili dodeljivanja objekata odgovarajućoj klasi - prilikom formiranja liste preporuka. Glavni cilj ovog podešavanja je postići odgovarajući nivo raznovrsnosti preporuka uz minimalnu degradaciju njihove tačnosti.

2.4.1.9 Rizik usled prihvatanja preporuke

Ponekad, prihvatanje preporuka može dovesti korisnike u neprijatnu situaciju ili ih čak i ugroziti [46]. Iako je rizik korišćenja preporučivača u nekim primenama, kao što je kupovina akcija, vidljiviji on postoji i kod personalizovanih programskih vodiča za digitalnu televiziju. Na primer, u slučajevima kada su korisnici sistema deca, u listi preporuka ne smeju se pronaći neprimereni TV sadržaji, koji bi potencijalno mogli da utiču na njihov razvoj. Mogući način da se konkretna situacija izbegne je korišćenje sistema za roditeljsku kontrolu. Imajući u vidu osetljivost ove kategorije korisnika, druga mogućnost je korišćenje sistema posebno razvijenih za njihove potrebe [47].

2.4.1.10 Robusnost sistema

Pod robusnošću podrazumeva se mera u kojoj maliciozni korisnik može uticati na preporuke sistema za posmatranog korisnika [48]. Ova osobina se obično posmatra samo kod sistema na bazi kolaborativnog filtriranja, jer se kod sistema na bazi filtriranja sadržaja preporuka formira samo na osnovu podataka pojedinačnog korisnika, pa podaci ostalih korisnika nemaju uticaj na sam proces pružanja preporuka. Kako postoje različiti scenariji napada na preporučivače, prilikom opisivanja njihove robusnosti, treba navesti i podatke o kom se konkretnom protokolu napada radi.

2.4.1.11 Zaštita privatnosti korisnika

Zaštita privatnosti korisnika je zbog značajnog pooštravanja regulative iz ove oblasti postala veoma značajna osobina sistema, te ćemo njoj posvetiti posebno poglavlje u doktorskoj disertaciji. U zavisnosti od pristupa koji se koristi prilikom formiranja preporuka postoje

različiti rizici za narušavanje privatnosti, bilo direktnim pristupom podacima ili donošenjem zaključaka na osnovu izlaza iz sistema.

2.4.1.12 Adaptivnost sistema

Adaptivnost sistema opisuje sposobnost sistema da se prilagodi bilo promenama u globalnim trendovima [28] ili promenama u korisničkim interesovanjima [49].

Promene u globalnim trendovima mogu dovesti do toga da sadržaj koji je nekada bio interesantan korisniku, ali je zastareo, usled nekog novog događaja ponovo postane interesantan. Na primer, ukoliko je korisnik ljubitelj filmova u kojima su akteri *Marvel*-ovi junaci, izlazak novog filma s ovom tematikom može dovesti do toga da korisnik poželi da ponovo odgleda neke od starijih. Brzina adaptivnosti sistema na ovakve promene se može povećati nauštrb tačnosti preporuka, tako što sistem može preporučivati objekte ovog tipa čak iako estimirana ocena nije dovoljno visoka [28].

S druge strane, jedna od pretpostavki našeg istraživanja je da su interesovanja korisnika sporo promenljiva. Stoga, da bi korisnik nastavio da koristi sistem za pružanje preporuka veoma je važno da isti nauči promene korisničkih interesovanja i prilagodi listu preporuka [49]. Sposobnost sistema da se prilagodi ovim promenama može se izmeriti razlikom u vrednosti *Shannon*-ove entropije:

$$H_o = -\sum_{o=1}^{O} p(o)\log\left(p(o)\right) \quad ,$$

2.23

pre i posle promene interesovanja, gde je $p(o)$ verovatnoća da se sadržaj *o* pojavi u listi preporuka, a *O* ukupan broj dostupnih sadržaja [50].

2.4.1.13 Skalabilnost sistema

Iako skalabilnosti nije pridavano toliko pažnje u ranijim godinama razvoja preporučivača [15], s porastom količine podataka koju sistemi prikupljaju ova osobina dobija na važnosti [30]. Konkretno, pod skalabilnošću sistema smatramo sposobnost sistema da nauči korisnička interesovanja i pruži pravovremene preporuke u uslovima kada količina prikupljenih podataka dostigne realne vrednosti s kojima se sistem sreće u produkcionom okruženju [28]. Već u fazi projektovanja sistema neophodno je posvetiti pažnju izboru algoritma učenja korisničkih interesovanja koji ima malu računsku složenost, ali kojim se može postići i dobra tačnost

preporuka. Kako bi se dokazala skalabilnost sistema najčešći način je da se za različitu količinu prikupljenih podataka prikažu utrošeni hardverski resursi [51]. Pored njih poželjno je prikazati i mere na osnovu kojih se može zaključiti o brzini učenja sistema, kao što je na primer, vreme treniranja neuralne mreže.

2.5 Pružanje preporuka grupi korisnika

Nasuprot dosadašnjem uopštenom razmatranju o preporučivačima, ne treba zaboraviti činjenicu da je gledanje televizije društvena aktivnost pa kod personalizovanih programskih vodiča za digitalnu televiziju treba razmotriti i mogućnost pružanja preporuka grupi korisnika [5].

Pružanje preporuka grupi korisnika je znatno složenije i predstavlja veliki izazov za preporučivače [52]. U literaturi koja se bavi ovom tematikom, obično se kreće od pretpostavke da su nam poznata interesovanja svakog od korisnika iz grupe, a zatim se primenjuju metode za agregaciju korisničkih interesovanja ili profila. Korisnici prilikom izbora sadržaja, u većini slučajeva teže da nijedan od članova grupe ne bude nezadovoljan, pa metode agregacije treba da simuliraju ovaj proces. U opštem slučaju, najbolje performanse se postižu korišćenjem agregacione metode koja grupnu ocenu formira množenjem pojedinačnih ocena svakog od korisnika, ali izbor najbolje agregacione metode za konkretni slučaj zavisi i od homogenosti grupe i međusobnog odnosa članova grupe [52].

Zadovoljstvo korisnika preporučenim sadržajem nije lako proceniti jer je promenjivo i na njega utiču emocije preostalih članova grupe i nagon pojedinca da bude društveno prihvaćen. Uticaj emocija zavisi od međusobnog odnosa korisnika, pa će tako osećaj zadovoljstva ili nezadovoljstva najboljeg prijatelja pre proizvesti ista kod korisnika nego u slučaju kada ona dolaze od poznanika iz grupe. S druge strane, zbog nagona da bude društveno prihvaćen korisnik će se neretko složiti s mišljenjem ostalih članova grupe, pa da bi se ovaj uticaj izbegao umesto direktnog merenja zadovoljstva korisnika koriste funkcije koje ga estimiraju.

Novija istraživanja pokušavaju da problem pružanja preporuka grupi korisnika prevaziđu korišćenjem agregacionih metoda u koje je ugrađena funkcija zadovoljstva korisnika i preporučivanjem sekvenci sadržaja tako da svaki od pojedinačnih korisnika na kraju bude zadovoljan [52]. Ukoliko se koristi pristup preporučivanja sekvenci sadržaja, posebnu pažnju treba posvetiti izboru redosleda sadržaja zbog dinamičkog članstva korisnika u grupi.

2.6 Rezime

U ovom poglavlju smo:

- predstavili preporučivače – širu klasu sistema u koje spadaju i personalizovani programski vodiči,
- diskutovali prednosti i mane različitih pristupa koji se koriste za formiranje preporuka,
- razmotrili aktuelne pravce istraživanja u ovoj oblasti,
- istražili koje su mere performansi pogodne za konkretnu primenu i
- prikazali mogućnosti koje postoje kod pružanja preporuka grupi korisnika.

U sledećem poglavlju ćemo detaljno istražiti rizike za narušavanje privatnosti koji postoje kod personalizovanih programskih vodiča i moguće načine za sprečavanje kompromitovanja podataka o ličnosti korisnika.

3. Zaštita privatnosti korisnika

Zaštita privatnosti korisnika je bitan preduslov svakog modernog telekomunikacionog servisa. Kod personalizovanih servisa, kao što je personalizovani programski vodič, privatnost korisnika može biti ugrožena usled neophodnog prikupljanja i čuvanja korisničkih podataka za potrebe personalizacije. Mere korišćene za zaštitu privatnosti i podaci koji se prikupljaju moraju da predstavljaju kompromis između želja korisnika da dobiju servis koji je što bolje prilagođen njihovim potrebama i težnje korisnika da ostave što manje svojih ličnih podataka [53].

3.1 Podaci o ličnosti korisnika

Podatak o ličnosti je svaka informacija koja se odnosi na fizičko lice koje se u nekom trenutku može identifikovati [54]. Kvalitet same informacije, odnosno da li je ona istinita ili lažna, kao i oblik u kome se informacija čuva (papirni ili digitalni, pa čak i enkriptovani), nisu od presudnog značaja za utvrđivanje da li informacija predstavlja podatak o ličnosti. Bitno je samo da informacija ima odgovarajuće značenje i smisao, poput informacije o žanru filmova koji korisnik voli da gleda.

Da bi predstavljala podatak o ličnosti, informacija se mora dovesti u relaciju sa fizičkim licem, bilo direktnom vezom ili indirektnom vezom, korišćenjem drugih informacija i drugih lica. Kvalitet posmatrane veze se mora bazirati na jednom od sledeća 3 elementa:

- sadržaju – informaciji koja opisuje korisnika (na primer, korisnik voli da gleda ljubavne filmove),
- svrsi – informaciji koja omogućava procenu i odgovarajući tretman lica (na primer korisnik voli da gleda crtane filmove, pa ga kolege na poslu mogu smatrati neozbiljnim),
- efektu – informaciji koja može imati uticaja na samo lice, njegov interes, slobodu i prava (na primer, korišćenje informacije o lokaciji korisnika kako bi se personalizovali ponuđeni servisi).

Poslednja karakteristika koju podatak o ličnosti mora da poseduje je element na osnovu kojeg se fizičko lice može identifikovati. Iako je u prošlosti naglasak bio na korišćenju JMBG broja i ličnog imena korisnika, u današnje vreme se kombinovanjem informacija koje ga direktno

ne identifikuju (pol, godine starosti, profesija) korisnik može identifikovati na indirektan način, korišćenjem međusobne povezanosti ovih informacija.

Jasno je da se mnoge informacije, na osnovu prethodne definicije podataka o ličnosti, mogu svrstati u ovu kategoriju. Međutim, na osnovu Zakona o zaštiti podataka o ličnosti (ZZPL) Republike Srbije [55] svi podaci o ličnosti se ne tretiraju isto. U ZZPL su s jedne strane jasno definisani podaci koji ne uživaju pravnu zaštitu i s druge strane naročito osetljivi podaci o kojima se mora posebno voditi računa.

Shodno članu 5 ZZPL Republike Srbije pravna zaštita korisnika u kontekstu ovog zakona se ne primenjuje nad:

- podacima koji su dostupni svakome i objavljeni u javnim glasilima i publikacijama, ili im se može pristupiti u arhivama, muzejima i drugim sličnim organizacijama,
- podacima koji se obrađuju za porodične i druge lične potrebe i nisu dostupni trećim licima,
- podacima koji se o članovima političkih stranaka, udruženja, sindikata, kao i drugih oblika udruživanja obrađuju od strane tih organizacija, pod uslovom da član da pismenu izjavu da određene odredbe ovog zakona ne važe za obradu podataka o njemu za određeno vreme, ali ne duže od vremena trajanja njegovog članstva,
- podacima koje je lice, sposobno da se samo stara o svojim interesima, objavilo o sebi.

Korisnik može potražiti pravnu zaštitu na osnovu Ustavom garantovanog prava na privatnost ukoliko dođe do dalje distribucije njegovih podataka koji su mimo njegove volje postali javno dostupni, kako bi sprečio njihovo širenje.

Shodno članu 16 ZZPL Republike Srbije definisani su posebno osetljivi podaci koji se moraju posebno označiti i nad kojima se moraju primeniti posebne mere zaštite. Razlikuju se dve grupe podataka:

- podaci koji se odnose na nacionalnu pripadnost, rasu, pol, jezik, veroispovest, pripadnost političkoj stranci, sindikalno članstvo, zdravstveno stanje, primanje socijalne pomoći, žrtvu nasilja, osudu za krivično delo i seksualni život koji se mogu obrađivati na osnovu slobodno datog pristanka lica, osim kada zakonom nije dozvoljena obrada ni uz pristanak;

- podaci koji se odnose na pripadnost političkoj stranci, zdravstveno stanje i primanje socijalne pomoći, koji se mogu obrađivati bez pristanka lica, samo ako je to zakonom propisano.

Iako je pol označen kao posebno osetljiv podatak, on se u većini slučajeva lako može saznati iz JMBG broja i ličnog imena osobe koji nisu svrstani u ovu kategoriju, pa se nameće pitanje opravdanosti ovakve klasifikacije [54].

Mere zaštite posebno osetljivih podataka shodno članu 16 ZZPL Republike Srbije uređuje Vlada. Do trenutka pisanja doktorske disertacije ovakva uredba nije doneta, pa je rukovaocima posebno osetljivih podataka ostavljena sloboda da primene mere za koje smatraju da su dovoljne.

3.2 Zakonski aspekti obrade podataka o ličnosti

Pod obradom podataka o ličnosti se može smatrati bilo koja radnja koja se primenjuje nad ovim podacima. Najčešće radnje su definisane u ZZPL Republike Srbije, od kojih je navedeno samo nekoliko reprezentativnih:

- prikupljanje, beleženje i čuvanje,
- obezbeđivanje i prikrivanje,
- organizovanje, objedinjavanje, upodobljavanje,
- stavljanje na uvid i objavljivanje,
- razvrstavanje i pretraživanje,
- kopiranje i umnožavanje,
- izmeštanje i činjenje nedostupnim na drugi način,
- iznošenje iz zemlje.

U zavisnosti od toga kakva ovlašćenja ima onaj koji obrađuje podatke, razlikujemo:

- rukovaoca podacima,
- obrađivača podataka i
- korisnika podataka.

Rukovalac je primarno odgovoran za obradu podataka i on određuje svrhu i način obrade, ukoliko to nije prethodno definisano nekim zakonom. Pod svrhom obrade podrazumeva se razlog zbog kojeg se podaci prikupljaju, dok način obrade definiše koji će se podaci

prikupljati, koliko dugo će trajati obrada podataka, ko će imati pristup podacima i ostale detalje povezane sa obradom.

Rukovalac može, na osnovu zakona ili ugovora, da poveri određene poslove u vezi sa obradom drugom pravnom entitetu koji se u terminologiji ZZPL naziva obrađivačem podataka. Obrađivač podataka, kao i rukovalac, vrši obradu podataka, ali ne definiše njihovu svrhu niti način obrade.

Korisnik podataka predstavlja pravni entitet koji na osnovu zakona ili pristanka fizičkog lica ima pravo da koristi podatke, ali ne određuje svrhu i način obrade podataka kao rukovalac, niti je ovlašćen od strane rukovaoca kao obrađivač podataka.

Definisanje različitih uloga pri obradi podataka je neophodno zbog određivanja odgovornosti prilikom složenih procesa obrade podataka. Na primer kod sistema koji koriste algoritme mašinskog učenja različiti entiteti mogu biti odgovorni za različite delove procesa [56]:

- prikupljanje korisničkih podataka,
- primenu algoritama mašinskog učenja i
- donošenje odluka na osnovu rezultata mašinskog učenja.

Prema ZZPL Republike Srbije rukovalac podacima je prvenstveno odgovoran za primenu i izvršavanje svih obaveza iz ovog Zakona. Rukovalac podacima ima sledeće obaveze [57]:

- da vrši obradu podataka u skladu sa Zakonom i načelima obrade podataka,
- da primenjuje organizacione i tehničke mere za zaštitu podataka,
- odgovornost za izbor obrađivača,
- da obavesti fizička lica o obradi pre njenog početka,
- da postupi u skladu sa zahtevima za ostvarivanje prava korisnika.

Dok su obaveze obrađivača podataka da [54]:

- vrši obradu podataka u skladu sa Zakonom i načelima obrade podataka,
- primenjuje organizacione i tehničke mere za zaštitu podataka i
- postupa u svemu u skladu s nalozima koje je zadao rukovalac.

Obaveza korisnika podataka je da vrši obradu podataka u skladu sa Zakonom i načelima obrade podataka [54].

Kako bi uopšte bila dozvoljena obrada podataka, tokom čitavog trajanja obrade mora da postoji zakonsko ovlašćenje na osnovu kojeg se ona vrši ili pristanak lica čiji se podaci o

ličnosti obrađuju. Postoje i slučajevi kada je moguće obaviti obradu bez ispunjenja ovih uslova, shodno čl. 12 i 13 ZZPL, u cilju zaštite životno važnih interesa lica, kao što su život, zdravlje i fizički integritet, ili ukoliko je obrada u cilju zaštite nacionalnih interesa, ali je predloženo rigorozno tumačenje ovih članova [54]. U slučaju pružalaca usluga distribucije medijskih sadržaja koji podatke prikupljaju za potrebe personalizovanog programskog vodiča za digitalnu televiziju, neophodno je da postoji pristanak lica. Ukoliko se planira iznošenje podataka iz Republike Srbije, pored pristanka korisnika, neophodno je obezbediti dozvolu od Poverenika za informacije od javnog značaja i zaštitu podataka o ličnosti. U skladu sa članom 53 ZZPL, dozvola može biti dobijena samo ukoliko je država članica Konvencije Saveta Evrope o zaštiti lica u odnosu na automatsku obradu ličnih podataka ili ukoliko je propisom ili ugovorom o prenosu podataka obezbeđen stepen zaštite u skladu sa Konvencijom.

Pre nego što fizičko lice da pristanak za obradu svojih ličnih podataka, mora biti upoznato sa svrhom i načinom obrade, ali i svim pravima korisnika u skladu sa ZZPL, kao što su pravo korisnika da opozove svoj pristanak i pravo da zatraži uvid u prikupljene podatke.

Pristanak se mora dati u pisanoj formi, koja prema članu 3 ZZPL podrazumeva i elektronski oblik pod uslovima iz zakona kojim je uređen elektronski potpis. Iako je danas dominantan način davanja pristanka na Internetu i kod mobilnih aplikacija konkludentnom radnjom (na primer klikom na “Pristajem” dugme u aplikaciji), aktuelni ZZPL Republike Srbije [55] ne dozvoljava ovu mogućnost.

3.2.1 Načela obrade podataka

Pored uslova koje pružalac usluga distribucije medijskih sadržaja mora da ispuni pre početka obrade podataka, postoje i načela koja se moraju poštovati tokom obrade korisničkih podataka:

- načelo zakonitosti i pravičnosti,
- načelo ograničenosti svrhe,
- načelo srazmernosti,
- načelo transparentnosti obrade,
- načelo tačnosti (kvaliteta informacija),
- načelo ograničenog zadržavanja i
- zabrana diskriminacije.

Načelo zakonitosti i pravičnosti proizilazi iz Ustava Republike Srbije, Konvencije Saveta Evrope o zaštiti lica u odnosu na automatsku obradu podataka i pojedinih odredbi ZZPL. Zakonitost obrade je ispunjena ukoliko su prilikom procedure prikupljanja podataka primenjene odredbe ZZPL, odnosno ukoliko postoji ovlašćenje ili pristanak korisnika za obradu. Načelo pravičnosti, s druge strane, predstavlja širi pojam koji podrazumeva da se tokom obrade mora voditi računa o interesima korisnika i da obrada nikako ne sme biti na njegovu štetu.

Članom 8, tačkama 2 i 3 ZZPL definisano je načelo ograničenosti svrhe. Obrada podataka nije dozvoljena ukoliko svrha obrade nije konkretna i unapred određena. Izuzetak od ovog pravila je obrada podataka u istorijske, statističke ili naučnoistraživačke svrhe, shodno članu 6 ZZPL, gde je moguće iskoristiti podatke koji su prikupljeni u druge svrhe.

Načelo srazmernosti dozvoljava obradu samo onih podataka koji su neophodni i relevantni za određenu svrhu. Shodno tačkama 6 i 7 člana 8 ZZPL, obrada se smatra nedozvoljenom ukoliko je podatak koji se obrađuje nepotreban ili nepodesan za ostvarenje svrhe obrade, ili su broj ili vrsta podataka koji se obrađuju nesrazmerni svrsi obrade.

Da bi obrada bila dozvoljena, neophodno je da ispuni i načelo transparentnosti. Potpuna transparentnost obrade se ostvaruje tako što se korisniku čiji se podaci obrađuju garantuje pravo obaveštenja o obradi, pravo na uvid u podatke i pravo na kopiju prikupljenih podataka. Način na koji korisnik može zahtevati ostvarivanje ovih prava opisan je u ZZPL.

Načelo tačnosti informacija garantuje da su podaci koji se koriste tokom obrade istiniti, potpuni i ažurni u kontekstu obrade. Kako rezultat obrade netačnih, zastarelih i nepotpunih podataka može direktno da ošteti interese korisnika, načelo transparentnosti obrade korisnicima daje pravo da kontrolišu da li se prilikom obrade podataka primenjuje načelo tačnosti. Koliko često će se podaci ažurirati zavisi od konteksta obrade, odnosno koliko često se ti podaci procesiraju u cilju donošenja odluka. Ukoliko korisnik utvrdi da je narušeno načelo tačnosti informacija, on može zahtevati ispravku, ažuriranje i dopunu podataka, ili čak prekid njihove obrade.

Načelo ograničenog zadržavanja predviđa definisanje roka u kojem se podaci o ličnosti obrađuju. Nakon isteka ovog roka pravni entitet koji obrađuje podatke mora da ih izbriše, ili trajno anonimizuje. Na osnovu prava o obaveštenju o obradi, fizičko lice treba da dobije i informaciju u kom vremenskom periodu se podaci obrađuju.

Kako je zaštita podataka o ličnosti univerzalno ljudsko pravo, svi korisnici moraju biti tretirani podjednako, bez obzira na lična svojstva koja ih razlikuju. Diskriminacija fizičkih lica koja se razlikuju bilo po državljanstvu, rasi, polu, seksualnom opredeljenju, imovinskom stanju ili nekoj drugoj karakteristici nije dozvoljena prilikom obrade.

Treba napomenuti da ZZPL pored gorenavedenih načela obrade predviđa da rukovalac i obrađivač moraju da preduzmu tehničke, kadrovske i organizacione mere kako bi sprečili zloupotrebu, uništenje, gubitak, neovlašćene promene ili pristup podacima. U daljem tekstu detaljnije će biti razmatrani rizici za narušavanje privatnosti i moguća rešenja.

3.3 Rizici za narušavanje privatnosti

Rizici za narušavanje privatnosti kod korišćenja sistema za pružanje preporuka, i načini za njihovo kontrolisanje, detaljno su razmatrani u radu [57]. Do narušavanja privatnosti dolazi ukoliko se neko od načela obrade ne poštuje, te prema tome razlikujemo narušavanje privatnosti nastalo usled direktnog pristupa podacima (narušavanje načela zakonitosti i pravičnosti, i načela srazmernosti), i usled sofisticirane obrade podataka koja može ugroziti prava korisnika (narušavanje načela ograničenosti svrhe). Takođe, u zavisnosti od koga nam preti narušavanje privatnosti, razlikujemo rizike koji dolaze od samog sistema za pružanje preporuka, od ostalih korisnika sistema, i od drugih eksternih entiteta.

Narušavanje privatnosti usled direktnog pristupa korisničkim podacima u sistemima za pružanje preporuka može nastati kao rezultat:

- prikupljanja neželjenih podataka,
- razmene podataka s trećim licima,
- neovlašćenog pristupa podacima od strane zaposlenih.

Korisnički uređaji koji se koriste u današnje vreme pružaju mogućnost prikupljanja velikog broja raznovrsnih podataka o korisniku. Imajući ovo u vidu, pružaoci usluge distribucije medijskog sadržaja mogu doći u iskušenje da prikupljaju i podatke koji nisu neophodni za posmatranu obradu, čime direktno krše načelo srazmernosti obrade. Ipak, korisnici nisu podjednako zabrinuti za sve podatke koji se mogu prikupljati. Istraživanje je pokazalo da su korisnici posebno osetljivi na prikupljanje podataka koji definišu kontekst u kome se nalaze, jer se korišćenjem ovih podataka može doći do zaključaka koji ih direktno kompromituju [58].

Pružalac usluge distribucije medijskog sadržaja može obavljati razmenu korisničkih podataka s trećim licima, bilo u svrhu saradnje sa akademskom zajednicom, prebacivanja dela procesa pružanja preporuka na nekog drugog, ili prodaje podataka.

S razvojem *cloud computing* tehnologija, veliki broj kompanija je počeo da nudi uslugu pružanja preporuka kao servis (*recommendation as service*), pa je razmena podataka neophodna ukoliko provajderi servisa žele da koriste ove usluge. Pored toga, imajući u vidu ekonomsku isplativost prodaje korisničkih podataka takozvanim „brokerima podacima“, nije neočekivano da provajderi servisa pokušaju da iskoriste priliku za dodatnu zaradu [59]. Iako se prilikom ovih razmena primenjuje proces anonimizacije podataka, on nije savršen i zato može doći do zloupotrebe originalnih podataka [60].

Prikupljeni podaci o korisnicima mogu biti interesantni i zaposlenima kod provajdera servisa. Najčešće mete ovakvih napada su slavne ličnosti čije podatke zaposleni mogu ukrasti, bilo zbog radoznalosti, ili materijalne koristi [60]. Tehničke i organizacione mere zaštite podataka moraju biti striktno primenjene kako bi se ova situacija izbegla.

S druge strane, narušavanje privatnosti usled sofisticirane obrade podataka u sistemima za pružanje preporuka može nastati kao rezultat:

- otkrivanja osetljivih informacija tokom obrade,
- ciljanog oglašavanja i
- diskriminacije korisnika.

Posebno osetljivi podaci o korisniku, kao što su pol i nacionalna pripadnost, mogu se odrediti primenom metoda mašinskog učenja nad podacima o korisničkim interakcijama sa sistemom [61]. Iako ovi podaci mogu ubrzati učenje korisničkih interesovanja i ublažiti problem hladnog starta, korisnici se vrlo retko odlučuju da ih podele, bilo zbog zabrinutosti za svoju privatnost, ili zbog toga što se zahteva eksplicitno unošenje ovih podataka. U istraživanju [61] je pokazano da se korišćenjem logističke regresije, pol korisnika može tačno odrediti u 80% slučajeva korišćenjem samo ocena filmova koje su korisnici gledali. Štaviše, informacija samo o tome da li je korisnik gledao određeni film, dovoljna je da se sa velikom preciznošću odredi pol korisnika. Po ZZPL Republike Srbije, informacija o polu spada u posebno osetljive podatke, i određivanje ove informacije može značajno ugroziti prava korisnika ako se istorijskom obradom podataka odredi da je korisnik u nekom trenutku promenio pol [54].

Sistemi za pružanje preporuka se mogu koristiti i za ciljano oglašavanje, gde se na osnovu prikupljenih podataka određuje za koju bi reklamu korisnik bio zainteresovan. Ovakvo oglašavanje može otkriti veoma osetljive informacije, pa se u literaturi navodi primer kada su roditelji saznali da im je maloletna ćerka trudna tako što su dobili kupone za kolevku i odeću za bebe [62].

Obrada podataka koji se prikupljaju u svrhu pružanja preporuka može dovesti i do diskriminacije između korisnika. Kao tipičan primer navode se sistemi koji pružaju preporuke na Internetu, i koji mogu iskoristiti podatke o korisniku kako bi prodali skuplje proizvode onima koji vole više da troše, ne obraćajući pritom pažnju na moguće uštede [63].

Do narušavanja privatnosti posmatranog korisnika može doći i usled aktivnosti ostalih korisnika u sistemu. Privatnost može biti ugrožena od strane korisnika koji su bliski s posmatranim, ili od strane potpuno nepoznatih i malicioznih osoba. U prvom slučaju, osobe s kojima delimo uređaj ili nalog mogu na osnovu liste preporuka da donesu određene zaključke o nama, i ovo se smatra blažim narušavanjem privatnosti, dok je drugi slučaj znatno ozbiljniji jer napadač pokušava da utiče na sam proces pružanja preporuka kako bi došao do naših podataka [60]. Na proces pružanja preporuka ostali korisnici mogu uticati samo kod sistema na bazi kolaborativnog filtriranja i hibridnih sistema jer se u samom procesu pored podataka posmatranog korisnika, koriste i podaci ostalih korisnika. Do zaključka o korisničkim interesovanjima može se doći bilo aktivnim ili pasivnim napadom, kao što je pokazano u [64]. Aktivni napad je demonstriran za kolaborativne sisteme koji formiraju listu preporuka na osnovu sadržaja koji su se svideli korisnicima sa sličnim interesovanjima. Kod ove vrste napada, najpre se kreiraju lažni profili korisnika koji su napravljeni tako da odgovaraju već poznatim interesovanjima posmatranog korisnika, a zatim se za svaki sadržaj koji se pojavi u listi preporuka lažnih korisnika smatra da će se svideti datom korisniku. Početne informacije se mogu prikupiti ili sa sajta na kome se sistem za pružanje preporuka koristi (ukoliko su javno dostupne), ili sa društvenih mreža gde korisnici često izražavaju svoje mišljenje o stvarima koje vole. U praksi je za većinu napada dovoljno imati informaciju o 8 sadržaja za koje je korisnik izrazio svoje mišljenje [64]. Ipak, kako pojedini sistemi mogu imati mehanizme za detekciju lažnih korisnika, razvijeni su i pasivni napadi. Pasivni napad je demonstiran na sistemu koji formira listu preporuka na osnovu sličnosti između sadržaja. Ukoliko se novi sadržaj pojavi u listi preporuka za nekoliko sadržaja za koje znamo da se sviđaju korisniku, pretpostavlja se da mu se i taj novi sadržaj sviđa. Iako je donošenje zaključaka primenom pasivnog napada znatno teže, moguće je postići dobre rezultate. Na

primer, za korisnike koji koriste *LibraryThing* sajt za preporučivanje knjiga pokazano je da je moguće doneti 58 zaključaka s tačnošću od 50%, ili 6 zaključaka s tačnošću od 90%, u zavisnosti od toga šta je cilj napada [64]. Treba naglasiti da broj zaključaka ne mora biti ključan parameter, jer je u većini slučaja dovoljno doći do jednog zaključka koji otkriva posebno osetljive podatke o ličnosti.

Do narušavanja privatnosti može doći i usled aktivnosti trećeg lica koje nije direktno povezano s procesom pružanja preporuka, ili sa samim sistemom za pružanje preporuka. Tipični primeri ovih napada su hakerski napadi koji imaju za cilj da ukradu podatke korisnika, kao i napadi deanonimizacije podataka [60]. U literaturi se u ovom kontekstu najčešće spominje *Netflix* takmičenje kada su anonimizirani podaci o filmovima koje je korisnik iznajmljivao pušteni u javnost u cilju pronalaženja najboljeg algoritma za pružanje preporuka. Pokazalo se da je uz pomoć adekvatanog algoritma, poznavanja osam ocena filmova - od kojih čak dve ne moraju biti tačne - i približnog datuma ocenjivanja filmova s greškom od 2 nedelje, moguće identifikovati čak 99% korisnika [65]. Autori ovog algoritma su pokazali da bi preturbacije potrebne za sprečavanje napada deanonimizacije uništile korisnost podataka, kao i da je napad moguće izvesti čak i u slučajevima kada su dostupni podaci samo o pojedinim korisnicima. Iako se podaci o filmovima koje je korisnik gledao ne mogu smatrati osetljivim, na osnovu njih je moguće doći do zaključaka kao što su seksualno i političko opredeljenje posmatranog korisnika.

3.4 Moguća rešenja za zaštitu privatnosti

Rešenja za zaštitu privatnosti mogu se grupisati u tehnička i netehnička rešenja [60]. Pod tehničkim rešenjima se podrazumevaju ona rešenja koja na sistemskom i algoritamskom nivou rešavaju problem zaštite podataka, dok se pod netehničkim rešenjima smatraju pravila obrade podataka koja pružalac usluge distribucije medijskog sadržaja mora da poštuje bilo zbog regulatornog okvira ili zato što su postala industrijski standard. Prilikom projektovanja sistema za pružanje preporuka neophodno je uzeti u obzir sve aspekte zaštite privatnosti i kombinovati ova rešenja. U okviru ovog odeljka svaka od ovih grupa rešenja biće detaljnije opisana.

3.4.1 Sistemska rešenja

Pod sistemskim rešenjima za zaštitu privatnosti podrazumevaju se sva rešenja kod kojih je sistem projektovan na takav način da ograničava obelodanjivanje i prosleđivanje podataka,

kao i dovođenje različitih skupova podataka u vezu sa istom osobom. Ovo se može postići korišćenjem softvera u koji korisnik ima poverenje, pružanjem korisniku potpune kontrole nad podacima i prebacivanjem dela ili celog procesa pružanja preporuka na korisničke uređaje [60].

Poverenje između softvera i korisnika može biti bazirano na reputaciji koju softver ima, sertifikatu koji je izdalo treće lice, ili samom procesu obrade koji garantuje poverenje [66]. Na primer, sam proces obrade kod aplikacije za kupovinu na mobilnom telefonu koji prikuplja minimalnu količinu podataka i pri svakom slanju podataka provajderu servisa zahteva dozvolu korisnika, može značajno uticati na formiranje međusobnog poverenja [67]. Na početku svake sesije kupovine kreira se privremena korpa i povezuje s novim pseudonimom, tako da provajder servisa ne može da prati ponašanje korisnika kroz više sesija. U toku svake sesije, aplikacija traži odobrenje pre nego što mobilni telefon počne da šalje prikupljene podatke. Na kraju sesije se privremena korpa i svi prikupljeni podaci brišu. Svaki od koraka ovog procesa se proverava u svakoj sesiji, tako da se i aplikacija na mobilnom telefonu i softver smatraju poverljivim.

Sistemi koji se trenutno koriste ne pružaju mogućnost korisniku da za potrebe novog servisa iskoristi podatke koji su već prikupljeni kod postojećeg servisa i da pritom ima potpunu kontrolu koji deo podataka želi da prebaci. U literaturi postoji predlog sistema koji bi omogućio prebacivanje podataka s jedne društvene mreže na drugu [68]. Ovaj sistem je baziran na tehnologiji semantičkog *weba* i ima jasno definisan format u kome se podaci čuvaju, način na koji se identifikuje profil, kao i vokabular pomoću kojeg korisnik definiše prava pristupa. Kako korisnici često svoje impresije dele na društvenim mrežama, mogućnost prebacivanja ovih podataka bi značajno pomogla u rešavanju problema hladnog starta.

Prebacivanje dela ili celog procesa pružanja preporuka na korisničke uređaje može značajno smanjiti rizik od narušavanja privatnosti sa strane provajdera servisa. Korišćenjem direktne komunikacije između korisnika (*peer-to-peer*) i mere sličnosti koja uzima u obzir privatnost korisnika, moguće je ceo proces pružanja preporuka prebaciti na korisničku stranu [69]. Konkretno za meru sličnosti izabran je procenat u kome se dva skupa ocena poklapaju. Ovakva mera sličnosti nam pruža mogućnost da se poređenje između dva korisnika ne vrši direktno, već da se svaki od skupova ocena uporedi s trećim, i da se na osnovu ovih poređenja donesu zaključci. Na ovaj način se sprečava identifikacija korisnika čiji su podaci poslati radi proračuna mere sličnosti, i samim tim štiti privatnost korisnika. Nasuprot ovome, neki autori

se odlučuju da zbog ograničenih resursa korisničkih uređaja samo deo procesa pružanja preporuka prebace na korisničku stranu. Na primer, sistem koji koristi faktorizaciju matrica može da na serverskoj strani čuva informacije o latentnim faktorima koji se tiču sadržaja, dok se na korisničkom uređaju čuvaju samo informacije o latentnim faktorima koji se tiču korisnika [70]. Kod sistema projektovanog na ovaj način moguće je postići zadovoljavajuće performanse, iako se korisnički podaci ne čuvaju na strani provajdera servisa, i ažuriranje podataka vrši samo na osnovu korisničkih interakcija.

Kao što se može uočiti iz dosadašnjeg izlaganja, sistemska rešenja je pogodno koristiti uglavnom u cilju zaštite od direktnog pristupa podacima.

3.4.2 Algoritamska rešenja

Algoritamska rešenja se mogu grupisati u sledeće četiri kategorije [60]:

- algoritmi bazirani na pseudonima ili anonimizaciji,
- algoritmi koji modifikuju podatke korisnika,
- algoritmi koji garantuju diferencijalnu privatnost i
- kriptografski algoritmi.

Algoritam koji koristi pseudonime može da ponudi korisniku da prilikom korišćenja servisa izabere kojom apstrakcijom želi da bude predstavljen [57]. Različite apstrakcije mogu biti povezane s različitim aktivnostima korisnika (kao što su na primer zabava, kupovina), pa provajderi servisa ne mogu tako lako povezati sve ove aktivnosti s posmatranim korisnikom.

Kod anonimizacije podataka cilj je sprečiti identifikaciju korisnika na osnovu prikupljenih podataka. Ipak, ovo nije tako lako postići jer pored identifikatora, kao što su ime i prezime osobe, mogu postojati i takozvani kvazi-identifikatori, kao što su demografske informacije, koji se mogu iskoristiti za pretraživanje neke eksterne baze podataka i otkrivanje identiteta korisnika [71]. Kako bi se ovo sprečilo najćešće se koriste algoritmi k-anonimizacije [72]. Glavni cilj ovih algoritama je da pomoću tehnika generalizacije i supresije postignu da se svaka kombinacija vrednosti kvazi-identifikatora može povezati s barem k osoba, pa tako nije moguće identifikovati pojedinačnu osobu. Iako su ovi algoritmi veoma popularni, zbog karakteristika podataka koji se prikupljaju kod sistema za pružanje preporuka, kao što su veliki broj dimenzija vektorskog prostora i mali broj ocenjenih sadržaja u odnosu na broj dostupnih, nije ih lako primeniti [65]. U praksi se koriste jednostavniji algoritmi pa su, na primer, kod podataka korišćenih za *Netflix* takmičenje u obzir uzeti samo identifikatori

korisnika koji su anonimizirani tako što su zamenjeni slučajnim brojevima. Kao što smo već diskutovali, ovakav način zaštite nije dovoljan i ovi podaci su deanonimizirani.

Algoritamska rešenja se mogu koristiti i za modifikovanje prikupljenih podataka, kako bi u slučaju krađe sprečili lopova da dođe do tačnih informacija. U literaturi postoji veliki broj ovih rešenja koja na različite načine dodaju slučajni šum, bilo na podatke o ocenama sadržaja, demografske podatke, ili opise sadržaja [60]. Iako je njihov primarni cilj zaštita privatnosti od direktnog pristupa podacima, moguće ih je koristiti i za zaštitu od zaključaka koji se mogu doneti nakon analize ovih podataka. Na primer, za potrebe zaštite od otkrivanja pola korisnika razvijen je algoritam koji u korisnički profil dodaje nove ocene sadržaja koji su visoko korelisani sa sadržajima koje osobe suprotnog pola vole da gledaju [61]. Na ovaj način, u nekim je slučajevima moguće uz dodatak od samo 1% novog sadržaja značajno smanjiti verovatnoću da se otkrije pol korisnika. Od toga koje se rešenje koristi, odnosno na koji se način i koliko se modifikuju podaci, zavisi da li će se i koliko degradirati performanse sistema za pružanje preporuka. Najveća mana ovih rešenja međutim, nije to što mogu dovesti do degradacije performansi, već to što nisu u skladu s načelom tačnosti informacija iz zakonske regulative i što mogu imati negativni psihološki uticaj na korisnika ukoliko bi se pojavile netačne kompromitujuće informacije u javnosti.

Kao što smo već pomenuli, privatnost korisnika može biti narušena bilo zbog direktnog pristupa podacima, bilo zbog zaključaka koji se mogu doneti njihovom obradom. Algoritmi koji garantuju diferencijalnu privatnost sprečavaju napadača da na osnovu prikupljenih podataka o izlazima iz sistema za pružanje preporuka dođe do zaključaka o korisniku. Tačnije, pod pojmom diferencijalna privatnost smatra se osobina sistema da njegov izlaz ne pruža mogućnost donošenja zaključaka o tome da li se određeni podaci koriste kao ulaz u sistem ili ne [60]. Algoritmi koji garantuju ovu osobinu sistema, za razliku od algoritama koji modifikuju podatke, dodaju šum (najčešće s *Laplaceovom* raspodelom) u same proračune koji se obavljaju u sistemu [73]. Količina šuma, odnosno degradacija performansi sistema, može se smanjiti ukoliko se pri proračunima umesto *Laplaceovog* koristi šum koji uzima u obzir promenljivost funkcije korisnosti sadržaja u okruženju prikupljenih podataka [74]. Degradacija performansi sistema u potpunosti nestaje kada se za proračune koristi dovoljna količina podataka [73]. Kao mana ovih rešenja navodi se da u dosadašnjoj literaturi ne postoje radovi koji ispituju mogućnost kontinualnog održavanja diferencijalne privatnosti bez izvršavanja proračuna nad svim prikupljenih podacima, iako takva rešenja već postoje u drugim oblastima primene [60].

Kriptografski algoritmi se koriste u cilju zaštite od direktnog pristupa podacima, bilo od strane ostalih korisnika u sistemu, ili od strane trećih lica zainteresovanih za ove podatke. U literaturi se najčešće koriste algoritmi bazirani na aditivnoj homomorfnoj enkripciji, kao što je Paillierov asimetrični kriptosistem [75], kod kojih nije potrebno vršiti dekripciju pojedinačnih činalaca kako bi se oni sabrali, već se enkriptovana suma može dobiti kao proizvod enkriptovanih činilaca. Ova osobina se može iskoristiti prilikom proračuna u sistemu za pružanje preporuka i na taj smanjiti računska složenost kriptografskih algoritama. Primena kriptografskih algoritama nije ograničena na jednu arhitekturu sistema, pa se oni mogu koristiti i kod klijent-server arhitektura, kao i kod arhitektura kod kojih korisnici direktno komuniciraju i razmenjuju podatke. Kao što je pokazano na primeru sistema s klijent-server arhitekturom, baziranog na *Slope One* algoritmu za pružanje preporuka, enkripcija se može koristiti kako bi zaštitila korisničke interakcije sa serverom [76]. Funkcionisanje ovog sistema se može podeliti na fazu učenja i fazu predikcije. U fazi učenja, kada korisnici šalju svoje podatke u cilju proračuna matrice prosečnih devijacija ocena sadržaja, primenjuju se prethodno opisani algoritmi za modifikovanje podataka ili se koriste pseudonimi. Enkripcija se koristi u fazi predikcije kada korisnici šalju vektor ocena sadržaja enkriptovan javnim ključem servera na osnovu kojeg aplikacija koristeći osobine aditivne homomorfne enkripcije proračunava enkriptovani vektor predikcije. Ovaj vektor se šalje korisničkom uređaju i on ga, koristeći svoj privatni ključ, dekriptuje i prezentuje korisniku konačni rezultat predikcije. Nasuprot ovome, kod reprezentativnog primera sistema kod kojeg korisnici direktno komuniciraju i razmenjuju podatke, aditivna homomorfna enkripcija se kombinuje s procedurom za siguran proračun u kome učestvuju više učesnika kako bi se kompletna komunikacija i sam proračun zaštitio od malicioznih korisnika [77]. Kako bi se iskoristile osobine aditivne homomorfne enkripcije, kao algoritam za pružanje preporuka korišćena je iterativna verzija algoritma SVD (*Singular Value Decomposition*), koja nelinearne proračune aproksimira serijom linearnih koraka. U zavisnosti od toga koju funkciju čvor obavlja u mreži, razlikujemo klijentske čvorove koji prosleđuju enkriptovane podatke o pojedinačnom doprinosu korisnika i čvorove koji obavaljaju agregaciju i proračune, pri čemu jedan uređaj u mreži može da obavlja obe funkcije. Kako bi se osiguralo da podaci budu bezbedni i da proračuni budu tačni, u okviru procedure za proračun u kome učestvuje više učesnika postoje mehanizmi koji proveravaju da li je posmatrani čvor maliciozan. Korišćenjem ove procedure moguće je zaštiti korisničke podatke od direktnog pristupa neovlašćenih lica, čak i u slučajevima kada postoje maliciozni korisnici u sistemu, tačnije kada je broj malicioznih čvorova manji od 50% ukupnog broja čvorova [77]. Svaki od klijenatskih čvorova ima deo

privatnog ključa koji se koristi za dekriptovanje rezultata proračuna. Kada dovoljno veliki procenat klijentskih čvorova dekriptuje proračun enkriptovan od strane čvora koji vrši proračun, taj isti čvor agregira pojedinačne dekripcije i na osnovu agregiranog rezultata formira listu preporuka. Kako bi se smanjili potrebni resursi pojedinačnog uređaja, proračun je distribuiran na više čvorova koji su odgovorni za različite klijentske čvorove. Međutim, i pored značajnog truda uloženog u istraživanje načina za smanjenje proračunskih zahteva, kriptografska rešenja se ne preporučuju za korišćenje kod sistema za pružanje preporuka od kojih se očekuje rad u realnom vremenu [60].

Ono što se može zaključiti iz prethodnog istraživanja je da postoji veliki broj raznovrsnih algoritamskih rešenja za zaštitu privatnosti i da se njihovim kombinovanjem može uticati kako na sprečavanje direktnog pristupa podacima, tako i na sprečavanje donošenja zaključaka o korisnicima.

3.4.3 Zakonska i regulatorna rešenja

Zakonska rešenja za zaštitu privatnosti u Republici Srbiji se baziraju na ZZPL [55] i poštovanju načela obrade podataka koji proizilaze iz njega. Kako su zakonski aspekti zaštite privatnosti već prethodno opisani, u ovom poglavlju im nećemo posvetiti dodatnu pažnju; Umesto toga, biće reči o pečatima poverenja i standardima za zaštitu privatnosti koji se mogu koristiti za samoregulaciju zaštite privatnosti korisnika.

Pečati poverenja, kao što je *BBBOnline* pečat [78], jedna su vrsta sertifikata kojim se korisnicima garantuje da se obrada podataka obavlja na način koji zadovoljava uslove neophodne za dobijanje pečata. Korišćenjem ovih pečata moguće je povećati poverenje korisnika u sistem i smanjiti zabrinutost korisnika za svoju privatnost [79].

Reprezentativan primer standarda za zaštitu privatnosti je P3P (*Platform for Privacy Preferences*) [80]. U okviru ovog standarda je definisan način na koji se proces prikupljanja podataka može opisati i sačuvati u formatu koji je čitljiv za mašinu, što omogućava korisniku da uz pomoć klijentskog programa proveri koliko se polisa kompanije slaže s njegovim preferencijama. Iako korišćenje ovog standarda može povećati poverenje korisnika u sistem, pokazano je da ipak ne smanjuje zabrinutost korisnika za privatnost [79].

Kao što se može videti iz dosadašnjih razmatranja zakonska i regulatorna rešenja za zaštitu privatnosti uglavnom se tiču zaštite od direktnog pristupa podacima, ali su kroz načelo

ograničenosti svrhe ZZPL obuhvaćena i zaštita od sofisticirane obrade podataka u cilju donošenja zaključaka.

3.5 Rezime

U ovom poglavlju smo predstavili:

- rizike za narušavanje privatnosti korisnika koji postoje kod preporučivača – s posebnim fokusom na personalizovane programske vodiče,
- zakonsku regulativu koja uređuje ovu oblast i
- moguća rešenja za zaštitu privatnosti.

Konkretno rešenje koje ćemo koristiti za naš personalizovani programski vodič biće prikazano u poglavlju u kojem je dat opis ovog sistema.

U sledećem poglavlju istražićemo literaturu koja se bavi uticajem kontekstualnih informacija na performanse sistema, ispitati načine na koje se one koriste, kao i o čemu sve treba voditi računa prilikom njihovog izbora.

4. Uticaj kontekstualnih informacija na performanse sistema

4.1 Definicija i podela kontekstualnih informacija

Intuitivno je jasno da na odluku korisnika koji će TV sadržaj izabrati, pored korisničkih interesovanja, mogu da utiču i informacije kao što su vreme, mesto i društvo s kojim se TV sadržaj bira. Kako ove informacije bliže opisuju kontekst u kome korisnik pristupa TV sadržaju nazivaju se kontekstualnim informacijama. Preciznije, pod kontekstualnim informacijama možemo smatrati bilo koje informacije koje se mogu koristiti za opisivanje situacije osobe, mesta ili objekta a koje se smatraju relevantnim za interakciju između korisnika i aplikacije [81].

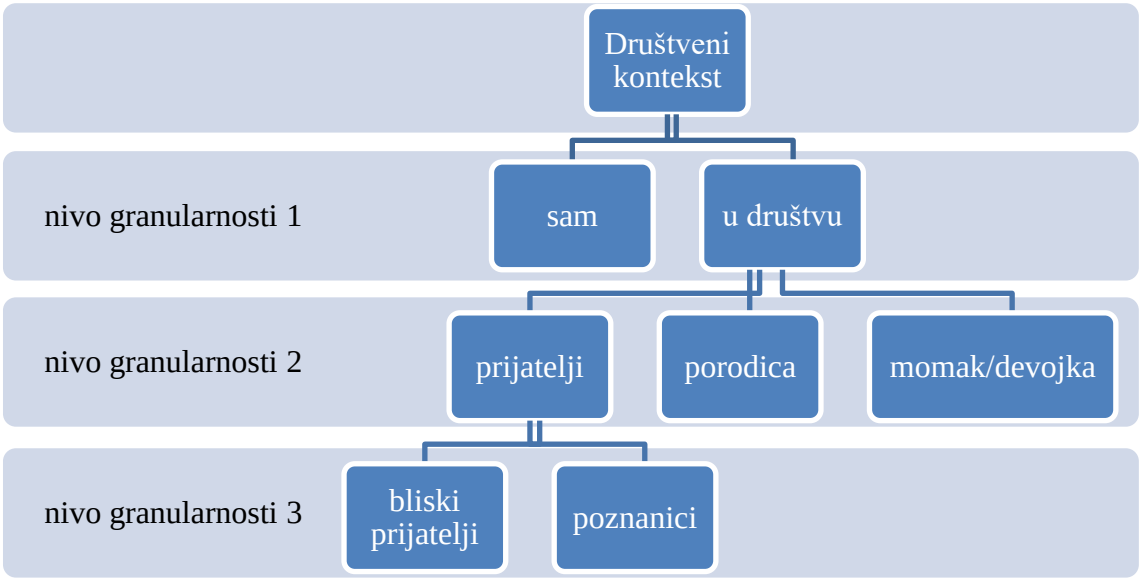
Kontekstualne informacije se mogu podeliti na informacije koje opisuju fizički, modalni, društveni kontekst i kontekst povezan s korisničkim uređajima i sadržajima kojima se pristupa [82]. Fizički kontekst može biti opisan informacijama poput vremena pristupa sadržaju, lokacije korisnika, trenutnih vremenskih uslova i godišnjeg doba. Modalni kontekst obuhvata informacije poput životnih ciljeva korisnika, njegovih iskustava, i trenutnog raspoloženja. Informacije o tome da li je korisnik sam ili u društvu drugih osoba, kao i informacije o odnosu korisnika s tim osobama definišu društveni kontekst. Kontekstualne informacije o uređaju koji se koristi za pristup sadržajima (npr. TV prijemnik, mobilni telefon, računar), kao i o vrsti sadržaja koja se bira (filmovi, muzika, serije) takođe mogu uticati na ponašanje korisnika, pa je potrebno i njih uzeti u obzir.

4.2 Izbor kontekstualnih informacija

Pri izboru kontekstualnih informacija koje će se koristiti u sistemu neophodno je razmotriti njihovu relevantnost i hijerarhijsku strukturu, kao i načine na koje se pojedinačne informacije mogu prikupiti [13].

Imajući u vidu veliki broj i raznolikost kontekstualnih informacija logično je da nisu sve informacije podjednako relevantne za svaku primenu. Na primer, u slučaju personalizovanog programskog vodiča za digitalnu televiziju, mala je verovatnoća da će informacija o vrednosti akcija na berzi direktno imati uticaja na korisnikov izbor TV sadržaja. Relevantnost pojedinačnih informacija moguće je odrediti na osnovu mišljenja eksperata iz oblasti primene sistema ili korišćenjem tehnika mašinskog učenja.

Većina kontekstualnih informacija ima hijerahijsku strukturu [83]. Na slici 4.1, prikazan je primer hijerarhijske strukture kontekstualnih informacija koje definišu društveni kontekst. Kako bi u potpunosti odredili kontekstualne informacije koje ćemo koristiti neophodno je izabrati nivo granularnosti informacija ili koristiti takozvanu automatsku generalizaciju konteksta [84]. Iako viši nivoi granularnosti informacija preciznije opisuju situaciju u kojoj se korisnik nalazi, izbor višeg nivoa granularnosti ne mora dovesti do poboljšanja performansi sistema. Na primer, posmatrajmo korisnika koji vikendom voli da gleda filmove sa svojom devojkom. Ukoliko umesto nižeg nivoa granularnosti informacije o danu u nedelji (radni dan, vikend, praznik), usvojimo viši nivo granularnosti (ponedeljak, utorak, sreda, četvrtak, petak, subota, nedelja), sistem neće bolje naučiti ponašanje korisnika. Čak šta više ovo može dovesti do degradacije performansi jer se povećava verovatnoća da usled nedovoljnog broja prikupljenih podataka za pojedinačne kontekste sistem ne može da nauči šablon korisničkog ponašanja.



Slika 4.1: Hijerarhijska struktura kontekstualnih informacija [13].

Automatska generalizacija konteksta predstavlja alternativu statičkom izboru nivoa granularnosti kontekstualnih informacija. Kod generalizacije konteksta u toku rada sistema bira se optimalni nivo granularnosti svake od korišćenih kontekstualnih informacija čime se postižu maksimalne performanse sistema. Da bi odredili optimalni nivo granularnosti informacija potrebno je ispitati sve kombinacije nivoa granularnosti svih kontekstualnih informacija. Imajući u vidu broj kontekstualnih informacija koje se mogu prikupiti i mogući

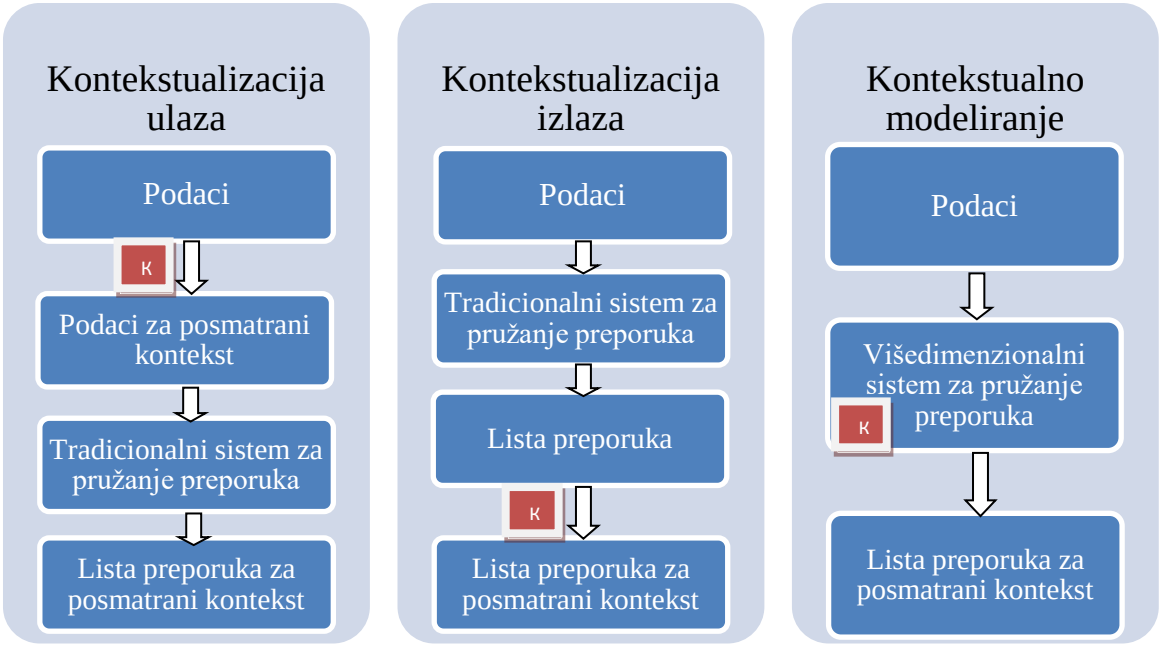
broj nivoa granularnosti svake od njih, korišćenje automatske generalizacije konteksta može značajno povećati složenost sistema.

Pri konačnom izboru kontekstualnih informacija koje će se koristiti neophodno je uzeti u obzir i načine na koje se pojedinačne informacije mogu prikupljati [13]. Pojedine informacije, kao što je na primer društvo u kojem se korisnik nalazi, mogu zahtevati eksplicitno učešće korisnika u procesu njihovog prikupljanja, bilo kroz popunjavanje upitnika ili kroz akciju pritiskanja dugmeta. Korišćenje informacija koje se eksplicitno prikupljaju, ometa uobičajen način gledanja televizije i može negativno uticati na zadovoljstvo korisnika. S druge strane, postoje informacije koje se mogu implicitno prikupljati bez ikakvog učešća korisnika korišćenjem samo podataka dostupnih na korisničkom uređaju. Kako se s razvojem korisničkih uređaja (posebno mobilnih telefona) broj i raznolikost dostupnih senzora značajno povećao sve je više informacija koje se mogu dobiti na ovaj način. Tipičan primer informacije koja se može implicitno prikupiti je vreme kada je korisnik gledao TV sadržaj. Pored direktnih (eksplicitnog i implicitnog) načina za prikupljanje informacija, primenom statističkih metoda i tehnika mašinskog učenja nad podacima koji opisuju ponašanje korisnika moguće je bliže odrediti kontekst u kome korisnik bira TV sadržaj. Na ovaj način se, po cenu povećanja složenosti sistema, može doći i do informacija kao što je omiljeni žanr filmova, čije bi prikupljanje inače zahtevalo eksplicitno učešće korisnika. Kako sakupljanje kontekstualnih informacija može zahtevati učešće korisnika i uticati na složenost sistema potrebno je izabrati samo one informacije koje omogućavaju implementaciju efikasnog personalizovanog programskog vodiča, a ne ometaju uobičajeni način gledanja televizije.

Treba napomenuti da pored goreopisanog shvatanja konteksta postoji i interaktivno shvatanje konteksta kod kojeg nije moguće definisati kontekst pomoću unapred izabranih kontekstualnih informacija, jer se smatra da se relevantni kontekst dinamički menja usled same aktivnosti korisnika [85]. Na primer, posmatrajmo situaciju u kojoj momak i devojka gledaju ljubavni film. U nekom trenutku devojka spomene momku kako je donet novi zakon koji može znatno uticati na poslovanje njegove firme i da će o tome biti više reči u vestima koje počinju za nekoliko minuta. Ovakav razvoj situacije može dovesti do toga da momak prebaci kanal na vesti i da kontekstualna informacija o društvu s kojim je izgubi relevantnost. Interaktivno shvatanje konteksta se ređe koristi i da bi se primenilo neophodno je koristiti modele iz oblasti psihologije.

4.3 Načini za korišćenje kontekstualnih informacija

U zavisnosti od toga u kojoj se fazi procesa pružanja preporuka koriste kontekstualne informacije razlikujemo kontekstualizaciju ulaza, kontekstualizaciju izlaza i kontekstualno modeliranje [13]. Na slici 4.2 prikazani su procesi pružanja preporuka koji odgovaraju ovim načinima za korišćenje kontekstualnih informacija.



Slika 4.2: Tri načina korišćenja kontekstualnih informacija u sistemu [13].

Kod kontekstualizacije ulaza informacije o kontekstu se koriste kako bi se izabrali samo oni podaci koji odgovaraju posmatranom kontekstu *K*, a zatim tradicionalni sistemi na osnovu ovih podataka mogu da formiraju listu preporuka. Na ovaj način problem pružanja preporuka u prisustvu konteksta se svodi na tradicionalni problem pružanja preporuka pa je moguće koristiti mnogobrojne sisteme koji su već razvijeni za ove potrebe [15]. Kako se kod kontekstualizacije ulaza vrši transformacija prostora (korisnici, TV sadržaji, kontekstualne informacije) u prostor (korisnici, TV sadržaji), ovi sistemi su posebno osetljivi na izbor nivoa granularnosti kontekstualnih informacija. Previše velika granularnost informacija može dovesti do toga da nema dovoljno podataka o posmatranom kontekstu, neophodnih da bi sistem naučio korisnička interesovanja, pa se predlaže korišćenje generalizacije konteksta i kombinovanje sa sistemima koji ne koriste kontekst [84].

Kod kontekstualizacije izlaza se lista preporuka najpre formira korišćenjem tradicionalnih sistema bez razmatranja konteksta, a zatim se modifikuje u skladu s posmatranim kontekstom

K. Modifikacija liste može biti takva da se TV sadržaji koji nisu relevantni za posmatrani kontekst izbace iz nje ili da se izvrši ponovno rangiranje TV sadržaja prema njihovoj relevantosti za posmatrani kontekst. Za proračun relevantnosti sadržaja u listi se mogu koristiti sistemi na bazi heuristike ili sistemi kod kojih se formira model korisničkog ponašanja. Relevantnost sadržaja se kod sistema koji koriste heuristiku može proceniti na osnovu sličnosti karakteristika TV sadržaja koje je korisnik gledao u datom kontekstu (na primer omiljeni glumci u filmovima koje korisnik gleda vikendom). S druge strane sistemi kod kojih se formira model korisničkog ponašanja mogu na svom izlazu da proračunaju verovatnoću izbora određenog tipa TV sadržaja u posmatranom kontekstu (na primer verovatnoća da će korisnik izabrati određeni žanr filmova u društvu devojke) i da ovu vrednost iskoriste kao relevantnost sadržaja. Kao i kod kontekstualizacije ulaza, za formiranje liste preporuka se mogu koristiti mnogobrojni tradicionalni sistemi za pružanje preporuka [15].

Kod kontekstualnog modeliranja informacije o posmatranom kontekstu *K* se direktno koriste u sistemima za pružanje preporuka. Da bi se ove informacije mogle iskoristi potrebno je da sám sistem pri proračunu korisnosti TV sadržaja može da uzme u obzir i kontekstualne informacije. Na ovaj način se, pored dimenzija koje definišu TV sadržaj i korisnike sistema, uvode dodatne dimenzije koje definišu kontekstualne informacije pa se ovi sistemi nazivaju višedimenzionalnim sistemima za pružanje preporuka. Treba naglasiti da ovo ne znači da nije moguće iskoristiti tehnike razvijene za potrebe tradicionalnih sistema već da ih je potrebno modifikovati. Kao i kod tradicionalnih sistema razlikujemo sisteme na bazi heuristike i sisteme kod kojih se formira model ponašanja korisnika. Višedimenzionalni sistemi na bazi heuristike mogu koristiti kontekstualne informacije za proračun višedimenzionalne mere sličnosti ili u nekom drugom koraku proračuna kao što je na primer kod koloborativnog filtriranja, izbor podataka o korisniku koje će se koristiti za formiranje liste suseda [13]. Tako se za svakog od korisnika mogu formirati kontekstualni profili, skupovi podataka koji odgovaraju pojedinačnim kontekstima, a zatim korišćenjem kosinusne sličnosti pronaći:

1. najsličniji kontekstualni profili,
2. podjednak broj najsličnijih profila za svaki od mogućih konteksta (bez obzira na to kom nivou granularnosti u hijerarhiji pripadaju),
3. najsličniji kontekstualni profili koji pripadaju istom nivou granularnosti u hijerahiji kojoj pripada trenutni kontekst,

4. podjednak broj najsličnijih profila za svaki od konteksta koji pripada istom nivou granularnosti u hijerarhiji kojoj pripada trenutni kontekst [86].

Lista preporuka se kod ovakvog pristupa formira na sličan način kao kod tradicionalnih kolaborativnih sistema, rangiranjem težinskih suma ocena sadržaja iz profila izabranih na jedan od goreopisanih načina. Kao težinski koeficijenti koriste se vrednosti kosinusne sličnosti između profila.

Nasuprot ovome, kod sistema kod kojih se formira model korisničkog ponašanja vektorski prostor kojim su definisani sadržaji lako se može proširiti dodatnim dimenzijama koje odgovaraju kontekstualnim informacijama, a zatim algoritmi mašinskog učenja pravolinijski iskoristiti kako bi sistem naučio korisnička interesovanja u ovom vektorskom prostoru. Na primer, SVM (*Support Vector Machine*) algoritam može se iskoristiti za klasifikaciju objekata koje preporučivač nudi (u posmatranom slučaju restorana) na one koje se korisniku sviđaju i na one koje mu se ne sviđaju [87]. Ponuđeni objekti predstavljeni su tačkama u višedimenzionalnom vektorskom prostoru čije koordinate između ostalog pružaju informaciju i o datumu, društvu i trenutnim vremenskim uslovima. Kako bi što bolje naučio korisnička interesovanja sistem ima za cilj da pronađe hiperravan koja s najvećom marginom razdvaja objekte koji se korisniku dopadaju od onih koji se korisniku ne dopadaju. Korišćenjem tehnike filtriranja sadržaja proračunati *SVM* modeli se mogu direktno iskoristiti za formiranje liste preporuka, ili se uzimajući u obzir koliko dobro *SVM* model jednog korisnika klasifikuju sadržaje koje je drugi korisnik ocenio može proračunati sličnost između korisnika i primeniti tehnika kolaborativnog filtriranja [87].

Uz pomoć veštačke neuralne mreže moguće je implementirati hibridni sistem za pružanje preporuka koji koristi kontekstualno modeliranje kako bi što bolje naučio ponašanje korisnika digitalne televizije [12]. Neuralna mreža na osnovu podataka o žanrovima ponuđenih TV sadržaja, grupnog profila korisnika, i raspoloženja korisnika pokušava da predvidi koje bi sadržaje korisnik voleo da gleda a koje ne. Grupni profil korisnika se najpre formira grupisanjem korisnika na osnovu podataka o njegovom životnom stilu, omiljenim žanrovima i demografskih podataka, a zatim se primenom kolaborativnog filtriranja pronalaze dodatni žanrovi koji bi mogli interesovati korisnika. Kontekstualna informacija o raspoloženju značajno utiče na izbor TV sadržaja, pa na osnovu postojećih istraživanja možemo očekivati da će korisnici koji su dobro raspoloženi birati sadržaje s dosta akcije i drame, dok će oni koji nisu raspoloženi u opštem slučaju izbegavati komedije [88]. Kako bi se ovi podaci prikupili, korisnici moraju eksplicitno da pritisnu jedno od dugmadi na daljinskom upravljaču koje

odgovaraju dobrom, lošem i neutralnom raspoloženju, što može ometati uobičajeni način gledanja televizije.

Jasan odgovor na pitanje na koji način je najbolje iskoristiti kontekstualne informacije u sistemu još uvek ne postoji. Koliko je nama poznato opsežno istraživanje postoji samo za sisteme koji koriste tehniku koloborativnog filtriranja kako bi preporučili objekte na sajtovima za prodaju preko Interneta [89]. Kod ovog poređenja performansi u obzir su uzeti:

- ciljevi sistema (da li je je cilj preporučiti nekoliko najboljih sadržaja ili sve sadržaje koji odgovaraju korisničkim interesovanjima),
- kontekstualne informacije (koja je namera korisnika pri kupovini proizvoda, informacija o godišnjem dobu i o kategoriji proizvoda za koje je korisnik zainteresovan),
- nivoi granularnosti kontekstualnih informacija,
- podaci o korisničkom ponašanju s različitim karakteristikama (s različitim procentom ukupnog sadržaja koji je ocenjen i različitim nivoima hetereogenosti korisničkog ponašanja).

Kao mere performansi posmatrane su tačnost pružanja preporuka i raznolikost sadržaja u listi. Ispitane su performanse koloborativnih sistema na bazi heuristike koji koriste:

- kontekstualizaciju ulaza bez automatske generalizacije konteksta,
- kontekstualizaciju izlaza s filtriranjem liste preporuka,
- kontekstualizaciju izlaza s ponovnim rangiranjem liste preporuka, i
- prethodno opisane verzije sistema na bazi kontekstualnih profila [86], kao predstavnike sistema koji koriste kontekstualno modeliranje.

Što se tiče tačnosti preporuka, istraživanje je pokazalo da se najbolje performanse mogu postići sistemima koji koriste kontekstualizaciju izlaza sa filtriranjem liste preporuka i sistemima koji koriste kontekstualne profile. Kod potonjih nijedan od načina za izbor liste suseda nije dominantan u svim slučajevima. S druge strane, što se tiče raznovrsnosti preporuka, u većini slučajeva najbolje performanse se postižu sistemima na bazi kontekstualizacije izlaza s ponovnim rangiranjem liste preporuka. U preostalim slučajevima, kada je veliki procenat ponuđenog sadržaja već ocenjen i ponašanje korisnika heterogeno, nije teško postići dobru raznovrsnost preporuka pa svi sistemi sem kontekstualizacije izlaza s filtriranjem preporuka postižu dobre performanse. Karakteristike prikupljenih podataka različito utiču na tačnost i na raznovrsnost preporuka, pa nije lako istovremeno postići dobre

vrednosti ovih mera performansi. Ukoliko je ponašanje korisnika, za koje su podaci prikupljeni, heterogeno, nije teško postići dobru raznovrsnost preporuka ali je zato znatno teže postići zadovoljavaljuću tačnost preporuka, i obrnuto. Dobar kompromis između ovih mera performansi se može postići korišćenjem sistema na bazi kontekstualnog modeliranja [89]. Što se tiče nivoa granularnosti kontekstualnih informacija, i u ovom istraživanju je potvrđena činjenica da previsok nivo granularnosti može negativno uticati na tačnost preporuka, dok na raznovrsnost preporuka nema nikakvog uticaja. Kao mogući pravci daljeg istraživanja predloženo je razmatranje mera performansi koje opisuju koliko je sadržaj u listi preporuka iznenađujuć i nov za korisnika, poređenje performansi sistema na bazi filtriranja sadržaja, i poređenje performansi sistema koji kombinuju različite načine za korišćenje konteksta.

Pored toga što se kontekstualne informacije mogu iskoristiti na različite načine u cilju proračuna korisnosti ponuđenog sadržaja, one se mogu iskoristiti i za određivanje najboljeg načina za predstavljanje sadržaja korisnicima. Na primer, u sistemu za pružanje preporuka multimedijalnog sadržaja na mobilnim telefonima, kontekstualne informacije o lokaciji korisnika i vremenu pristupanja sadržaju se mogu iskoristiti u procesu proračuna korisnosti sadržaja, a kontekstualne informacije koje opisuju mogućnosti korisničkog uređaja za izbor nivoa kvaliteta sadržaja koji će se ponuditi korisniku [90]. Korisnost sadržaja se najpre računa koristeći tradicionalnu tehniku filtriranja sadržaja bez korišćenja kontekstualnih informacija. Kontekstualne informacije kojima je opisan fizički kontekst se zatim zajedno s karakteristikama ponuđenog sadražaja (žanr, glumci, opis) koriste u *Naïve Bayes* klasifikatoru kako bi se proračunala verovatnoća da posmatrani sadržaj bude izabran u datom kontekstu. Konačna lista preporuka se formira rangiranjem ponderisane sume sličnosti proračunate primenom filtriranja sadržaja i vrednosti dobijene korišćenjem *Naïve Bayes* klasifikatora. Korišćenjem metoda za kreiranje pravila i kontekstualnih informacija o mogućnostima korisničkog uređaja i brzini pristupa mreži određuje se format sadržaja (video, slika ili tekst) kao i kvalitet sadržaja koji će se ponuditi korisniku (npr. običan ili visokokvalitetan video snimak).

Veoma je važno istražiti i objasniti načine na koje kontekstualne informacije mogu poboljšati performanse sistema. U jednom od praktičnih istraživanja sistema za pružanje preporuka TV sadržaja pokazano je da kontekstualne informacije o vremenu kada je korisnik gledao TV sadržaj mogu značajno poboljšati rangiranje TV sadržaja koje korisnik nije gledao ili čije epizode (u slučaju TV serija) nije gledao [91]. Ukoliko je korisnik odgledao barem jednu epizodu neke serije smatralo se da je korisnik upoznat s radnjom i kvalitetom te serije.

Nasuprot ovome, korišćenje kontekstualnih informacija kod pružanja preporuka TV sadržaja (na primer serija i utakmica) koje korisnik ima običaj da gleda ne poboljšava značajno performanse sistema. Kao mogući razlog istaknuto je to da je problem pružanja preporuka ovih sadržaja značajno jednostavniji pa tradicionalni sistemi, koji ne koriste kontekstualne informacije, mogu dovoljno dobro da nauče ponašanje korisnika. Kao što je u ovom istraživanju naglašeno gledanje televizije je društvena i porodična aktivnost, pa nije redak slučaj da više korisnika, bilo istovremeno ili u različitim terminima, koriste isti uređaj za pristup TV sadržajima. Imajući ovo u vidu performanse sistema za pružanje preporuka mogu biti degradirane ukoliko sistem ne može da zaključi kad koji korisnici gledaju televiziju i greškom preporuči TV sadržaje koji odgovaraju interesovanjima drugih korisnika koji trenutno ne koriste ovu uslugu. Kombinovanjem informacija o vremenu emitovanja TV sadržaja i trenutnom sadržaju kojem se pristupa, sistem može bliže odrediti koji korisnik ili korisnici trenutno gledaju televiziju. Na primer, posmatrajmo situaciju u kojoj otac i dete imaju običaj da gledaju televiziju ujutru, a majka uveče. Ukoliko osoba u jutarnjim časovima trenutno gleda crtani film sistem će pretpostaviti da je pred televizorom dete, dok ukoliko se u istom vremenskom terminu trenutno gledaju vesti sistem će pretrepostaviti da je trenutni korisnik servisa otac. S druge strane, ukoliko osoba u večernjim časovima pristupa sadržajima koji odgovaraju interesovanjima odrasle osobe, sistem će identifikovati majku kao trenutnog korisnika. Informacije o trenutnom kontekstu se u ovom istraživanju koriste kroz dvostruku primenu kontekstualizacije izlaza s ponovnim rangiranjem liste preporuka. Pri formiranju inicijalne liste preporuka korisnost TV sadržaja proračunata tehnikom kolaborativnog filtriranja na bazi faktorizacije matrica množi se sa težinskim faktorom koji odgovara verovatnoći da će se posmatrani sadržaj gledati nakon trenutnog TV sadržaja. Konačna lista preporuka se formira na osnovu vrednosti dobijene množenjem korisnosti sadržaja proračunate u prethodnom koraku ss težinskim faktorom koji odgovara verovatnoći da će se posmatranom sadržaju pristupiti u posmatranom vremenskom terminu.

4.4 Rezime

U ovom poglavlju smo:

- definisali pojam kontekstualnih informacija,
- predstavili njihovu hijerarhijsku strukturu,
- diskutovali njihovu relevantnost i granularnost,
- istražili načine za njihovo prikupljanje i korišćenje,

- prikazali na koji način one poboljšavaju performanse preporučivača i
- dali smernice na osnovu kojih se biraju kontekstualne informacije.

U narednom ćemo opisati opšte karakteristike neuralnih mreža i specifičnosti njihove primene u preporučivačima.

5. Primena neuralnih mreža u preporučivačima

Neuralna mreža predstavlja sistem za obradu informacija koji pokazuje osobine učenja, memorisanja i generalizacije na osnovu podataka kojima se mreža obučava [92]. Ideja za razvoj ovakvog načina obrade u računarstvu je potekla od nervnog sistema čoveka. Veštačka neuralna mreža, kako je drugačije nazivaju, sastoji se od čvorova (neurona) grupisanih u slojeve na način koji omogućava paralelnu obradu. Svaki od čvorova koji učestvuje u proračunima računa ponderisanu sumu ulaznih podataka i na njih primenjuje aktivacionu funkciju, imitirajući sam proces funkcionisanja ljudskih neurona:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \qquad 5.1$$

gde je x_i *i*-ti ulaz u čvor, w_i težina koja odgovara vezi između *i*-tog ulaza u čvor i posmatranog čvora, b tzv. *bias* posmatranog čvora, y njegov izlaz i f aktivaciona funkcija.

U zavisnosti od funkcije samog sloja razlikujemo

- ulazni,
- skriveni i
- izlazni sloj.

Čvorovi neuralne mreže koji pripadaju ulaznom sloju (ulazni čvorovi) ne učestvuju u samim proračunima, već imaju za cilj da prihvate ulazne podatke u mrežu. Nasuprot ovome, čvorovi koji pripadaju skrivenom sloju (skriveni čvorovi) i izlaznom sloju (izlazni čvorovi) vrše goreopisano procesiranje i međusobno se razlikuju po tome što izlaz iz izlaznih neurona predstavlja konačni rezultat obrade, što nije slučaj kod skrivenih čvorova. U mreži može postojati samo jedan ulazni i izlazni sloj, ali i veliki broj skrivenih [93]. U zavisnosti od primene neuralne mreže bira se arhitektura neuralne mreže.

5.1 Arhitektura neuralne mreže

Prema tome kako su slojevi međusobno povezani [94], neuralne mreže se mogu podeliti na:

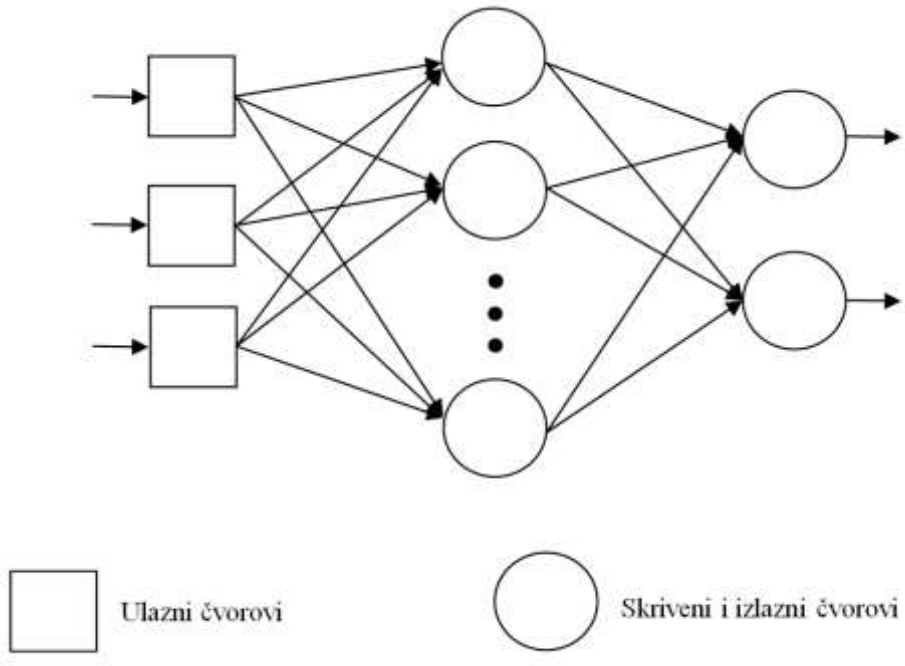
- neuralne mreže s tokom podataka u jednom smeru (*feedforward*),
- parcijalno rekurentne neuralne mreže i
- rekurentne neuralne mreže.

Neuralne mreže s tokom podataka u jednom smeru koriste se u slučajevima kada se očekuje da izlaz zavisi samo od ulaznih podataka, a ne i od redosleda kojim se oni koriste.

Parcijalno rekurentne neuralne mreže koriste se u situacijama kada se očekuje da i sam redosled korišćenja ulaznih podataka ima uticaja na rezultat, pa se kod njih podaci koji su trenutno na ulazu kombinuju s informacijama o prethodnim vrednostima ulaza. Sama distribucija informacija o prethodnim ulazima vrši se kroz rekurentne veze iz skrivenog ili izlaznog sloja do ulaznog sloja.

Rekurentne neuralne mreže karakteriše postojanje dvosmernih veza između svih slojeva. Iako se mogu koristiti za znatno složenije probleme nego neuralne mreže s tokom podataka u jednom smeru, vreme potrebno za treniranje ove vrste neuralne mreže može biti veoma dugo i nije ga lako proceniti.

Koji će se način povezivanja slojeva koristiti, zavisi od cilja preporučivača. U ovome poglavlju, naše razmatranje ćemo ograničiti na neuralne mreže s tokom podataka u jednom smeru, koje je pogodno koristiti u slučajevima kada se vrši dugoročna predikcija sporo promenljivih korisničkih interesovanja [95]. Njihova opšta struktura prikazana je na slici 5.1.



Slika 5.1: Neuralna mreža s tokom podataka u jednom smeru.

Iako je parcijalno rekurentne neuralne mreže pogodno koristiti u slučajevima kada postoje informacije o redosledu pristupanja sadržajima i kad kontekstualne informacije imaju

presudnu ulogu prilikom izbora sadržaja [96], one su ostavljene za buduća istraživanja. S druge strane, rekurentne neuralne mreže, nisu pogodne za korišćenje u svrhu pružanja preporuka zbog nepredvidivog i potencijalno dugog trajanja učenja korisničkih interesovanja.

Kako je broj ulaznih čvorova jednak dimenziji vektorskog prostora nad kojim su definisani ulazni podaci, da bi u potpunosti odredili arhitekturu mreže neophodno je definisati još i brojeve izlaznih čvorova, skrivenih slojeva i čvorova u svakom od njih.

Broj izlaznih čvorova zavisi od ciljeva preporučivača koji su detaljnije razmotreni u drugom poglavlju, gde je diskutovan izbor mera performansi. Imajući u vidu raznolikost primena, u doktorskoj disertaciji usvojili smo da je primarni cilj sistema da odluči koje bi TV sadržaje korisnik voleo da gleda a koje ne, te se problem pružanja preporuke svodi na problem klasifikacije ponuđenog sadržaja [24], [97]. U ovim primenama, broj čvorova određen je brojem klasa [98].

Što se tiče broja skrivenih slojeva i broja čvorova u njima, oni moraju biti što manji da bi neuralna mreža mogla da nauči korisnička interesovanja u realnom vremenu, ali i dovoljno veliki da ne bi došlo do degradacije performansi preporučivača. Kako neuralna mreža s jednim skrivenim slojem i sigmoidnom aktivacionom funkcijom skrivenih čvorova može dovoljno dobro aproksimirati bilo koju nelinearnu zavisnost između ulaza i izlaza mreže [99], odlučili smo se za korišćenje ove arhitekture.

Sigmoidna funkcija definisana je izrazom:

$$f(x) = \frac{1}{1+e^{-x}}, \tag{5.2}$$

Ovo je jedna od najčešće korišćenih aktivacionih funkcija skrivenih čvorova, pa ne čudi da je to slučaj i u našoj primeni [24].

Broj skrivenih čvorova zavisi i od količine dostupnih podataka. Ukoliko je dostupna velika količina podataka i izabrani broj skrivenih čvorova previše mali, neuralna mreža neće moći adekvatno da modelira posmatrani problem. Nasuprot tome, ukoliko je dostupna mala količina podataka i izabrani broj skrivenih čvorova je previše veliki, neuralna mreža će previše dobro naučiti podatke kojima je trenirana i imati lošu sposobnost generalizacije [99]. Gubitak sposobnosti generalizacije neuralne mreže usled neadekvatnog izbora broja skrivenih čvorova naziva se problemom previše istrenirane neuralne mreže.

Iako postoji više teorijskih načina za izbor broja skrivenih čvorova u praksi se najčešće on dobija na osnovu eksperimentalnih rezultata. Brojevi skrivenih čvorova, međutim, dobijeni na neki od teorijskih načina se često koriste kao gornja granica pa ih je pogodno spomenuti. Jedan od teorijskih načina za izbor broja čvorova je *Hecht-Nielsenova* teorema koja kaže da se bilo koja kontinualna funkcija može aproksimirati sa *feedforward* neuralnom mrežom s jednim skrivenim slojem i $2N+1$ skrivenih čvorova, gde N predstavlja broj ulaznih čvorova [100]. Na konkretnom primeru personalizovanog programskog vodiča za digitalnu televiziju koristićemo najmanji broj skrivenih čvorova koji daje zadovoljavajuće performanse.

Vreme treniranja preporučivača može se dodatno skratiti smanjenjem broja veza između slojeva neuralne mreže ili smanjenjem dimenzije vektorskog prostora kojim su opisani podaci. U radu [24] koristi se arhitektura kod koje su ulazni čvorovi na koje dolaze podaci relevantni za kolaborativno filtriranje i ulazni čvorovi na koje dolaze podaci relevantni za filtriranje sadržaja povezani s različitim delovima skrivenog sloja mreže. Na ovaj način moguće je smanjiti broj veza između čvorova, a samim tim i ubrzati trening neuralne mreže, uz očuvanje performansi sistema. Nasuprot ovome, kako je kod sistema na bazi filtriranja sadržaja broj ulaznih čvorova određen dimenzionalnošću vektorskog prostora kojim su opisani ponuđeni sadržaji, vreme treniranja može se teorijski smanjiti primenom linearne transformacije nad originalnim vektorskim prostorom koja bi smanjila njegove dimenzije, a samim tim i broj veza između ulaznog i skrivenog sloja [101]. Ovaj pristup smo detaljnije ispitali i u našim eksperimentima.

5.2 Algoritmi učenja neuralne mreže

Koliko je nama poznato, opsežno istraživanje koje se bavi izborom algoritma učenja preporučivača baziranih na neuralnim mrežama ne postoji.

U opštem slučaju, neuralna mreža uči o nekom problemu kroz podešavanje težina koje odgovaraju vezama između čvorova. U zavisnosti od toga koji su podaci dostupni i šta je cilj učenja razlikujemo algoritme nadgledanog i nenadgledanog učenja. Dok su kod nadgledanog učenja pored ulaznih podataka, dostupni i podaci o očekivanim izlazima, pa je moguće uporediti dobijene izlaze s očekivanim i na osnovu proračunate greške, podesiti težine na taj način da se greška minimizira, to kod nenadgledanog učenja nije slučaj. Kod nenadgledanog učenja, poznati su samo ulazi, pa mreža uči o problemu samo na osnovu karakakteristika i regularnosti koje postoje u podacima. Iako, se jedinom prednošću personalizovanih

programskih vodiča koji koriste nenadgledano učenje može smatrati to što neće ometati uobičajen način gledanja televizije, adekvatnim izborom načina za prikupljanje korisničkih interakcija [42], ona se gubi. Stoga smo se u doktorskoj disertaciji fokusirali na algoritme nadgledanog učenja.

Kao tipične predstavnike različitih klasa algoritama nadgledanog učenja, razmotriću:

- RP (*Resilient backPropagation*) [102],
- SCG (*Scaled Conjugate Gradient*) [103],
- LM (*Levenberg-Marquardt*) [104] i
- ELM (*Extreme Learning Machine*) [105].

Prva tri navedena algoritma koriste tradicionalni načini treniranja kod kojeg sve težine tretiraju jednako i iterativno podešavaju tako da minimiziraju funkciju greške $E(\mathbf{w})$, krećući se unazad od težina koje spajaju izlazni i skriveni sloj ka težinama koje spajaju ulazni i skriveni sloj mreže. Funkcija greške $E(\boldsymbol{w})$ na izlazu iz mreže se dobija na osnovu izraza

$$E(\mathbf{w}) = \sum_{p=1}^{P}\sum_{m=1}^{M}\left(z_{mp} - y_{mp}\right)^2, \tag{5.3}$$

gde je z_{mp} očekivani izlaz m-tog izlaznog čvora za p-ti podatak u trening skupu, y_{mp} dobijeni izlaz m-tog izlaznog čvora za p-ti podatak u trening skupu, M broj izlaznih čvorova i P broj podataka u trening skupu. Sve težine neuralne mreže bez obzira između kojih slojeva se nalazile kod tradicionalih algoritama predstavljene su s matricom težina $\mathbf{w}$.

Nasuprot ovome pristupu, ELM algoritam se bazira na novoj metodologiji za treniranje neuralnih mreža s jednim skrivenim slojem kod koje se prilikom treninga podešavaju samo izlazne težine neuralne mreže i ne koristi iterativno podešavanje.

Svaki od algoritma detaljnije smo opisali u daljem tekstu disertacije.

5.2.1 RP algoritam

RP algoritam je jedna od najčešće korišćenih varijanti BP (*BackPropagtion*) algoritma. Kao i kod svih varijanti BP algoritma težine neuralne mreže se podešavaju u smeru najstrmije opadajućeg gradijenta funkcije greške, ali za razliku od osnovnog algoritma veličina gradijenta nema direktan uticaj na promenu vrednosti težina - kod RP algoritma se u obzir

uzima samo znak gradijenta koji ukazuje na smer u kom treba modifikovati vektor težina. Modifikacija se vrši prema formuli

$$\Delta w_{ij}^{\,k} = \begin{cases} -\Delta_{ij}^{k}, \text{ za } \dfrac{\partial E}{\partial w_{ij}}^{k} > 0 \\ +\Delta_{ij}^{k}, \text{ za } \dfrac{\partial E}{\partial w_{ij}}^{k} < 0 \\ 0, \text{inače} \end{cases}, \qquad 5.4$$

gde je $\Delta w_{ij}^{\,k}$ vrednost promene težine između *i*-tog čvora posmatranog sloja i *j*-tog čvora narednog sloja u *k*-toj iteraciji algoritma. Adaptivno podešavanje same vrednosti obavlja se na osnovu sledećeg izraza

$$\Delta_{ij}^{k} = \begin{cases} \eta^{+} \cdot \Delta_{ij}^{k-1}, \text{ za } \dfrac{\partial E}{\partial w_{ij}}^{k-1} \cdot \dfrac{\partial E}{\partial w_{ij}}^{k} > 0 \\ \eta^{-} \cdot \Delta_{ij}^{k-1}, \text{ za } \dfrac{\partial E}{\partial w_{ij}}^{k-1} \cdot \dfrac{\partial E}{\partial w_{ij}}^{k} < 0 \\ \Delta_{ij}^{k-1}, \text{inače} \end{cases}. \qquad 5.5$$

pri čemu mora da važi sledeći odnos između parametara: $0 < \eta^{-} < 1 < \eta^{+}$. Ovi parametri služe za kontrolisanje koraka promene težine. Ukoliko je ovaj korak suviše mali, algoritmu treba dosta vremena da smanji vrednost funkcije greške, dok s druge strane za prevelik korak postoji mogućnost da nikada ne dođe do konvergencije algoritma. Kod RP algoritma se ovaj problem rešava praćenjem promene znaka gradijenta funkcije greške odnosno parcijalnih izvoda koji odgovaraju svakoj od težina. Svaki put kada izvod promeni znak, to znači da je promena suviše velika i da je preskočen lokalni minimum, pa se zato korak promene težine smanjuje za faktor η^{-}. S druge strane, ukoliko se znak izvoda ne menja u dve uzastopne iteracije, korak promene težine se skalira s faktorom η^{+} da bi se ubrzala konvergencija algoritma u delovima gde je funkcija greške ravna. Kako su u originalnom radu date i preporučene vrednosti ovih parametara, za korišćenje algoritma potrebno je kao inicijalne parametre odrediti samo početnu i maksimalnu vrednost veličine koraka promena težina. Zanimljivo je primetiti da loš izbor parametara koje korisnik definiše nema značajan uticaj na performanse algoritma [102].

5.2.2 SCG algoritam

SCG algoritam je izabran kao predstavnik CG (*Conjugate Gradient*) algoritama. Ovi algoritmi su prvenstveno razvijeni za potrebe numeričke optimizacije, a kako se treniranje neuralne mreže može posmatrati kao problem optimizacije bez ograničenja tako su CG algoritmi pronašli primenu u i ovoj oblasti. Za podešavanje težina koristi se izraz

$$\mathbf{w}^{k+1} = \mathbf{w}^{k} - \eta_k \mathbf{d}_k , \tag{5.6}$$

gde je $\mathbf{w}^{k}$ vektor težina u *k*-toj iteraciji, η_k brzina učenja u *k*-toj iteraciji i $\mathbf{d}_k$ smer promene vektora težina u *k*-toj iteraciji. Smer promene težina se dobija na osnovu izraza

$$\mathbf{d}_k = -\nabla \mathbf{E}\left(\mathbf{w}^{k}\right) + \alpha_{k-1}\mathbf{d}_{k-1}, \tag{5.7}$$

gde α_{k-1} definiše formulu koja se koristi za proračun srodnog smera. Iako u dostupnoj literaturi postoji veliki broj formula, uprkos tome što nema svojstvo globalne konvergencije, PR (*Polak-Ribiere*) formula se najčešće koristi jer omogućava efikasnije treniranje mreže od ostalih [17]. PR formula je data sledećim izrazom:

$$\alpha_k^{PR} = \frac{\left(\mathbf{G}_k - \mathbf{G}_{k-1}\right)^T \mathbf{G}_k}{\mathbf{G}_{k-1}^T \mathbf{G}_{k-1}} , \tag{5.8}$$

gde je $\mathbf{G}_k = \nabla \mathbf{E}\left(\mathbf{w}^{k}\right)$ gradijent funkcije greške po težinama u *k*-toj iteraciji.

Glavna razlika između SCG algoritma i ostalih CG algoritama je u načinu na koji se parametar brzine učenja η_k određuje u svakoj od iteracija. Većina CG algoritama za pronalaženje odgovarajuće vrednosti ovog parametara koristi tehnike linijskog pretraživanja [106]. Kako ove tehnike mogu uneti značajnu kompleksnost u algoritam, kod SCG algoritma koristi se drugačiji pristup. Parametar brzine učenja se računa na osnovu izraza

$$\eta_k = \frac{-\mathbf{d}_k^{T} \nabla \mathbf{E}\left(\mathbf{w}^{k}\right)}{\mathbf{d}_k^{T} \mathbf{s}_k} . \tag{5.9}$$

Parametar $\mathbf{s}_k$ u opštem slučaju zahteva proračun *Hessian* matrice $\nabla^2 \mathbf{E}\left(\mathbf{w}^{k}\right)$ koja nosi informaciju o drugom izvodu funkcije greške po težinama neuralne mreže. Kako je ovaj proračun kompleksan, SCG algoritam koristi sledeći izraz kao aproksimaciju $\mathbf{s}_k$ parametra

$$\mathbf{s}_k = \frac{\nabla \mathbf{E}\left(\mathbf{w}^k + \frac{\sigma}{|\mathbf{d}_k|}\mathbf{d}_k\right) - \nabla \mathbf{E}\left(\mathbf{w}^k\right)}{\frac{\sigma}{|\mathbf{d}_k|}} + \lambda_k \mathbf{d}_k \,. \tag{5.10}$$

Parametar σ se mora definisati prilikom korišćenja SCG algoritma, ali kao što je pokazano u literaturi [103] sve dok je njegova vrednost manja od 10^{-4} nije kritičan za performanse algoritma. Nasuprot tome, parametar λ_k se koristi da bi aproksimacija *Hessian* matrice bila pozitivna i konačna matrica. Pokazalo se da ukoliko to nije slučaj, algoritam ne uspeva da adekvatno minimizira funkciju greške [103]. Shodno tome, ukoliko je ispunjen sledeći uslov

$$\mathbf{d}_k{}^T \mathbf{s}_k \le 0, \tag{5.11}$$

neophodno je prilagoditi vrednost parametra λ_k prema sledećoj formuli:

$$\overline{\lambda_k} = 2\left(\lambda_k - \frac{\mathbf{d}_k{}^T \mathbf{s}_k}{|\mathbf{d}_k|^2}\right), \tag{5.12}$$

gde $\overline{\lambda_k}$ predstavlja novu vrednost parametra λ_k u *k*-toj iteraciji algoritma.

Parametar λ_k ima inverzan uticaj na parametar brzine učenja η_k. Korišćenjem ovog mehanizma skaliranja, SCG algoritam izbegava upotrebu vremenski zahtevnih tehnika linijskog pretraživanja i na taj način skraćuje vreme treniranja neuralne mreže.

5.2.3 LM algoritam

LM algoritam je jedan od najbržih algoritama za neuralne mreže koje nemaju više od nekoliko stotina veza [104]. Težine neuralne mreže se podešavaju na osnovu izraza

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \left(\mathbf{J}_k^T \mathbf{J}_k + \mu_k \mathbf{I}\right)^{-1} \mathbf{J}_k^T \mathbf{e}_k \,, \tag{5.13}$$

gde je $\mathbf{I}$ jedinična matrica, μ_k parametar učenja u *k*-toj iteraciji, $\mathbf{J}_k$ *Jacobian* matrica u *k* -toj iteraciji, i $\mathbf{e}_k$ vektor greške u *k*-toj iteraciji**.** *Jacobian* matrica $\mathbf{J}_k$ i vektor greške $\mathbf{e}_k$ definisani su na sledeći način:

$$\mathbf{J}_k = \begin{bmatrix} \frac{\partial e_{11}}{\partial w_1} & \frac{\partial e_{11}}{\partial w_2} & \cdots & \frac{\partial e_{11}}{\partial w_{\tilde{N}}} \\ \frac{\partial e_{12}}{\partial w_1} & \frac{\partial e_{12}}{\partial w_2} & \cdots & \frac{\partial e_{12}}{\partial w_{\tilde{N}}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial e_{1M}}{\partial w_1} & \frac{\partial e_{1M}}{\partial w_2} & \cdots & \frac{\partial e_{1M}}{\partial w_{\tilde{N}}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial e_{P1}}{\partial w_1} & \frac{\partial e_{P1}}{\partial w_2} & \cdots & \frac{\partial e_{P1}}{\partial w_{\tilde{N}}} \\ \frac{\partial e_{P2}}{\partial w_1} & \frac{\partial e_{P2}}{\partial w_2} & \cdots & \frac{\partial e_{P2}}{\partial w_{\tilde{N}}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial e_{PM}}{\partial w_1} & \frac{\partial e_{PM}}{\partial w_2} & \cdots & \frac{\partial e_{PM}}{\partial w_{\tilde{N}}} \end{bmatrix} \quad \mathbf{e}_k = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1M} \\ \vdots \\ e_{P1} \\ e_{P2} \\ \vdots \\ e_{PM} \end{bmatrix}, \tag{5.14}$$

gde je N ukupni broj težina i *bias*-a čvorova. Elementi vektora greške $\mathbf{e}_k$ koji se pojavljuju i u parcijalnim izvodima matrice $\mathbf{J}_k$ definisani su izrazom

$$e_{pm} = z_{mp} - y_{mp}\,. \tag{5.15}$$

U svakoj iteraciji LM algoritam automatski podešava parametar μ_k kako bi se postigla konvergencija. Za velike vrednosti parametra μ_k, LM algoritam se ponaša kao BP algoritam.

5.2.4 ELM algoritam

ELM algoritam se bazira na novoj paradigmi po kojoj nema potrebe za podešavanjem svih težina neuralne mreže. Naime, u radu [105] rigorozno je dokazano da ukoliko je aktivaciona funkcija čvorova skrivenog sloja beskonačno diferencijabilna, onda se ulazne težine i *biasi* skrivenog sloja mogu slučajno izabrati i nije ih potrebno podešavati u toku treninga. Jedna od aktivacionih funkcija koja zadovoljava ovaj uslov je i sigmoidna aktivaciona funkcija [107] koja je korišćena u našem istraživanju, pa se ELM algoritam može koristiti za trening neuralne mreže i u našoj primeni.

Kada se ulazne težine i *biasi* skrivenog sloja slučajno izaberu, neuralna mreža s jednim skrivenim slojem može se smatrati linearnim sistemom jednačina opisanim sa

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{Z}\,, \tag{5.16}$$

gde je $\mathbf{H}$ matrica izlaza skrivenog sloja, $\boldsymbol{\beta}$ matrica izlaznih težina i $\mathbf{Z}$ matrica očekivanih izlaza.

Matrica izlaza skrivenog sloja $\mathbf{H}$ definisana je izrazom:

$$\mathbf{H}(\mathbf{w}_1,\ldots,\mathbf{w}_S,b_1,\ldots,b_S,\mathbf{x}_1,\ldots,\mathbf{x}_P)=\begin{bmatrix} f(\mathbf{w}_1\mathbf{x}_1+b_1) & \ldots & f(\mathbf{w}_S\mathbf{x}_1+b_S) \\ \ldots & \ddots & \ldots \\ f(\mathbf{w}_1\mathbf{x}_P+b_1) & \ldots & f(\mathbf{w}_S\mathbf{x}_P+b_S) \end{bmatrix}, \quad 5.17$$

na taj način da njena *s*-ta kolona odgovara izlazu *s*-tog skrivenog čvora za pojedinačne sekvence $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P$ iz trening skupa. Za razliku od tradicionalnih algoritama gde su sve težine tretirane isto i označavane oznakom $\mathbf{w}$, kod ELM algoritma pod vektorom $\mathbf{w}_s$ podrazumevali smo samo težine između ulaznih čvorova (od prvog do N-tog) i *s*-tog skrivenog čvora:

$$\boldsymbol{w}_s = \left[w_{1s}, w_{2s}, \ldots, w_{NS}\right]^T. \quad 5.18$$

Matrica izlaznih težina $\boldsymbol{\beta}$ definisana je izrazom

$$\boldsymbol{\beta}=\begin{bmatrix} \beta_{11},\beta_{12} & \ldots & \beta_{1M} \\ \ldots & \ddots & \ldots \\ \beta_{S1},\beta_{S2} & \ldots & \beta_{SM} \end{bmatrix}, \quad 5.19$$

gde elementi matrice, β_{sm} predstavljaju težinu između *s*-tog skrivenog čvora i *m*-tog izlaznog čvora.

Matrica očekivanih izlaza $\mathbf{Z}$ je

$$\mathbf{Z}=\begin{bmatrix} z_{11},z_{21} & \ldots & z_{M1} \\ \ldots & \ddots & \ldots \\ z_{1P},z_{2P} & \ldots & z_{MP} \end{bmatrix}. \quad 5.20$$

U slučaju kada je broj skrivenih čvorova *S* jednak broju trening sekvenci *P*, matrica skrivenih izlaza $\mathbf{H}$ je kvadratna i postoji njena inverzna matrica pa, kao što je pokazano u radu [105], neuralna mreža može aproksimirati ove sekvence bez greške. Izlazne težine $\boldsymbol{\beta}$, kao rešenje sistema jednačina (5.16) se u ovom slučaju računaju na osnovu formule

$$\boldsymbol{\beta} = \mathbf{H}^{-1}\mathbf{Z} \quad 5.21$$

Ipak, kako je u većini slučajeva broj skrivenih čvorova mnogo manji od broja trening sekvenci, matrica $\mathbf{H}$ nije kvadratna, te se rešenje sistema jednačina računa na osnovu formule

$$\boldsymbol{\beta} = \mathbf{H}^{+}\mathbf{Z} = \left(\mathbf{H}\mathbf{H}^{\mathbf{T}}\right)^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{Z} \qquad 5.22$$

gde se umesto inverzne matrice koristi *Moore-Penroseova* generalizovana inverzna matrica $\mathbf{H}^{+}$.

Ova matrica karakteristična je po sledećim osobinama [105]:

1. Rešenje $\boldsymbol{\beta} = \mathbf{H}^{+}\mathbf{Z}$ je jedno od rešenja dobijeno metodom najmanjih kvadrata, pa se najmanja greška pri treniranju mreže može postići korišćenjem ovog rešenja

$$\left\|\mathbf{H}\boldsymbol{\beta} - \mathbf{Z}\right\| = \left\|\mathbf{H}\mathbf{H}^{+}\mathbf{Z} - \mathbf{Z}\right\| = \min_{\boldsymbol{\beta}}\left\|\mathbf{H}\boldsymbol{\beta} - \mathbf{Z}\right\|. \qquad 5.23$$

2. Rešenje $\boldsymbol{\beta} = \mathbf{H}^{+}\mathbf{Z}$ ima najmanju normu od svih rešenja linearnog sistema $\mathbf{H}\boldsymbol{\beta} = \mathbf{Z}$ dobijenih metodom najmanjih kvadrata

$$\left\|\boldsymbol{\beta}\right\| = \left\|\mathbf{H}^{+}\mathbf{Z}\right\| \le \left\|\boldsymbol{\beta}\right\|, \ \forall\boldsymbol{\beta}\check{n}\left\{\boldsymbol{\beta} : \left\|\mathbf{H}\boldsymbol{\beta} - \mathbf{Z}\right\| \le \left\|\mathbf{H}\mathbf{q} - \mathbf{Z}\right\|, \forall\mathbf{q}\check{n}\mathbf{R}^{SxP}\right\} \qquad 5.24$$

3. Rešenje linearnog sistema $\mathbf{H}\boldsymbol{\beta} = \mathbf{Z}$ dobijeno metodom najmanjih kvadrata koje ima najmanju normu je jedinstveno i iznosi $\boldsymbol{\beta} = \mathbf{H}^{+}\mathbf{Z}$.

5.3 Sposobnost generalizacije neuralne mreže

Sposobnost generalizacije neuralne mreže kod tradicionalnih algoritama zavisi od trenutka završetka treninga, odnosno do kog nivoa se smanjuje funkcija greške. Suviše dugo treniranje može dovesti do previše istrenirane neuralne mreže koja pruža dobre performanse samo za podatke koji su korišćeni prilikom treniranja.

Kod personalizovanih programskih vodiča na bazi neuralne mreže u literaturi moguće je pronaći dva pristupa:

1. korišćenje kriterijuma ranog zaustavljanja [97] i
2. korišćenje broja epoha kao kriterijuma za prekid treninga [98].

Korišćenje kriterijuma ranog zaustavljanja zahteva da se deo prikupljenih podataka izdvoji za takozvani validacioni skup. Trening se prekida ukoliko funkcija greške proračunata za validacioni skup nastavi da raste u predefinisanom broju iteracija. Na ovaj način se sprečava

dalje smanjenje funkcije greške za trening skup koje može dovesti do povećanja funkcije greške za podatke koji nisu korišćeni pri treningu i poboljšava sposobnost generalizacije neuralne mreže. Treba naglasiti da postoji i varijacija pod nazivom unakrsna validacija koja može poboljšati generalizaciju mreže nauštrb povećanja vremena treniranja. Kod unakrsne validacije skup podataka se deli na predefinisan broj delova od kojih se jedan deo uzima za validacioni skup dok se od ostalih formira trening skup, a zatim neuralna mreža trenira onoliko puta koliko ima delova, vodeći računa o tome da se pri svakom treningu za validacioni skup izabere različit deo podataka.

S druge strane, trajanje treninga neuralne mreže kod tradicionalih algoritama može se ograničiti brojem epoha. Pod završetkom jedne epohe se podrazumeva trenutak kada su svi podaci iz trening skupa po jednom iskorišćeni za trening i nova epoha započinje ponovnim korišćenjem nekog od podataka. Broj epoha nakon kojeg će se trening završiti može se odrediti empirijski na različite načine, pa se tako, na primer, prilikom testiranja sistema može koristiti validaciona procedura, a zatim u produkcionom okruženju kao kriterijum koristiti prosečan broj epoha dobijen tokom faze testiranja [24].

U našem istraživanju smo u slučaju tradicionalnih algoritama učenja koristili kriterijum ranog zaustavljanja s jednim validacionim skupom, jer smatramo da bi korišćenje više validacionih skupova znatno povećalo vreme treniranja na uređajima sa ograničenim resursima, ali ne isključujemo ni mogućnost korišćenja broja epoha na način opisan u radu [24].

Nasuprot ovome, u slučaju kada se koristi ELM algoritam, osobine *Moore-Penroseove* generalizovane inverzne matrice $\mathbf{H}^{+}$ utiču na sposobnost generalizacije neuralne mreže. Korišćenjem ELM algoritma, za slučajno izabrane ulazne težine, postiže se najmanja vrednost norme vektora težina neuralne mreže, pa prema *Bartlettovoj* teoriji generalizacionih performansi *feedforward* neuralnih mreža [108] treba očekivati da ova mreža ima najbolje generalizacione performanse. Konkretno, što *feedforward* neuralna mreža, koja za trening skup postiže niske vrednosti greške, ima manju vrednost norme težina neuralne mreže to će njena generalizaciona sposobnost biti bolja. Dodatno, kako nema potrebe za korišćenjem tehnike ranog zaustavljanja, nema potrebe ni za izdvajanjem dela podataka za validacioni skup pa se i ovi podaci mogu koristiti za treniranje neuralne mreže. Ukoliko oni sadrže podatke o nešto drugačijem ponašanju korisnika u odnosu na dosadašnje to može dovesti do poboljšanja generalizacionih sposobnosti mreže. Činjenica da se podaci ne izdvaju za

validacioni skup može biti od posebne važnosti za primene u kojima je količina podataka ograničena ili samo prikupljanje podataka nije lako realizovati.

Generalizacione sposobnosti neuralne mreže se u opštem slučaju mogu poboljšati i korišćenjem regularizacije čiji je cilj da smanji kompleksnost funkcije koju mreža simulira između svojih ulaza i izlaza – što doprinosi smanjenju verovatnoće učenja specifičnosti trening skupa. Najčešće se realizuje dodavanjem novog uslova, koji sadrži informacije o težinama, u funkciju koju treba optimizovati. U zavisnosti da li se koristi euklidska norma ili apsolutna suma težina razlikujemo l_2 [109] i l_1 regularizaciju [110], respektivno. Podešavanjem parametra regularizacije kontroliše se relativni uticaj težina mreže i proračunate greške na sam proces učenja.

5.4 Disbalans klasa

Neravnomerna raspodela podataka između klasa, ili takozvani disbalans klasa, može značajno degradirati performanse sistema, te spada među deset najozbiljnijih problema koji se javljaju kod sistema na bazi mašinskog učenja [11]. Usled tendencije gledalaca digitalne televizije da mnogo češće pružaju informacije o sadržajima koje vole da gledaju u odnosu na one koje ne žele da gledaju [42], [111], jasno je da problem disbalansa klasa postoji i u našoj primeni.

Mogući razlog za pojavu ovakve raspodele je taj što gledaoci najčešće ili prate manji broj njima omiljenih TV kanala, ili koriste personalizovani programski vodič s dobrim performansama, te je verovatnoća prikupljanja informacija o sadržaju koji ne vole veoma mala. Ovaj problem, je po našem mišljenju, izraženiji kada se preporučivači koriste, jer ukoliko sistem funkcioniše kako treba, korisnik će u većini slučajeva prihvatiti i pozitivno oceniti preporučen sadržaj, dok će samo u retkim slučajevima ocena biti negativna.

Štaviše ponašanje pojedinih korisnika, bilo u ranoj fazi prikupljanja podataka ili tokom celokupnog korišćenja sistema, može dovesti do ekstremnog slučaja kada prikupljeni podaci sadrže informacije samo o TV sadržajima koje korisnik voli da gleda, te sistem mora biti u mogućnosti da pruži pouzdane preporuke i ovim korisnicima.

U daljem tekstu najpre ćemo razmotriti mehanizam delovanja disbalansa klasa, a zatim i potencijalna rešenja ovog problema.

5.4.1 Mehanizam delovanja disbalansa klasa

Pretpostavimo da se 99% prikupljenih podataka odnosi na sadržaje koje korisnik voli da gleda, a da se preostalih 1% odnosi na sadržaje koje korisnik ne voli da gleda. Pod ovakvim uslovima neuralna mreža koja ima za cilj da minimizira grešku klasifikacije, zbog znatno većeg uticaja sadržaja koje korisnik voli da gleda na funkciju greške, biće loša u klasifikaciji sadržaja koje korisnik ne želi da gleda. Česta pogrešna klasifikacija ovih sadržaja može dovesti do značajnog smanjenja poverenja korisnika u sistem jer će se oni usled greške u klasifikaciji pojavljivati u listi preporuka.

S druge strane, ukoliko se prikupljeni podaci odnose samo na sadržaje koje korisnik voli da gleda, problem pružanja preporuka se svodi na problem binarne klasifikacije na osnovu informacija o samo jednoj klasi tzv. *one-class* klasifikacije [112], te je potrebno prilagoditi arhitekturu i algoritme učenja neuralne mreže.

5.4.2 Metode za borbu protiv problema disbalansa klasa

Metode za borbu protiv problema disbalansa klasa mogu se grubo podeliti na one koje se primenjuju na nivou podataka i one koje se primenjuju na algoritamskom nivou [29]. Metode koje se primenjuju na nivou podataka mogu generisati nove podatke u okviru klase s manjim brojem podataka ili odbaciti podatke iz klase s većim brojem. U svom osnovnom obliku, metode koje generišu nove podatke, u stvari, na slučajan način biraju već postojeće podatke u klasi s manjim brojem podataka koje će duplirati u trening skupu. Ovo može dovesti do toga da sistem nauči konkretno te podatke, pa će imati lošiju sposobnost generalizacije za klasu s manjim brojem podataka. Naprednije metode koje generišu nove podatke, kao što je, na primer, SMOTE (*Synthetic Minority Oversampling Technique*) [113], najpre određuju najbliže susede posmatrane sekvence koji pripadaju klasi s manjim brojem podataka, potom slučajno biraju jednog od njih, te generišu novu trening sekvencu koja se u vektorskom prostoru nalazi između posmatrane i izabranog suseda. Iako je ovaj pristup pokazao dobre performanse u velikom broju primena, smatramo da nije pogodan za korišćenje kod personalizovanih programskih vodiča, jer ga zbog malog broja prikupljenih korisničkih interakcija nije moguće primeniti od početka rada sistema. Nasuprot ovome, metode koje odbacuju podatke iz klase s većim brojem podataka mogu dovesti do gubitka podataka koji su značajni za učenje korisničkih interesovanja, pa se iz tog razloga nismo odlučili ni za njih.

Odgovarajuću metodu za borbu protiv problema disbalansa klasa, u našem istraživanju potražili smo u okviru algoritamskih rešenja jer je i empirijski pokazano da se njihovim

korišćenjem u slučaju sistema na bazi neuralnih mreža mogu postići bolji rezultati nego kada se koriste metode koje se primenjuju na nivou podataka [114]. Algoritamska rešenja uvode dodatne težinske faktore koji se mogu koristiti za modifikaciju [115]:

1. verovatnoće vrednosti izlaza neuralne mreže,
2. očekivane vrednosti izlaza neuralne mreže,
3. parametra brzine učenja, odnosno samog algoritma učenja, ili
4. greške koja se minimizira,

kako bi izjednačila uticaj pojedinačnih klasa na sam proces mašinskog učenja. Konkretno rešenje koje smo izabrali biće opisano u poglavlju u kojem je dat predlog dizajna našeg personalizovanog programskog vodiča za digitalnu televiziju.

S druge strane, u literaturi koja se tiče preporučivača ne postoji veliki broj radova koji se bave istraživanjem problema binarne klasifikacije na osnovu informacija o samo jednoj klasi. Kao reprezentativan primer može se navesti korišćenje *one-class* SVM algoritma [116]. Imajući u vidu da vreme potrebno za treniranje SVM algoritma, po našem mišljenju, ne može da zadovolji zahtev za učenje korisničkih interesovanja i pružanje preporuka u realnom vremenu, proširili smo dalje teorijsko istraživanje na srodne oblasti.

U istraživanju [117] upoređene su performanse velikog broja klasifikatora za problem klasifikacije dokumenata na Internetu, pri čemu se posebna vrsta neuralnih mreža, tzv. autoenkoder pokazao posebno pogodnim, nadmašajući i performanse *one-class* SVM algoritma [116].

Autoenkoder neuralna mreža se odlikuje tokom podataka u jednom smeru, pri čemu su izlazi mreže jednaki ulazima u nju [93]. Arhitektura autoenkoder neuralne mreže s jednim skrivenim slojem i brojem skrivenih čvorova koji je manji od dimenzija ulaznog vektorskog prostora posebno je pogodna za korišćenje kod problema *one-class* klasifikacije [117]. Korišćenjem ovakve arhitekture, sprečava se situacija u kojoj se skriveni sloj ponaša kao jedinična matrica i neuralna mreža može da nauči generalne karakteristike TV sadržaja koje korisnik voli da gleda. Na osnovu greške na izlazu neuralne mreže:

$$RMSE = \sqrt{\sum_{p=1}^{P}\sum_{m=1}^{M}\left(z_{mp} - y_{mp}\right)^2}\,, \qquad 5.25$$

estimira se da li ponuđeni sadržaj, opisan vektorom na ulazu mreže, treba dodeliti klasi sadržaja koje korisnik želi da gleda ili klasi sadržaja koje korisnik ne želi da gleda. Ukoliko je greška manja od unapred definisanog praga, ponuđeni sadržaj se klasifikuje u klasu sadržaja koje korisnik voli da gleda, dok se u suprotnom smatra da pripada klasi sadžaja koje korisnik ne voli da gleda.

Pored arhitekture mreže, u slučaju kada ne postoje informacije o sadržajima koje korisnik ne voli da gleda neophodno je modifikovati i način na koji se biraju optimalni parametri sistema. Korišćenjem metoda doslednosti (*consistency-based method*) [118] moguće je izabrati parametre sistema samo na osnovu informacija o jednoj klasi [117]. Kako bi se postigle dobre generalizacione sposobnosti neuralne mreže, u slučaju kada postoje informacije samo o klasi sadržaja koje korisnik voli da gleda, parametri sistema se biraju na taj način da određeni broj sadržaja iz trening skupa bude dodeljen klasi sadržaja koje korisnik ne voli da gleda. Za ove sadržaje smatra se da odstupaju od dominantnih korisničkih interesovanja i oni se koriste za simuliranje postojanja klase sadržaja koje korisnik ne voli da gleda. Procenat sadržaja iz trening skupa koji će se namerno pogrešno klasifikovati, p_e, određuje se empirijski i predstavlja jedini parametar metoda doslednosti. Kako veliki broj parametara sistema može dovesti do ovakve klasifikacije sadržaja, cilj metoda doslednosti je izabrati najkompleksniji klasifikator (s najmanjom greškom u toku treninga) koji je i dalje pouzdan (ima dobre generalizacione sposobnosti za klasu sadržaja koje korisnik voli da gleda). Kompleksnost klasifikatora je određena brojem skrivenih čvorova *S*, [112], a klasifikator se smatra pouzdanim ukoliko zadovoljava sledeći uslov [118]

$$p_v < p_e + 2\sqrt{N_v p_e (1-p_e)}, \qquad 5.26$$

gde je p_v procenat sadržaja iz validacionog skupa koje korisnik voli da gleda, a koje je klasifikator dodelio klasi sadržaja koje korisnik ne voli da gleda, dok je N_v ukupan broj korisničkih interakcija u validacionom skupu. Ukoliko, nakon izbora broja skrivenih čvorova, više vrednosti regularizacionih parametara ispunjava uslov pouzdanog klasifikatora, koristićemo najveću.

5.5 Rezime

U ovom poglavlju opisali smo:

- karakteristike neuralnih mreža koje se koriste u preporučivačima,

- njihovu sposobnost generalizacije,
- dostupne arhitekture i algoritme učenja,
- kao i uticaj disbalansa klasa na njihove performanse i moguća rešenja ovog problema.

U sledećem poglavlju ćemo razmotriti pregled aktuelnog stanja u oblasti, uporediti i diskutovati rezultate reprezentativnih primera preporučivača.

6. Pregled aktuelnog stanja u oblasti

Imajući u vidu veliki broj raznovrsnih pravaca istraživanja koji postoje u oblasti personalizovanih programskih vodiča, nije lako pronaći radove koji bi opisali aktuelno stanje. Iz tog razloga, u ovom poglavlju ćemo predstaviti tri rada koji su, po nama, relevantni za različite aspekte razmatrane prilikom projektovanja i diskutovati kako se njihovi pristupi razlikuju od onog za koji smo se mi odlučili.

U radu [9] predstavljen je preporučivač filmova koji je, za razliku od velikog broja sistema u literaturi, u potpunosti implementiran lokalno na korisničkom uređaju. Autori su se za ovakav način implementacije odlučili usled eksplicitnog zahteva proizvođača opreme za kojeg je ovaj sistem projektovan. Kako nije moguće prikupiti informacije o interesovanjima ostalih korisnika, osim onoga koji koristi uređaj, preporuke su formirane korišćenjem tehnike filtriranja sadržaja. Kao glavna prednost lokalne implementacije sistema ističe se zaštita privatnosti korisnika.

Posebna pažnja posvećena je izboru algoritma učenja jer, kao što je naglašeno u samom radu, usled ograničenih hardverskih resursa nije moguće koristiti kompleksnije varijante. Stoga su se autori odlučili za heuristički pristup problemu pružanja preporuka i koristili kosinusnu sličnost kako bi odredili u kojoj meri posmatrani film odgovara korisničkim interesovanjima.

Za svaki od dostupnih filmova, korišćenjem *The Movie Database* repozitorijuma, prikupljene su informacije o njihovom žanru, scenaristima, režiserima, godini i dekadi objavljivanja, žanrovima, ključnim rečima koji opisuju radnju filma, glumcima, producentima i studiju u kome je sniman. Ipak, u konkretnoj implementaciji korišćeni su samo podaci o žanrovima i dekadi kojoj film pripada, jer je procenjeno da upotreba preostalih informacija, kao što su one o glumcima, zbog njihovog broja (reda veličine nekoliko hiljada), prevazilazi mogućnosti hardverskih resursa korisničkih uređaja [9]. Ukupna sličnost između filmova se računa kao ponderisana suma sličnosti između njihovih žanrova i između dekada u kojoj su oni snimljeni, pri čemu je prvima dodeljen znatno veći težinski faktor u iznosu od 0.8, a drugima samo 0.2.

Na početku korišćenja sistema, inicijalni profil korisnika formira se na osnovu ankete o omiljenim žanrovima, koju svaki korisnik mora da popuni. Na ovaj način su autori pokušali da reše problem hladnog starta i prikupe što više podataka na početku.

Podaci o ponašanju korisnika prikupljaju se u toku rada sistema i eksplicitno i implicitno, pri čemu je vodič projektovan tako da preferira direktnu interakciju korisnika. Od njih se očekuje da nakon pogledanog filma ocene sadržaj s vrednošću između 1 i 5, ali ukoliko korisnik to ne učini, preporučivač će implicitno dodeliti ocenu filmu koristeći sledeću ugrađenu logiku. Kada je korisnik odgledao film do kraja, dodeliće mu se ocena 4, dok ukoliko to nije uradio ni u nekoliko narednih dana od početka gledanja, pridružiće mu se vrednost ocene 2. Dodatno, ocena 4 se dodeljuje i omiljenim žanrovima koji su izabrani prilikom formiranja inicijalnog profila. U slučajevima kada postoji i implicitna i eksplicitna ocena – kada je korisnik najpre odgledao film i nije ga ocenio, a zatim ponovo odgledao i dao ocenu, kao validna se uzima ocena dobijena eksplicitnim putem. S druge strane, ako je film ocenjen nekoliko puta s različitim ocenama usvaja se poslednja.

Profil korisnika, u smislu omiljenih žanrova i dekada filmova, ažurira se dodavanjem ocena novo odgledanih filmova, na svaki od posmatranih aspekata, pri čemu se u zavisnosti od broja prikupljenih interakcija različiti težinski faktori pridružuju trenutnom profilu i oceni. Autori su posmatrali slučajeve kada je korisnik odgledao:

1. između 0 i 5 filmova,
2. između 6 i 49 filmova,
3. i više od 50 filmova.

U fazi hladnog starta (do 5 ocenjenih filmova), prikupljenim ocenama je dodeljen znatno veći značaj od profila, dok se s porastom broja prikupljenih interakcija smatra da trenutni profil bolje opisuje interesovanja korisnika, te je uticaj pojedinačne ocene manji.

U ovome radu, konkretno nije usvojena ni jedna mera performansi, ali je funkcionisanje sistema opisano na primeru jednog korisnika kroz preporuke koje on dobija ukoliko oceni unapred izabrane sadržaje s posmatranom ocenom. Iako je sistem prvenstveno projektovan za filmove koji se nalaze lokalno na korisničkom uređaju ili eksternoj memoriji i u tu svrhu razvijen deo koji dodeljuje opise ovim sadržajima, jasno je naglašeno da se može koristiti i za ostale multimedijalne sadržaje dostupne putem televizije ili Interneta.

Premda su u radu [9] razmatrane veoma važne osobine personalizovanih programskih vodiča, kao što su mogućnost rada na uređajima s ograničenim hardverskim resursima i zaštita privatnosti; uticaj kontekstualnih informacija i disbalansa klasa nije uzet u obzir. Takođe, ni naprednija arhitektura sistema, koja može modelirati različite načine na koje korisnici donose odluke, kao što je na primer ona koja koristi neuralne mreže, nije uzeta u razmatranje. U ovoj doktorskoj disertaciji posvetili smo značajnu pažnja ovim aspektima.

Na kraju, treba spomenuti da iako autori rada [9], objavljenog 2015. godine, naglašavaju da je njihova implementacija jedna od prvih koja je prilagođena radu na mobilnim uređajima, deo našeg istraživanja koji se tiče ove tematike predstavljen je široj javnosti već 2012. godine.

S druge strane, u radu [12], koji smo odlučili da predstavimo u ovom poglavlju, glavni fokus je na primeni neuralnih mreža u preporučivačima za digitalnu televiziju i korišćenju kontekstualnih informacija.

Konkretno, predložen je hibridni sistem kod kojeg se na osnovu informacija o omiljenim aktivnostima i generalnih interesovanja korisnika, njihovog raspoloženja, podataka o TV sadržajima kojima je pristupao u prošlosti i demografskih informacija formira grupni profil korisnika.

Na početku korišćenja personalizovanog programskog vodiča, podaci o omiljenim žanrovima, aktivnostima i interesovanjima, kao i demografski podaci se eksplicitno prikupljaju uz pomoć upitnika, a zatim se na osnovu njih, korišćenjem *K-mean* klasterizacije, detektuju grupe korisnika i određuje pripadnost grupi. Aktivnosti i opšta interesovanja korisnika izabrani su kako bi bolje opisali njihov životni stil za koji autori smatraju da značajno pomaže u određivanju sličnih korisnika.

Dostupni sadržaji predstavljeni su vektorom odlika čije kordinate opisuju pripadnost sledećim žanrovima: *Education*, *Drama*, *Shopping*, *Entertainment*, *Sport/Healthcare*, *Cartoons*, *Fashion* i *News*. Vrednost odgovarajuće koordinate iznosi 1 u slučajevima kada sadržaj pripada posmatranom žanru, dok je u suprotnom jednaka 0. Na osnovu opisa TV sadržaja, raspoloženja i ocena korisnika iz iste grupe, neuralna mreža procenjuje koje bi sadržaje oni voleli da gledaju, a koje ne. Korisnik klikom na posebno dugme modifikovanog daljinskog upravljača pruža informaciju o tome da li je u posmatranom trenutku srećan, nesrećan ili mu je dosadno, dok TV sadržaje uz pomoć istog interfejsa eksplicitno ocenjuje na skali od 1 do 5.

Kao algoritam učenja usvojen je *backpropagation* u svom osnovnom obliku, jer se on, kao što je naglašeno u radu [12], najčeće primenjuje kod neuralnih mreža. Kako bi pronašli optimalnu arhitekturu, u ovome istraživanju isprobali su verzije sistema s različitim brojem skrivenih slojeva, ali nažalost podatak o broju skrivenih slojeva koji su korišćeni u svakom od njih nije prikazan. Performanse različitih verzija upoređene su, u smislu tačnosti klasifikacije i korena srednje kvadratne greške, za slučajeve kada se koriste kontekstualne informacije o raspoloženju i kada ih nema, respektivno.

Dobijeni rezultati pokazuju da, bez obzira koja se od razmatranih metrika koristi, personalizovani programski vodič koji koristi kontekstualne informacije o raspoloženju korisnika postiže bolje performanse.

Ukoliko se kao mera performansi koristi koren srednje kvadratne vrednosti, kao optimalna izabrana je arhitektura s 10 skrivenih slojeva, dok je u slučaju tačnosti klasifikacije najbolja ona s 8. Iako su autori, najavili da će u budućim istraživanjima prilagoditi njihov sistem uslovima mobilne televizije, zbog velikog broja skrivenih slojeva i neadekvatnog izbora algoritma učenja, smatramo da ga bez značajnijih modifikacija nije moguće koristiti u okruženju s ograničenim hardverskim resursima.

Nasuprot ovome, disbalans klasa, niti metode za borbu s ovim problemom nisu razmatrane ni u ovome radu, iako raspodela korišćenih podataka, prikazana u tabeli 6.1, potvrđuje tendenciju korisnika da znatno ređe pružaju informacije o TV sadržajima koje ne vole da gledaju.

Tabela 6.1. Raspodela ocena korisnika [12]

Ocena	1	2	3	4	5
Procenat od ukupnog broja ocena (%)	0.003	0.015	33.74	50.66	13.69

Premda, goreopisani sistem koristi kontekstualne informacije njihov izbor i način prikupljanja nije detaljnije diskutovan niti istražen. Razlozi zašto su baš koren srednje kvadratne greške i tačnost klasifikacije izabrane kao mere performansi, nisu navedeni u radu [12].

Iako je u prethodna dva rada razmotren veliki broj aspekata koje bi po našem mišljenju trebalo uzeti u obzir prilikom projektovanja personalizovanog programskog vodiča, uticaj disbalansa klasa nije posmatran ni u jednom od njih. Ovaj problem detaljno je istražen u radu [27], opisanom u nastavku teksta.

Sistem na bazi faktorizacije matrica, koji je predložen, osmišljen je tako da može da funkcioniše u okruženjima gde je raspodela implicitno prikupljenih podatke o korisničkim interesovanjima takva da dovodi do problema disbalansa klasa. Kao što su autori jasno naglasili, očekivano je da u većini slučajeva preporučivači rade baš u ovim uslovima.

Kako bi pružili preporuke čak i korisnicima sa specifičnim interesovanjima, pored matrice malog ranga, koja se, kao što je u poglavlju 2 doktorske disertacije napisano, uobičajeno koristi za aproksimaciju originalne matrice ocena i proračun latentnih faktora, formira se i matrica s malim brojem nenultih vrednosti koja detektuje baš ovakva interesovanja.

Problem pronalaženja latentnih faktora iz sume ovih dveju matrica se zatim rešava korišćenjem algoritma na bazi proračuna gradijenta funkcije greške koja je modifikovana tako da u obzir uzima i nejednaku količinu podataka u pojedinačnim klasama.

Različite verzije CSRR (*Robust Cost-Sensitive Learning for Recommendation*) algoritma izvedene su za slučajeve kada se:

- maksimizira ponderisana suma odziva i tačnosti klasifikacije sadržaja koje korisnik ne želi da gleda,
- minimizira ukupna cena grešaka prilikom klasifikacije dobijena kao ponderisana suma broja sadržaja za koje je greškom proglašeno da se sviđaju korisniku i broja sadržaja za koje je greškom proglašeno da mu se ne sviđaju.

Kako bi se poboljšala sposobnost generalizacije sistema, koristi se i regularizacija. Podešavanjem regularizacionog parametra utiče se na to koliko će sistem biti prilagođen specifičnim interesovanjima pojedinaca, a koliko zajedničkim interesovanjima sličnih korisnika.

Iako smo u dosadašnjem tekstu doktorske disertacije, uvek isticali da negativno ocenjeni sadržaji pripadaju klasi s manjom količinom podataka, to nije slučaj u ovome istraživanju. Autori rada [27] posmatraju matricu ocena, u prostoru korisnik – sadržaj, kod koje su s 1 označeni sadržaji koji im se sviđaju, dok su oni koji im se ne sviđaju i kojima nije pristupano označeni s 0. Imajući u vidu činjenicu da je jedna od glavnih pretpostavki kod problema pružanja preporuka da korisnik u realnom vremenu ne može da pretraži dostupne sadržaje, ne čudi da disbalans klasa kod kolaborativnog filtriranja s ovakvim načinom označavanja dobija drugačiju dimenziju. Klasa s manjom količinom podataka, u ovom slučaju formira se od sadržaja koje je korisnik pozitivno ocenio.

Performanse opisanog preporučivača ispitane su korišćenjem različitih skupova podataka koji sadrže informacije o ocenama filmovima. Kako su korisnici prilikom iskazivanja svog mišljenja u ovim skupovima koristili diskretnu skalu od 1 do 5, način interakcije prilagođen je binarnom sistemu korišćenom u radu [27]. Ukoliko je vrednost ocene veća od 3, smatralo se da se film sviđa korisniku, i obrnuto ukoliko nije, da mu se ne sviđa.

Kao mere performansi korišćene su preciznost, odziv, F1 i NDCG za liste preporuka s 5, 10 i 15 filmova, respektivno.

Predloženi sistem upoređen je s personalizovanim programskim vodičima:

- koji preporučuju najpopularnije filmove,
- koji koriste WRMF (*Weighted Regularized Matrix Factorization*) s težinskim faktorima prilagođenim nivoima pouzdanosti ocene korisnika,
- koji koriste BPRMF (*Bayesian Personalized Regularized Matrix Factorization*), algoritam projektovan tako da direktno rangira dostupne sadržaje i
- koji koriste MC-Shift algoritam, karakterističan po tome da uči o korisničkim interesovanjima samo na osnovu pozitivnih interakcija.

Dobijeni rezultati pokazuju značajno poboljšanje performansi u odnosu na sisteme koji koriste algoritme kod kojih se, bez obzira kojoj klasi sadržaj pripada, greške tretiraju podjednako. U slučaju kada se koristi *MovieLens-100K* skup podataka, CSRR algoritam u odnosu na algoritam koji preporučuje najpopularnije filmove, BPRMF, i MC-Shift postiže 84.5%, 17.2% i 4.03% veću vrednost NDCG metrike za listu od 5 preporuka. S druge strane, pod istim uslovima dobijena vrednost za WRMF algoritam koji koristi težinske faktore u procesu učenja je za 3.46 % manja od predloženog. Sličan odnos postiže se i za ostale razmatrane mere performansi i skupove podataka, i drugu verziju predloženog algoritma.

Iako WRMF dodeljuje različite težinske faktore korisničkim interakcijama, razlog zašto se oni koriste nije isti kao kod CSRR algoritma. Prvi ih koristi kako bi prilagodio sistem različitom nivou pouzdanosti implicitno prikupljenih ocena, dok se kod drugog primenjuju u cilju prevazilaženja problema disbalansa klasa.

Odgovarajućim podešavanjem težinskih faktora prema greški koja se minimizira ili sumi koja se maksimizira, postiže se pomeranje granice odlučivanja i povećava uticaj sadržaja koje korisnik voli da gleda na proces pružanja preporuka.

Kao što se može primetiti, shvatanje uticaja disbalansa klasa u radu [27], razlikuje se od naših stavova. Autori ovoga rada smatraju da će ukoliko personalizovani programski vodič preskoči da preporuči neki od sadržaja koje korisnik želi da gleda, to imati znatno veći uticaj na degradaciju njegovog zadovoljstva nego ukoliko se neki od sadržaja koje ne voli da gleda pojavi u listi preporuka. Premda, mi podržavamo suprotnu hipotezu, jer smatramo da je grešku prilikom klasifikacije sadržaja koje korisnik ne želi da gleda lakše primetiti, treba naglasiti da sistem mora biti dovoljno dobar u predikciji obeju klasa, a da usled uticaja disbalansa klasa treba primeniti fina podešavanja.

6.1 Rezime

U ovome poglavlju predstavili smo pregled aktuelnog stanja u oblasti, uporedili i diskutovali razlike i sličnosti s pristupom koji smo mi koristili pri projektovanju personalizovanog programskog vodiča za digitalnu televiziju. U sledećem ćemo opisati konkretan predlog sistema i predstaviti rezultate eksperimenata koji su doveli do njega.

7. Opis predloženog sistema

Kako bi došli do konačnog predloga našeg personalizovanog programskog vodiča za digitalnu televiziju sprovešćemo veliki broj eksperimentalnih i teorijskih istraživanja koja se tiču:

- zaštite privatnosti korisnika,
- mogućnosti rada sistema na uređajima s ograničenim resursima,
- izbora i načina korišćenja kontekstualnih informacija,
- problema disbalansa klasa i
- izbora mera performansi.

Problem pružanja preporuka, u svim istraživanjima, posmatraćemo kao problem klasifikacije ponuđenih TV sadržaja na one koji se korisniku sviđaju i na one koji mu se ne sviđaju. Smatramo da prikupljanje podataka koji su potrebni za estimiranje tačne ocene TV sadržaja može ometati uobičajeni način gledanja televizije, te nije pogodno koristiti sisteme koji funkcionišu na ovaj način, ali i da lista preporuka ne treba biti previše dugačka, te nije pogodno koristiti ni sisteme koji ponuđene sadržaje rangiraju u listi. Dodatno, ovakvim izborom cilja sistema postiže se određena fleksibilnost sistema jer sam pružalac usluge distribucije medijskog sadržaja može definisati pravila na osnovu kojih će se iz klase sadržaja koje korisnik voli da gleda formirati lista preporuka.

U cilju ispitivanja različitih aspekata sistema koristićemo tri skupa podataka.

MovieLens skup podataka o pogledanim filmovima prikupljen je u okviru *GroupLens Research* projekta na Univerzitetu u Minesoti [119]. Vektor odlika kojim su opisani filmovi u ovom skupu određen je s sledećih 18 koordinata: *action*, *adventure*, *animation*, *children's*, *comedy*, *crime*, *documentary*, *drama*, *fantasy*, *film noir*, *horror*, *musical*, *mystery*, *romance*, *Sci-Fi*, *thriller*, *war* i *sport*, koje imaju vrednost 1 ukoliko film pripada posmatranom žanru, odnosno 0 ukoliko to nije slučaj. Od 6040 gledalaca sa 1000298 ocena za 3900 filmova kreirali smo slučajni uzorak s 45 gledalaca koji su ukupno generisali 20140 interakcija, i njega koristili u našim istraživanjima. Ocene od 1 do 5, koje su korišćenje u ovom skupu, prilagodili smo načinu interakcije opisanom u radu [120]. Ukoliko je korisnik ocenio film sa "1" ili "2" smatrali smo da mu se film ne sviđa, dok ukoliko ga je ocenio sa "3", "4" ili "5" smatrali smo da mu se sviđa i da bi voleo da gleda slične sadržaje. Iako se za filmove koji su ocenjeni sa "3" može smatrati da je korisniku svejedno, ili čak da mu se ne sviđaju, ovakav način

grupisanja izabran je jer smanjuje potrebu za direktnim učešćem korisnika pri interakciji sa sistemom [42].

Mana ovog skupa podataka je to što ne sadrži informacije o kontekstu u kome je pristupano filmovima, niti informacije o preostalim sadržajima (npr. serijama, vestima, emisijama) koji se mogu pronaći na televiziji, te ga nećemo koristiti u eksperimentima koji zahtevaju ove podatke. Takođe, nećemo ga koristiti ni u eksperimentima koji se tiču disbalansa klasa jer mera u kojoj će se ovaj problem javiti može u mnogome zavisiti od slučajnog izbora korisnika iz skupa.

ETF skup podataka formirali smo od TV sadržaja kojima su studenti Elektrotehničkog fakulteta u Beogradu pristupali u periodu od aprila do juna 2012. godine. Ovaj skup podataka smo sami prikupili za potrebe istraživanja jer u to vreme nije postojao nijedan reprezentativan skup kojim bi bilo opisano ponašanje gledalaca televizije, a koji je sadržao kontekstualne informacije. Dostupne TV sadržaje predstavili smo 24 dimenzionalnim vektorom odlika čije koordinate odgovaraju posmatranim žanrovima: *action, adventure, animation, children's, comedy, crime, documentary, drama, fantasy, film noir, horror, musical (movie), mystery, romance, Sci-Fi, thriller, war, western, fun, music* (*show*), *talk show, lifestyle, news/info* i *sport*, koji smo formirali na isti način kao kod eksperimenata koji koriste *MovieLens skup podataka*. Pored opisa TV sadržaja, ovaj skup sadrži i informacije o danu u nedelji (radni dan, vikend ili praznik) kada im je korisnik pristupao i vremenu njihovog emitovanja (jutro, podne, poslepodne, veče ili noć). Ovi nivoi granularnosti izabrani su jer sam smatrao da se na ovaj način može prikupiti dovoljno podataka za svaki od konteksta. Usvojili smo način interakcije sa sistemom iz rada [42], kod kojeg se podaci o sadržajima koje korisnik voli da gleda prikupljaju implicitno, a podaci o onima koje ne voli da gleda eksplicitno, jer se na ovaj način prikupljaju pouzdani podaci i ne ometa uobičajeni način gledanja televizije. Ukupno je prikupljeno 1219 interakcija, i kao što se može videti iz tabele 7.1, u kojoj je za svakog od korisnika prikazan broj pozitivnih i negativnih interakcija, postoji problem disbalansa klasa.

Prednost ETF skupa podataka, u odnosu na *MovieLens* skup, je ta što pored filmova sadrži podatke i o ostalim TV sadržajima, kao i to što je pogodan za istraživanja koja se tiču uticaja konteksta i disbalansa klasa. Ipak, kako su kontekstualne informacije ograničene na dan i vreme kada je korisnik pristupao sadržaju, za opsežnija istraživanja konteksta pogodno je koristiti skupove podataka koji sadrže veći broj informacija ovog tipa.

Tabela 7.1. Karakteristike prikupljenih podataka

Korisnik	Broj pozitivnih interakcija	Broj negativnih interakcija
1	30	1
2	30	1
3	30	1
4	62	0
5	30	1
6	57	5
7	62	0
8	30	1
9	90	3
10	24	7
11	30	1
12	31	0
13	28	3
14	29	2
15	31	0
16	60	2
17	61	1
18	57	5
19	30	1
20	91	2
21	17	5
22	28	2
23	25	5
24	27	4
25	22	2
26	29	1
27	30	1
28	21	10
29	28	1
30	29	2

U tu svrhu koristili smo *LDOS – CoMoDa skup podataka* o odgledanim filmovima prikupljen na Elektrotehničkom fakultetu, Univerziteta u Ljubljani [121]. Dostupni filmovi, u njemu, opisani su 25-dimenzionalnim vektorima žanrova čije koordinate imaju sledeća značenja: *Action, Adult, Adventure, Animation, Art, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Short, Sport, Thriller, War* i *Western*, respektivno. Pored opisa filmova, ovaj skup sadrži i veliki broj kontekstualnih informacija, koje smo naveli u tabeli 7.2.

Kako su pri prikupljanju *LDOS – CoMoDa* podataka, korisnici eksplicitno ocenjivali filmove numeričkim ocenama od “1” do “5” na isti goreopisani način, kao i kod istraživanja kod kojih

smo koristili *MovieLens podatke*, prilagodićemo način interakcije. Za naše potrebe iz celokupnog skupa, korišćenjem metode slučajnog uzorka, izdvojili smo podatke o 20 korisnika s ukupno 965 interakcija.

Tabela 7.2. Kontekstualne informacije

Kontekstualne informacije	**Moguće vrednosti**
Vreme	Jutro, poslepodne, veče, noć
Dan	Radni dan, vikend, praznik
Godišnje doba	Proleće, leto, jesen, zima
Lokacija	Kuća, javno mesto, Prijateljeva kuća
Vremenski uslovi	Sunčano, kišovito, pljusak, sneg, oblačno
Društvo	Sam, partner, prijatelj, kolega, roditelji, nepoznati ljudi, porodica
Emocija nakon gledanja programa	Tužan, srećan, uplašen, iznenađen, besan, zgrožen, neutralno
Dominantna emocija	Tužan, srećan, uplašen, iznenađen, besan, zgrožen, neutralno
Raspoloženje	Pozitivno, neutralno, negativno
Fizičko stanje	Zdrav, bolestan
Odluka	Naša odluka, tuđa odluka
Interakcija	Prva interakcija, ostale interakcije

Najveća prednost ovog skupa u odnosu na podatke koje smo mi prikupili je to što sadrži veliki broj raznovrsnih kontekstualnih infromacija, dok je njegova mana to što je ograničen samo na podatke o filmovima.

U daljem tekstu ispitaćemo redom jedan po jedan aspekte sistema koji smo uzeli u obzir prilikom projektovanja. Sve simulacije realizovali smo u Matlabu na desktop računaru s Intel(R) Core(TM)2 T5250 1.5GHz procesorom i 1.5 GB RAM memorije.

7.1 Zaštita privatnosti

Veoma je važno da se umesto dosadašnje prakse dodavanja funkcionalnosti zaštite privatnosti u već isprojektovane sisteme, ona uzme u obzir već u fazi projektovanja personalizovanog programskog vodiča. Od ovakvog pristupa zaštiti privatnosti očekuje se da postane dominantan, jer GDPR (*General Data Protection Regulation*) uredba o zaštiti privatnosti donešena na nivou Evropske Unije, predstavlja zakonsku implementaciju ovog principa [10].

U skladu s važećim ZZPL [55], predlažemo da se u okviru ugovora koji korisnik potpisuje s pružaocem usluge distribucije medijskog sadržaja ubace odredbe koje bi se odnosile na

pristanak korisnika, svrhu i način obrade podataka za potrebe personalizovanog programskog vodiča. U okviru ugovora treba ubaciti i obaveštenje o obradi podataka, vremenski period koliko se podaci čuvaju, kao i prava korisnika i način žalbe. Na ovaj način bi bila ispoštovana načela zakonitosti, ograničenosti svrhe, transparentnosti obrade i ograničenog zadržavanja iz ZZPL Republike Srbije.

Zaštita privatnosti u našim radovima [122]–[124] koji su nastali kao rezultat istraživanja u okviru doktorskih studija, bazira se na prebacivanju celog procesa pružanja preporuka na korisničke uređaje. Jedan od razloga zašto smo izabrali ovakvo rešenje je taj što je zbog velikog broja novih TV sadržaja (premijera) koji se emituju pogodno koristiti sisteme na bazi filtriranja sadržaja, a koji se mogu implementirati lokalno na korisničkom uređaju. Drugi razlog je taj što ovako izabrano rešenje za zaštitu privatnosti poštuje i drugi princip na kome se bazira GDPR uredba, a koji podrazumeva da su mere za zaštitu privatnosti primenjene i za one korisnike koji nisu svesni postojanja opasnosti za njeno narušavanje. U kontekstu ZZPL Republike Srbije, podrazumevana zaštita svih korisnika se može bliže povezati s načelom srazmernosti obrade, odnosno s tim da je dozvoljeno prikupljati samo one podatke koji su neophodni i relevantni za određenu svrhu obrade. Poseban deo našeg istraživanja biće posvećen ovom aspektu zaštite privatnosti, kroz izbor samo onih kontekstualnih informacija koje su neophodne da bi se ubrzalo učenje korisničkih interesovanja, a koje se ne smatraju posebno osetljivim informacijama. Pored dobrih strana koje ovo rešenje nudi, razmotrili smo i način da se prevaziđe njegova najveća mana, a to je ograničenost resursa koji su dostupni na korisničkom uređaju. U daljem tekstu disertacije razmotrićemo izbor algoritma učenja za sistem baziran na neuralnoj mreži koji se može implementirati čak i na korisničkim uređajima s ograničenim hardverskim resursima. Iako provajder servisa nema pristup podacima korisnika i u nekim tumačenjima možda ne treba ni primenjivati ZZPL, smatramo da je primena načela obrade podataka koji iz njega proizilaze neophodna, imajući u vidu cilj zbog kog se podaci prikupljaju, a i iz razloga da prosečan korisnik nije svestan do kog nivoa narušavanja privatnosti može doći.

Korišćenjem sistema na bazi filtriranja sadržaja lokalno implementiranog na korisničkom uređaju, smanjuju se rizici koji potiču od ostalih korisnika sistema i zaposlenih kod pružaoca usluge distribucije medijskih sadržaja. Do narušavanja privatnosti kod ovih sistema može doći samo ukoliko više korisnika koristi isti uređaj; s druge strane rizik od krađe podataka usled otuđenja ili neovlašćenog pristupa uređaju je znatno veći nego kada se koristi tradicionalna klijent-server arhitektura, pa je potrebno razmotriti dodatne mere zaštite. Jedan od načina

kako bi pružalac usluge distribucije medijskih sadržaja mogao da zaštiti ove podatke je uz pomoć kriptografskih algoritama. Iako ovi algoritmi mogu značajno opteretiti procesorske resurse uređaja, imajući u vidu da isti zahtevi postoje kod IoT (*Internet of Things*) mreža koje su u fokusu istraživanja naučne zajednice, očekujemo razvoj kriptografskih algoritama koji bi zadovoljili ove zahteve [125]. Ipak, sam izbor kriptografskog algoritma prevazilazi temu doktorske disertacije i ostavljen je za buduća istraživanja.

Što se tiče ostalih rešenja za zaštitu privatnosti od direktnog pristupa podacima, smatramo da algoritme koji modifikuju podatke svakako ne treba primenjivati jer nisu u skladu s načelom o tačnosti, odnosno kvalitetu informacija. Nasuprot ovome, zaštita od donošenja osetljivih zaključaka ostavljena je za buduća istaživanja, pri čemu očekujemo da u zaštiti ovog aspekta privatnosti algoritmi koji garantuju diferencijalnu privatnost mogu biti dobro rešenje.

Za kraj ovog dela doktorske disertacije treba spomenuti i da se uskoro očekuje usaglašavanje regulative Republike Srbije s GDPR uredbom, pa pri projektovanju novih personalizovanih programskih vodiča za digitalnu televiziju treba uzeti u obzir i nove zakonske odredbe.

7.2 Mogućnost rada sistema na uređajima s ograničenim resursima

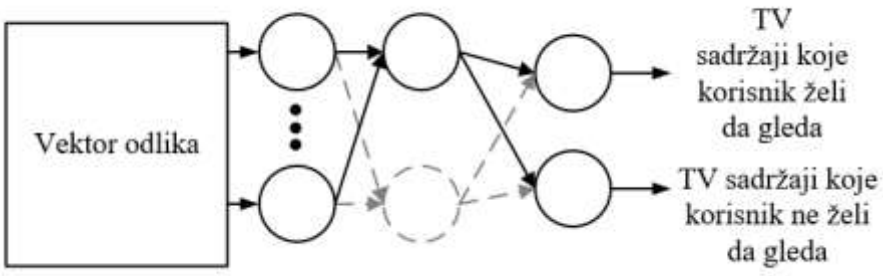
Da bi lokalno implementirani sistem na bazi neuralnih mreža mogao da nauči korisnička interesovanja i pruža preporuke u realnom vremenu, neophodno je izabrati odgovarajući algoritam učenja. Kako se od mobilnih uređaja, s ograničenim hardverskim resursima, očekuje da postanu dominantan način pristupa digitalnoj televiziji, izabrani algoritam mora biti jednostavan.

Koliko je nama poznato, opsežno poređenje performansi personalizovanih programskih vodiča za različite algoritme učenja, pre našeg, u literaturi nije postojalo.

Kao osnovu sistema koristićemo neuralnu mrežu s jednim skrivenim slojem, a zatim menjati broj skrivenih čvorova i ispitati različite algoritme učenja.

Prilikom izbora broja skrivenih čvorova težili smo da vreme treniranja mreže bude što manje, tako da smo u obzir uzeli varijante sistema s jednim i s dva skrivena čvora (slika 7.1). S druge strane, prilikom izbora algoritama koje ćemo uporediti vodili smo računa da svi predstavnici različitih klasa algoritama nadgledanog učenja budu uključeni, te smo kao reprezentativne izabrali RP, SCG, LM i ELM algoritme. Kao meru performansi usvojili smo tačnost

klasifikacije i uporedićemo je za goreopisane varijante sistema koristeći *MovieLens* skup podataka.



Slika 7.1: Blok šema sistema.

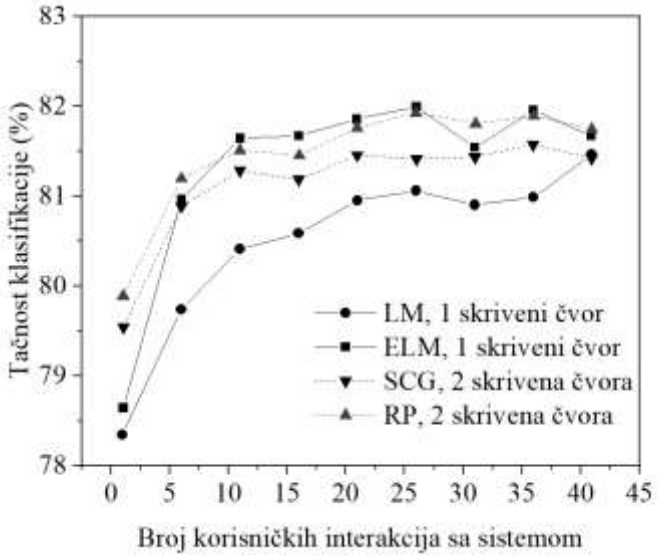
Pored toga u cilju daljeg smanjenja vremena treniranja neuralne mreže primenićemo i linearnu transformaciju koja je predložena u radu [101]. Ona se bazira na množenju vektora odlika s transformacionom matricom čime se originalni vektorski prostor projektuje u prostor čije koordinate označavaju koliko je neki sadržaj emotivan, informativan i opuštajući. Kako transformaciona matrica zavisi od dimenzionalnosti korišćenih podataka i podneblja korisnika, među studentima Elektrotehničkog fakulteta u Beogradu sproveli smo anketu, te formirali adekvatnu matricu, prikazanu u tabeli 7.3.

Tačnost klasifikacije koju sistem postiže kada se optimalni broj čvorova koristi za svaki od razmatranih algoritama prikazana je na slici 7.2.

Kao što se može videti sa slike 7.2, bez obzira na to koji se algoritam učenja koristio, postoji strmi deo krive koji odgovara tkz. hladnom startu, kada korisnik još uvek nema dovoljno interakcija sa sistemom kako bi ovaj naučio njegova interesovanja, i ravni deo krive kada sistem počinje da pruža pouzdane preporuke. Iako se korišćenjem ELM algoritma i arhitekture mreže s jednim skrivenim čvorom postižu nešto lošije performanse prilikom hladnog starta, već nakon 5 do 10 korisničkih interakcija ova varijanta sustiže performanse konfiguracije s dva skrivena čvora, trenirane sa RP algoritmom, te su u ravnom delu krive obe podjednako dobre. Sistem s dva skrivena čvora treniran SCG algoritmom je nešto lošiji od prethodnih, dok se najgore performanse postižu u slučaju kada se koristi LM algoritam i konfiguracija s jednim skrivenim čvorom.

Tabela 7.3. Transformaciona matrica

Ulaz	Izlaz		
	Emotivan	**Informativan**	**Opuštajući**
Action	29	90	30
Adventure	44	82	34
Animation	21	93	30
Children's	23	93	30
Comedy	12	95	29
Crime	60	71	37
Documentary	97	23	10
Drama	31	37	87
Fantasy	26	85	45
Film noir	28	75	60
Horror	27	41	87
Musical	21	86	45
Mystery	51	66	54
Romance	10	44	89
Sci-Fi	62	70	37
Thriller	42	63	65
War	74	39	54
Sport	46	78	42



Slika 7.2: Tačnost klasifikacije u zavisnosti od broja korisničkih interakcija za optimalne konfiguracije broja skrivenih čvorova razmatranih algoritama.

Imajući u vidu da razlike u postignutoj tačnosti klasifikacije nisu toliko velike, vreme treniranja je odlučujući faktor prilikom izbora algoritma i arhitekture mreže personalizovanog

programskog vodiča projektovanog za rad na mobilnim uređajima. Relativna vremena treniranja, proračunata na osnovu simulacija različitih varijanti sistema za 10 korisničkih interakcija prikazana su u tabeli 7.4.

Tabela 7.4. Relativno vreme treniranja neuralne mreže u procentima

	ELM	**LM**	**RP**	**SCG**
Jedan skriveni čvor	0.58	100	83	97
Dva skrivena čvora	0.70	100	83	93

ELM algoritam ubedljivo dominira u ovom aspektu, s više od stotinu puta kraćim vremenom treniranja u odnosu na ostale kandidate. S obzirom na tačnost klasifikacije koja se može postići neuralnom mrežom s jednim skrivenim čvorom treniranom ovim algoritmom, ELM se nameće kao najbolji izbor u konkretnoj primeni.

Na kraju treba napomenuti i da korišćenje transformacione matrice nije imalo značajnog uticaja na vreme treniranja, ali ni na tačnost klasifikacije. Mogući razlog za ovo je relativno mali broj dimenzija originalnog vektorskog prostora (18), te ne treba zanemariti ovu činjenicu prilikom tumačenja dobijenih rezultata. U radu [101] u kojem je originalno predložena ova vrsta linearne transformacije korišćen je 256-dimenzionalni vektorski prostor.

Rezultate ovog istraživanja, objavili smo u radu [53].

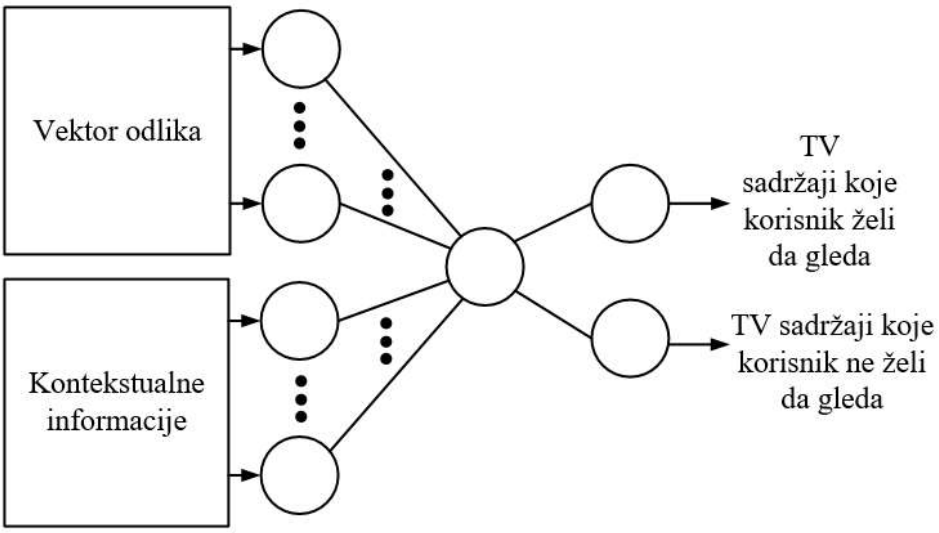
7.3. Izbor i način korišćenja kontekstualnih informacija

Kao što se iz pregleda literature, u poglavlju 4 doktorske disertacije, može zaključiti nesumljivo je da kontekstualne informacije imaju uticaja na to koji će TV sadržaj korisnik izabrati u posmatranoj situaciji. Da bi u potpunosti definisali personalizovani programski vodič, koji ih koristi, neophodno je odrediti koje će se informacije prikupljati i na koji način će biti uključene u proces pružanja preporuka. Iz tog razloga sprovešćemo istraživanje koje obuhvata oba ova aspekta.

U prvoj seriji eksperimenata koristićemo ETF skup podataka za potrebe:

- ispitivanja uticaja kontekstualnih informacija o danu i vremenu emitovanja TV sadržaja na tačnost klasifikacije sistema koji koristi kontekstualno modeliranje i
- poređenja tačnosti klasifikacije sistema za različite načine korišćenja ovih informacija.

Korišćenje kontekstualnog modeliranja usvojeno je za prvu fazu istraživanja jer se polazeći od arhitekture koja je izabrana za optimalnu u smislu mogućnosti rada na uređajima s ograničenim resursima [53] može lako primeniti, dodavanjem novih ulaznih čvorova za svaku od kontekstualnih informacija. Blok-šema tako dobijenog sistema prikazana je na slici 7.3.



Slika 7.3: Uprošćena blok šema predloženog sistema koji koristi kontekstualno modeliranje.

S druge strane, informacije o danu i vremenu emitovanja korišćene su jer ih je lako dobiti iz sistemskog sata korisničkog uređaja, pa nema potrebe da korisnik eksplicitno učestvuje u njihovom prikupljanju.

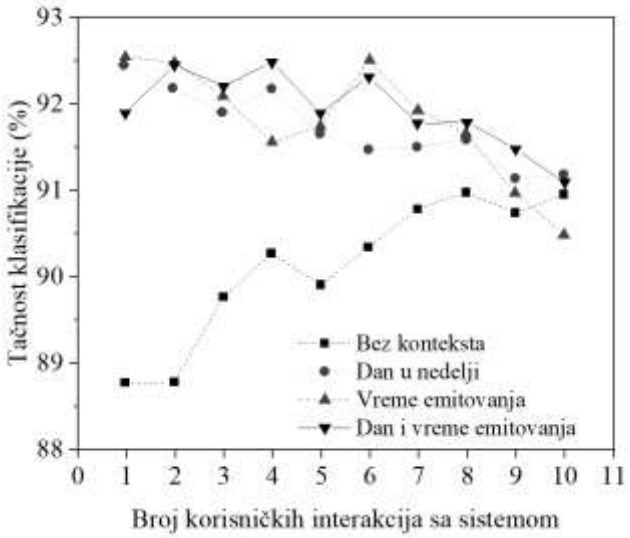
Kao i kod istraživanja koje se tiče ispitivanja mogućnosti rada sistema na uređajima s ograničenim resursima, na isti način formiraćemo i primenićemo transformacionu matricu, ovoga puta prilagođenu ETF skupu podataka. Matrica koju smo dobili primenom metode anketiranja data je u tabeli 7.5.

Na slici 7.4 prikazali smo tačnost klasifikacije, dobijenu kroz proces simulacije sistema, za slučajeve kada:

- se ne koriste kontekstualne informacije,
- koriste informacije o danu u nedelji,
- koriste informacije o vremenu emitovanja TV sadržaja i
- kada se koriste obe kontekstualne informacije.

Tabela 7.5. Transformaciona matrica

Ulaz	**Izlaz**		
	Opuštajući	**Informativan**	**Emotivan**
Action	24	47	29
Adventure	25	46	29
Animation	33	47	20
Children's	25	47	28
Comedy	24	48	28
Crime	24	49	27
Documentary	24	28	48
Drama	22	50	28
Fantasy	22	51	27
Film noir	22	52	26
Horror	23	50	27
Musical	25	48	27
Mystery	25	47	28
Romance	25	46	29
Sci-Fi	24	47	29
Thriller	24	46	30
War	23	48	29
Western	23	51	26
Fun	85	5	10
Music show	52	0	48
Talk show	40	1	59
Lifestyle	72	19	9
News/info	0	95	5
Sport	21	19	60



Slika 7.4: Zavisnost tačnosti klasifikacije od broja korisničkih interakcija sa sistemom u slučajevima kada se koriste i kada se ne koriste kontektualne informacije o danu u nedelji i vremenu emitovanja.

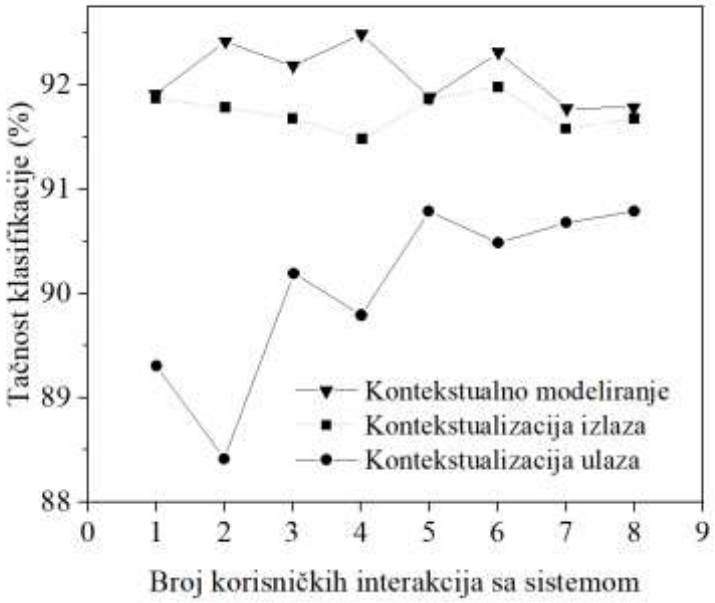
Na osnovu dobijenih rezultata može se zaključiti da se sve verzije sistema koje koriste konktekstualne informacije ponašaju slično i imaju bolje performanse od onih koje sistem postiže kada ih ne koristi. Ovo poboljšanje je posebno vidljivo u delu krive koji odgovara hladnom startu sistema, te se posmatranjem konteksta može ubrzati učenje korisničkih interesovanja i smanjiti broj interakcija potreban da bi personalizovani programski vodič počeo s pružanjem pouzdanih preporuka.

Ukoliko obratimo pažnju samo na verzije koje koriste kontekst, možemo primetiti da se najbolje performanse postižu u slučaju kada su u proces pružanja preporuka uključene informacije o danu u nedelji i vremenu emitovanja TV sadržaja. Štaviše može se uočiti da su performanse ove verzije sistema jednake performansama verzije s jednom kontekstualnom informacijom koja je u posmatranom trenutku bolja. Kako ne postoji nikakav razlog da se ne koriste obe kontekstualne informacije, izabrali smo ovu verziju sistema kao osnovu za dalje projektovanje i istraživanje.

Detaljnija analiza slike otkriva još dva fenomena koji se javljaju kod personalizovanih programskih vodiča. Opadanje performansi do kojeg dolazi u slučajevima kad sistem koristi veći broj interakcija može se povezati s takozvanom prespecijalizacijom koja je karakteristična za sve preporučivače na bazi filtriranja sadržaja. Ukoliko se ovi sistemi treniraju s previše korisničkih interakcija, oni će preporučivati samo TV sadržaje koji su jako slični sadržajima koji su se u prošlosti svideli korisniku. Jedan od načina da se ovo prevaziđe je ograničavanjem njihovog broja, te se dobar kompromis između tačnosti klasifikacije i brzine učenja korisničkih interesovanja može postići sistemom treniranim s 4 interakcije. Nasuprot ovome, razlog oscilacija u performansama sistema koji koristi obe kontekstualne informacije su reprize TV sadržaja, jer su korisnici koji su propustili emitovanje u regularnom terminu često ostajali budni do kasno u noć čekajući reprizu.

U prvoj fazi istraživanja uporedićemo i različite načine korišćenja konteksta. Za razliku od goreopisanog sistema, baziranog na kontekstualnom modeliranju, kod verzije koja koristi kontekstualizaciju ulaza, nema čvorova mreže kroz koje bi se kontekstualne informacije propagirale, već se za svaku kombinaciju ovih informacija trenira posebna neuralna mreža s podacima o TV sadržajima kojima je pristupano u posmatranom kontekstu. S druge strane, kod verzije koja koristi kontekstualizaciju izlaza, isto kao i kod kontekstualizacije ulaza, nema dodavanja ulaznih čvorova, ali se zato svim dostupnim podacima trenira jedna neuralna mreža, a prilikom formiranja liste preporuka u obzir uzimaju samo oni sadržaji koji

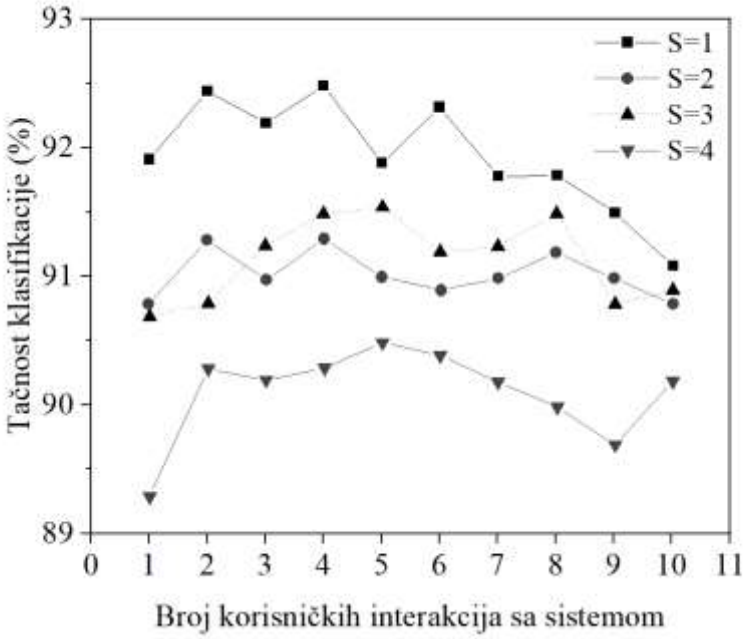
odgovaraju posmatranom kontekstu. Rezultati poređenja ovih verzija sistema prikazani su na slici 7.5.



Slika 7.5: Poređenje sistema koji na različite načine koriste kontekstualne informacije.

Kao što se može uočiti, primena kontekstualizacije izlaza, kao i kontekstualnog modeliranja, ubrzava učenje korisničkih interesovanja, dok to nije slučaj za kontekstualizaciju ulaza jer sistem u kome se na ovaj način koriste kontekstualne informacije ima slične performanse kao i personalizovani programski vodič gde se one ne koriste. Primenom kontekstualnog modeliranja postižu se nešto bolje performanse nego u slučaju kada se koristi kontekstualizacija izlaza, te se naš inicijalni izbor pokazao odličnim i koristićemo sisteme bazirane na njemu u nastavku doktorske disertacije. Koliko je nama poznato, u vreme objavljivanja ovog istraživanja nije postojalo slično koje se bavilo izborom načina korišćenja konteksta kod sistema na bazi filtriranja sadržaja, što se može i proveriti u radu [13] gde je naglašeno da ovaj aspekt konteksta nije dovoljno istražen.

Na kraju prve faze eksperimenata, ispitali smo i da li će izabrana arhitektura mreže, koja je optimalna u smislu tačnosti klasifikacije kada se ne koristi kontekst, optimalna u slučaju kada se on koristi. Performanse sistema koji kroz kontekstualno modeliranje koristi obe kontekstualne informacije za različite brojeve skrivenih čvorova predstavljene su na slici 7.6.



Slika 7.6: Zavisnost tačnosti preporuke od broja skrivenih čvorova neuralne mreže za sistem koji kombinuje kontektualne informacije o danu u nedelji i vremenu emitovanja.

Izbrana arhitektura neuralne mreže pokazala se optimalnom i u slučaju kada se koristi kontekst.

Rezultate prve faze istraživanja konteksta predstavili smo javnosti u radu [126].

U drugoj fazi eksperimenata, nastavićemo s ispitivanjem uticaja kontekstualnih informacija na tačnost klasifikacije i detaljnije ćemo istražiti koje od njih bi trebalo uključiti u proces pružanja preporuka.

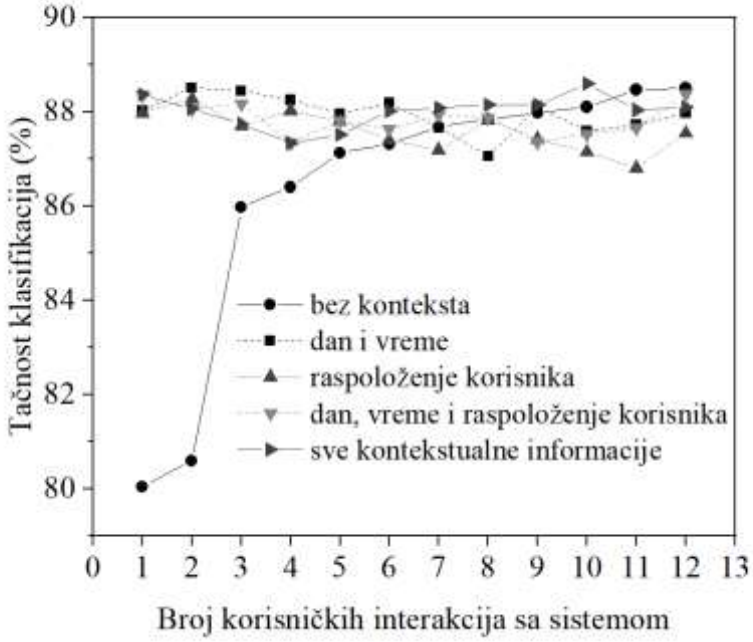
U tu svrhu koristićemo *LDOS-CoMoDa* skup podataka koji sadrži veliki broj raznovrsnih kontekstualnih informacija od kojih ćemo izabrati one koje su po našem mišljenu reprezentativne za dalje istraživanje.

Konkretno, uporedićemo performanse u slučajevima kada se koriste informacije:

- o danu i vremenu emitovanja,
- informacije o raspoloženju korisnika,
- informacije o danu, vremenu emitovanja, i raspoloženju korisnika,
- sve dostupne informacije u skupu i
- kada se ne koriste kontekstualne informacije.

Za razliku od goreopisanog personalizovanog programskog vodiča, na bazi kontekstualnog modeliranja, kod sistema kojeg ćemo koristiti u ovoj fazi istraživanja nismo primenili transformacionu matricu jer se njen dosadašnji uticaj na performanse pokazao zanemarljivim, a njeno kreiranje iziskuje ponovno sprovođenje procedure anketiranje korisnika. U svim ostalim aspektima posmatrani sistem je isti i u potpunosti odgovara blok-šemi na slici 7.3.

Dobijene rezultate poređenja objavili smo u radu [127], a u doktorskoj disertaciji prikazali smo ih na slici 7.7.



Slika 7.7: Zavisnost tačnosti klasifikacije od broja korisničkih interakcija sa sistemom u slučajevima kada se koriste i kada se ne koriste kontekstualne informacije.

Činjenica da se korišćenjem kontekstualnih informacija može ubrzati učenje korisničkih interesovanja potvrđena je i u slučaju sistema treniranog s *LDOS-CoMoDa* skupom podataka. Takođe, primetno je da nema značajnijih razlika u performansama ukoliko se pored dana i vremena emitovanja dodatno koriste i one informacije koje zahtevaju eksplicitno učešće korisnika ili za čije se prikupljanje moraju koristiti statističke metode i tehnike mašinskog učenja, te ćemo u daljim istraživanja ostati pri inicijalnom izboru. Korišćenje informacija koje se ne mogu implicitno prikupiti može negativno uticati na uobičajeni način gledanja televizije ili povećati kompleksnost sistema što bi ugrozilo mogućnost rada na uređajima s ograničenim resursima.

7.4 Problem disbalansa klasa

Problemu disbalansa klasa kod personalizovanih programskih vodiča u literaturi nije posvećeno previše pažnje. Mogući razlog za to je što predikcija TV sadržaja koje korisnik ne želi da gleda nije primarni cilj ovih sistema. Ipak, kako se pogrešno klasifikovani sadržaji koji se ne sviđaju korisniku mogu pojaviti u listi preporuka i direktno uticati na smanjenje njegovog zadovoljstva, sistem mora biti dobar u predikciji obeju klasa. Štaviše, ukoliko vodič greškom dodeli sadržaj koji korisnik želi da gleda klasi onih koje ne želi, zbog velikog broja dostupnih programa, gledaoci televizije to neće tako lako primetiti, kao u suprotnom slučaju.

U poglavlju 5 doktorske disertacije, u kome je dat pregled literature koja se tiče primene neuralnih mreža kod preporučivača, već smo spomenuli da postoje dva scenarija koja se mogu javiti kod problema disbalansa klasa, te će i samo naše istraživanje biti strukturirano na ovaj način.

U slučajevima kada je i pored problema disbalansa klasa prikupljeno dovoljno negativnih interakcija koje se mogu uključiti u proces treniranja neuralne mreže, ispitaćemo učinkovitost korišćenja WELM (*Weighted Extreme Learning Machine*) algoritma učenja [126]. On se bazira na istoj paradigmi kao i osnovni ELM, kod kojeg se ulazne težine na početku slučajno izaberu i ne podešavaju u toku treninga, samo se razlikuje izraz na osnovu kojeg se računaju izlazne težine. U WELMu, ovaj izraz glasi

$$\boldsymbol{\beta} = \begin{cases} \mathbf{H}^T \left(\frac{1}{R}\mathbf{I} + \mathbf{B}\mathbf{H}\mathbf{H}^T \right)^{-1} \mathbf{B}\mathbf{Z}, \text{ kada je } P < S \\ \left(\frac{1}{R}\mathbf{I} + \mathbf{H}^T\mathbf{B}\mathbf{H} \right)^{-1} \mathbf{H}^T\mathbf{B}\mathbf{Z}, \text{ kada je } P > S \end{cases}, \qquad 7.1$$

gde je $\mathbf{I}$ jedinična matrica, $\mathbf{B}$ dijagonalna matrica sa težinskim faktorima B_{pp}, p=1,...,P pomoću koje se u proces učenja uključuju informacije o nejednakom broju podataka u pojedinačnim klasama [126] i $\frac{1}{R}$ pozitivna vrednost koja se koristi za regularizaciju [109]. Verzija ELM algoritma kod koje je u izrazu (7.1) matrica $\mathbf{B}$ zamenjena jediničnom matricom $\mathbf{I}$ naziva se RELM (*Regularized Extreme Learning Machine*).

Adekvatnim izborom matrice $\mathbf{B}$ moguće je pomeriti ravan odlučivanja prema tačkama koje odgovaraju klasi s većim brojem podataka, tako da možemo očekivati da se tačnost predikcije

sadržaja koje korisnik ne želi da gleda poveća. U našem istraživanju, ispitaćemo dve šeme za proračun težinskih faktora koje su i predložene u originalnom radu koji opisuje WELM algoritam [126]:

$$\textbf{\textit{šema}}1: B_{pp} = \begin{cases} \dfrac{1}{mes\{N_p\}}, \text{ za podatke iz pozitivne klase} \\ \dfrac{1}{mes\{N_n\}}, \text{ za podatke iz negativne klase} \end{cases}, \qquad 7.2$$

$$\textbf{\textit{šema}}2: B_{pp} = \begin{cases} \dfrac{1}{mes\{N_p\}}, \text{ za podatke iz pozitivne klase} \\ \dfrac{0.618}{mes\{N_n\}}, \text{ za podatke iz negativne klase} \end{cases}. \qquad 7.3$$

Šema 1 ima za cilj da izjednači uticaj pojedinačnih klasa, dok šema 2 teži da postigne odnos uticaja 1:0.618 između klase s većim i klase s manjim brojem podataka, što odgovara tkz. zlatnom preseku koji postoji u prirodi. Za korišćenje varijante ELM algoritma, kao metode za borbu protiv disbalansa klasa odlučili smo se zbog značajno kraćeg vremena treniranja koje ovi algoritmi postižu u odnosu na ostale razmatrane.

Pored algoritma učenja, zbog prisustva problema disbalansa klasa, prilagodićemo i samu meru tačnosti pružanja preporuka. Usvojili smo korišćenje G-mean metrike jer je znatno pogodnija za korišćenje u ovakvim situacijama od tačnosti klasifikacije, ali ćemo dodatno performanse za svaku od pojedinačnih klasa ispitati i na ROC grafiku.

Iz ETF skupa podataka izdvojili smo 15 korisnika s ukupno 664 interakcije, od čega je 605 pozitivnih, a 59 negativnih. Detaljniji opis podataka za svakog od korisnika može se pronaći u tabeli 7.1. U obzir smo uzeli samo one koji imaju barem 2 negativne interakcije, jednu koja će se koristiti za treniranje mreže i još jednu za proračun mera performansi. Trening skup formirali smo tako da s povećanjem broja interakcija koje se posmatraju, odnos broja pozitivnih i negativnih u ovom skupu teži odnosu koji postoji u prikupljenim podacima.

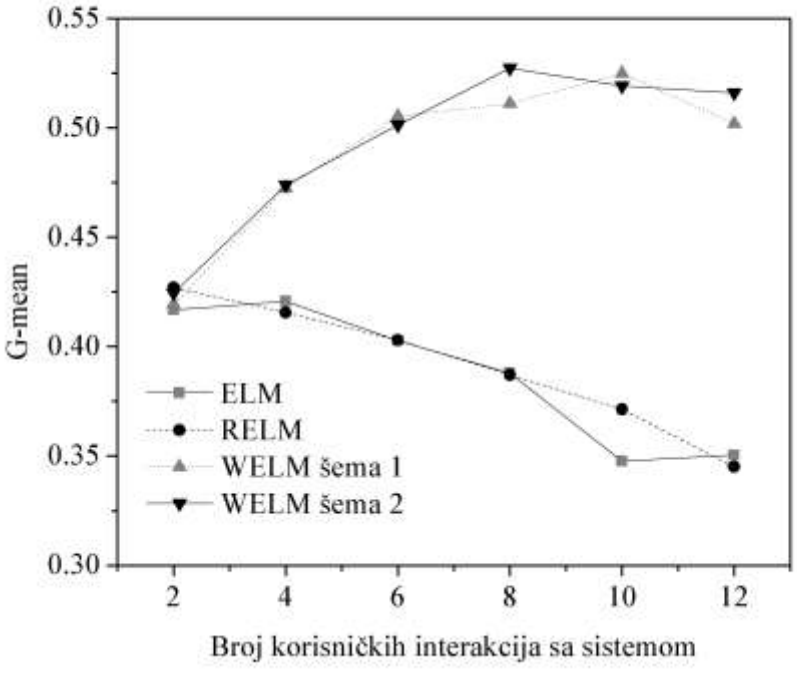
Blok-šema sistema koju ćemo koristiti razlikuje se samo po broju skrivenih čvorova od one koja je prikazana na slici 7.3 - koristeći metod simulacije odredićemo ovaj broj tako da maksimizira vrednost G-mean metrike za različite algoritme. Na ovaj način dobijene

optimalne vrednosti parametara za ELM, RELM, i obe šeme WELM algoritma prikazane su u tabeli 7.6, a samo poređenje G-mean metrike sistema koji ih koriste na slici 7.8. Pored osnovnog ELM, i WELM algoritma, performanse smo ispitali i za RELM algoritam kako bi ustanovili da li je poboljšanje sposobnosti generalizacije mreže dobar način da se ublaži problem disbalansa klasa.

Tabela 7.6. Parametri testiranih algoritama

Algoritam	**Broj skrivenih čvorova,** S	**Parametar regularizacije,** R
ELM	15	-
RELM	16	2^{48}
WELM šema 1	10	2^{2}
WELM šema 2	15	2^{2} ili 2^{51}

Kao što se može videti iz tabele, u slučaju kada se koristi šema 2 WELM algoritma neophodno je definisati dva parametra regularizacije, jedan za slučaj jednakog broja podataka u klasama (2^{51}) i jedan za slučaj kada je on nejednak (2^{2}), kako bi se postigle optimalne performanse.

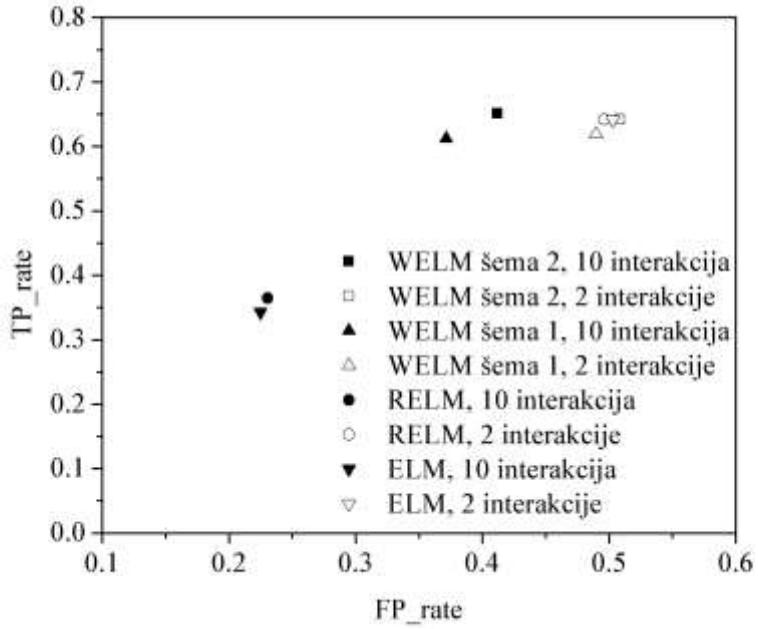


Slika 7.8: Zavisnost G-mean metrike od broja korisničkih interakcija.

Slika 7.8 jasno pokazuje da nakon 2 korisničke interakcije, kada trening skup ima podjednak broj podataka u svakoj od klasa, sve varijante sistema imaju veoma slične performanse. Nasuprot tome, za veći broj interakcija, kada problem disbalansa klasa postane dominantan,

sistem treniran bilo kojom varijantom WELM algoritma ima značajno bolje performanse u odnosu na one koje su dobijene u slučajevima kada se koriste preostali algoritmi. Ovaj rezultat je dobar pokazatelj činjenice da poboljšanje generalizacionih sposobnosti mreže nije dovoljno za borbu protiv problema disbalansa klasa. S druge strane, ukoliko uporedimo samo predložene šeme WELM algoritma, može se primetiti da je razlika između odgovarajućih vrednosti G-mean metrike zanemariva.

Performanse sistema u slučajevima kada je broj podataka u svakoj od klasa podjednak (nakon 2 interakcije) i kada je problem nejednakog broja podataka u pojedinačnim klasama dominantan (nakon 10 interakcija) detaljnije ćemo ispitati i na ROC grafiku prikazanom na slici 7.9.

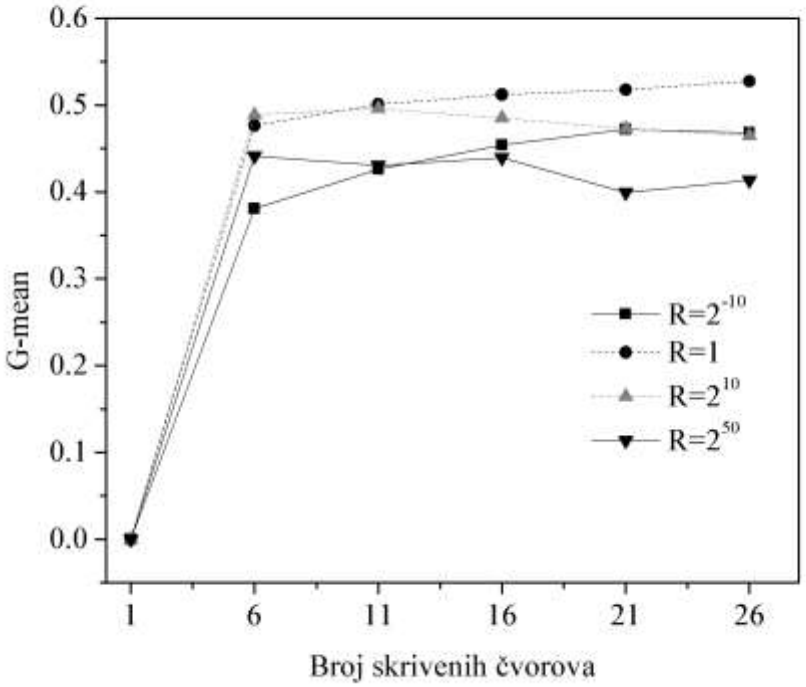


Slika 7.9: ROC grafik za posmatrane sisteme nakon 2 i 10 korisničkih interakcija.

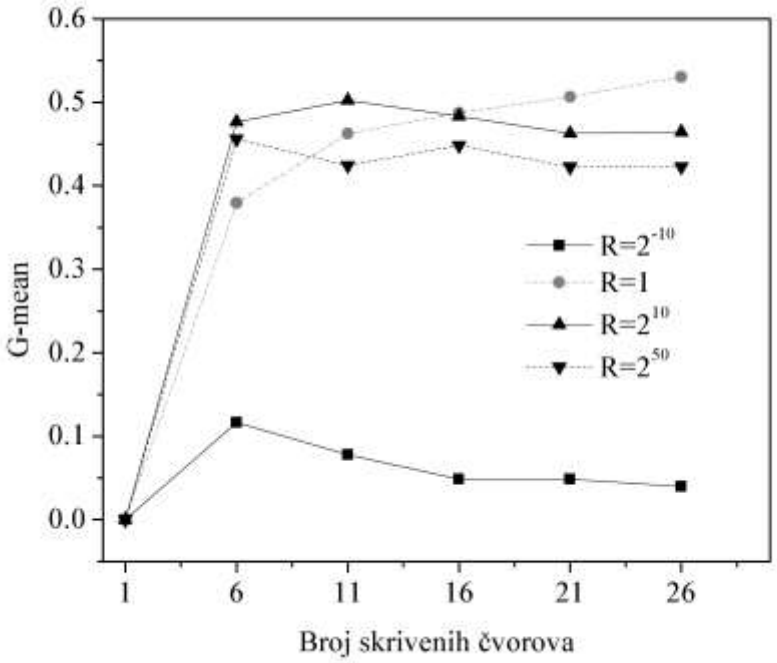
Dobijeni rezultati su očekivani. U slučaju nejednakog broja podataka u pojedinačnim klasama, sistem koji koristi ELM ili RELM algoritam imaće manje *TP_rate* i *FP_rate* vrednosti u odnosu na onog koji je treniran WELM algoritmom. Ukoliko neki od metoda borbe protiv posmatranog problema nije primenjen, disbalans klasa degradiraće tačnost klasifikacije sadržaja koje korisnik ne želi da gleda (što dovodi do manje *TP_rate* vrednost), a poboljšati tačnost klasifikacije sadržaja koje korisnik želi da gleda (što kao rezultat daje manju *FP_rate* vrednost). Nasuprot ovome, bez obzira na to koja se od predloženih šema koristila, upotrebom WELM algoritma postiže se bolja klasifikacija sadržaja koje korisnik ne

želi da gleda (veća *TP_rate* vrednost) po cenu nešto lošije klasifikacije sadržaja koje korisnik želi da gleda (veća *FP_rate* vrednost). Ovo može dovesti do toga da sistem treniran WELM algoritmom ne uključi u listu preporuka baš sve sadržaje koje bi korisnik želeo da gleda, ali će zato značajno smanjiti verovatnoću pojavljivanja u listi sadržaja koje korisnik ne želi da gleda. Naglašavamo da često pojavljivanje sadržaja koje korisnik ne želi da gleda u listi preporuka može znatno lošije da utiče na korisničko zadovoljstvo nego kada sistem povremeno preskoči da preporuči sadržaj koji korisnik želi da gleda.

Kako se ni na ROC grafiku ne mogu uočiti značajnije razlike između sistema koji koriste različite šeme WELM algoritma, prilikom izbora optimalne uzeli smo u obzir i složenost odabira njihovih parametara. Zavisnost vrednosti G-mean metrike koja se dobija nakon 10 korisničkih interakcija od broja skrivenih čvorova i parametra regularizacije za šemu 1 i šemu 2 prikazana je na slici 7.10 i slici 7.11, respektivno.



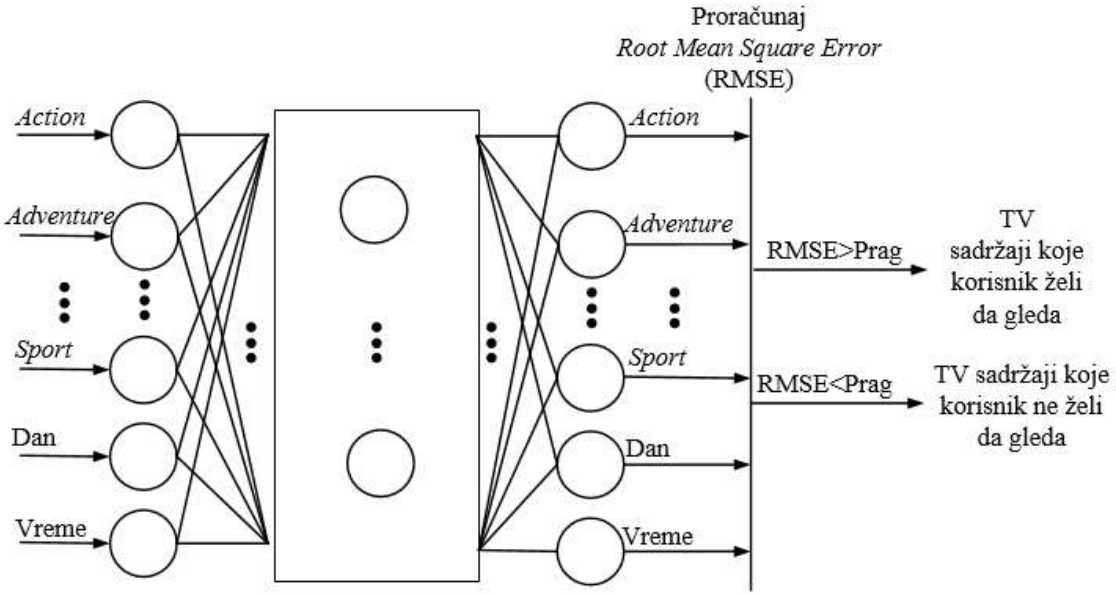
Slika 7.10: Šema 1 - Zavisnost G-mean metrike od broja skrivenih čvorova i parametra regularizacije nakon 10 interakcija.



Slika 7.11: Šema 2 – Zavisnost G-mean metrike od broja skrivenih čvorova i parametra regularizacije nakon 10 interakcija.

Bez obzira na to koja se šema koristi, vidi se da parametar regularizacije ima znatno veći uticaj na performanse sistema nego broj skrivenih čvorova. Ukoliko se on adekvatno izabere, broj skrivenih čvorova potrebno je samo izabrati tako da bude dovoljno velik. Ipak, evidentno je da uticaj parametra regularizacije nije podjednak za obe šeme. Pogrešan izbor ovog parametra u slučaju šeme 2 može znatno više degradirati performanse nego kada se koristi šema 1, te smo se, imajući ovo u vidu, prilikom implementacije našeg personalizovanog programskog vodiča odlučili za šemu 1. Rezultate istraživanja uticaja nejednakog broja podataka u pojedinačnim klasama, koje smo do sada predstavili, objavili smo u radu [123].

U istraživanju koje se tiče disbalansa klasa ispitaćemo i slučaj kada zbog ponašanja samih korisnika sistem nije u mogućnosti da prikupi podatke o sadržajima koji im se ne sviđaju, te je problem pružanja preporuka pogodno posmatrati kao problem binarne klasifikacije na osnovu informacija o samo jednoj klasi. Teorijsko istraživanje, koje smo sproveli u poglavlju 4, pokazalo je da je tkz. autoenkoder arhitektura neuralne mreža odgovarajuća za ovu primenu, pa smo je i usvojili za osnovu našeg sistema. Blok-šema predloženog personalizovanog programskog vodiča koji je koristi prikazana je na slici 7.12.



Slika 7.12: Blok-šema predloženog sistema.

Da bi u potpunosti definisali predloženi sistem, kroz seriju eksperimenata ispitaćemo koji je algoritam učenja najpogodniji za treniranje autoenkoder mreže.

Imajući u vidu zahtev za malim vremenom treniranja koje se očekuje čak i na uređajima s ograničenim hardverskim resursima, rešenje smo potražili među sistemima koji koriste algoritme bazirane na ELM paradigmi. U trenutku pisanja doktorske disertacije i izvođenja eksperimenata, u literaturi su se kao reprezentativni izdvojili tkz. *ELM autoenkoder* [128] i *ELM sparse autoenkoder* [129], te ćemo ih detaljnije istražiti.

Kod *ELM autoenkodera* proces treniranja neuralne mreže posmatra se kao l_2 optimizacioni problem, kod kojeg je cilj minimizirati izraz:

$$\boldsymbol{\beta} = argmin(\mathbf{H}\boldsymbol{\beta} - \mathbf{Z}^2 + \frac{1}{R}\boldsymbol{\beta}^2).$$

7.4

Kao njegovo rešenje dobija se izraz za proračun izlaznih težina koji je identičan kao kod RELM algoritma:

$$\boldsymbol{\beta} = \mathbf{H}^{+}\mathbf{Z} = \left(\frac{1}{R}\mathbf{I} + \mathbf{H}\mathbf{H}^{\mathbf{T}}\right)^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{Z}, \qquad 7.5$$

s tim da se kod autoenkodera [128] nakon slučajno izabranih ulaznih težina i *bias*-a skrivenih čvorova primenjuje algoritam ortogonalizacije vektora sačinjenog od ovih vrednosti kako bi se poboljšale generalizacione sposobnosti neuralne mreže. Konkretno u našoj implementaciji autoenkodera za ovu svrhu koristićemo *Gram-Schmidtov* algoritam.

Nasuprot ovome, proces treniranja kod *ELM sparse autoenkodera* [129] posmatra se kao l_1 optimizacioni problem, kod kojeg je cilj minimizirati sledeći izraz:

$$\boldsymbol{\beta} = argmin\left(\mathbf{H}\boldsymbol{\beta} - \mathbf{Z}^2 + \frac{1}{R}|\boldsymbol{\beta}|_1 \right), \tag{7.6}$$

i ne primenjuje se ortogonalizacija ulaznih težina i *bias*a skrivenih čvorova. Sa $|\boldsymbol{\beta}|_1$, u ovom izrazu, označena je apsolutna suma pojedinačnih vrednosti u matrici izlaznih težina $\boldsymbol{\beta}$.

Za treniranje *ELM sparse autoenkodera* koristi se FISTA algoritam (*Fast Iterative Shrinkage-Thresholding algorithm*) [130], male računske složenosti, koji je pogodan za rešavanje l_1 optimizacionih problema.

Da bismo opisali proces treniranja kod FISTA algoritma, najpre označimo deo optimizacionog problema koji odgovara minimiziranju greške proračunate u toku treniranja sa

$$m(\boldsymbol{\beta}) = \mathbf{H}\boldsymbol{\beta} - \mathbf{Z}^2, \tag{7.7}$$

a deo koji odgovara l_1 regularizaciji [110] sa

$$n(\boldsymbol{\beta}) = \frac{1}{R}|\boldsymbol{\beta}|_1. \tag{7.8}$$

Tada se ovaj algoritam može opisati sledećom procedurom:

1. Izračunaj *Lipschitzovu* konstantu λ gradijenta funkcije $\nabla m(\boldsymbol{\beta})$;

2. Prvu iteraciju započni s sledećim vrednostima $\mathbf{y}_1 = \boldsymbol{\beta}_0$, $t_1 = 1$, a zatim za svaku sledeću iteraciju izvrši sledeće proračune:

   a) $\boldsymbol{\beta}_k = g_\lambda(\mathbf{y}_k),\ g_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{\lambda}{2}\boldsymbol{\beta} - \left(\mathbf{y}_k - \frac{1}{\lambda}\nabla m(\mathbf{y}_k) \right)^2 + n(\boldsymbol{\beta}) \right\},$ (7.9)

b) $t_k = \frac{1+\sqrt{1+4t_k^{\,2}}}{2}$, 7.10

c) $\mathbf{y}_{k+1} = \boldsymbol{\beta}_k + \frac{t_k - 1}{t_{k+1}}\left(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}\right)$, 7.11

gde je s $\boldsymbol{\beta}_k$ označena matrica izlaznih težina u *k*-toj iteraciji. Iako je po svojoj prirodi iterativan, FISTA algoritam ima globalnu brzinu konvergencije $\frac{1}{k^2}$ [130], pa u praksi već nakon 50 iteracija algoritam postiže zadovoljavajuće performanse [129]. Ovu vrednost broja iteracija kao kriterijum prestanka treniranja ćemo i mi usvojiti u našim eksperimentima.

Kako *ELM autoenkoder* koristi l_2 regularizaciju kod njega možemo očekivati mrežu s velikim brojem malih izlaznih težina, dok kod *ELM sparse autoenkodera*, zbog primene l_1 regularizacije očekujemo mrežu s većim brojem težina koje su jednake nuli, ali će zato preostale težine imati veće vrednosti nego u slučaju potonjeg.

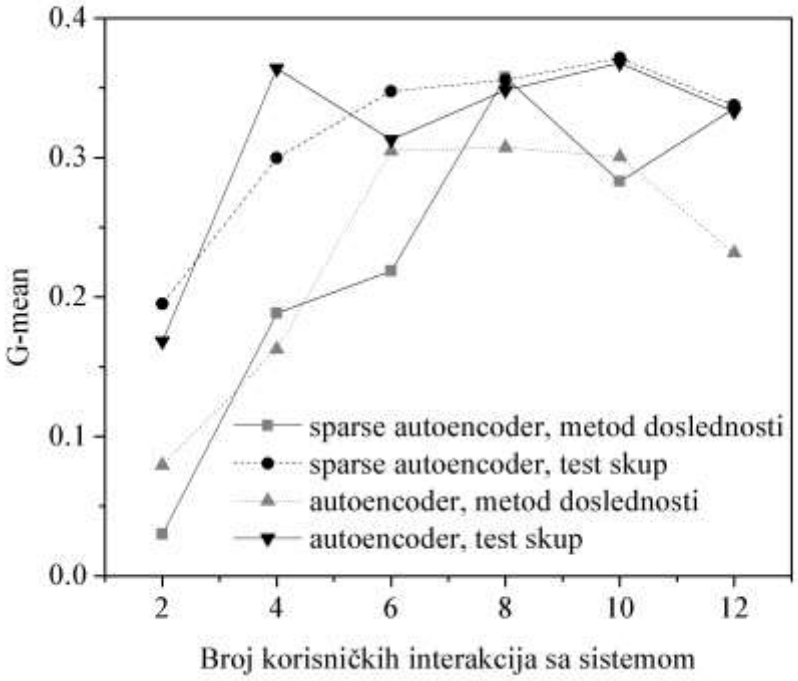
Kao i u slučajevima kada je prikupljeno dovoljno negativnih interakcija, za testiranje performansi sistema, u ovom delu istraživanja koristićemo ETF skup podataka, ali ćemo u obzir uzeti i korisnike koji su barem jedan sadržaj negativno ocenili. Iako sam sistem, opisan blok-šemom na slici 7.12, može da pruža preporuke i u situacijama kada su dostupne samo informacije o poželjnim TV sadržajima, barem jedna negativna ocena mora da postoji u test skupu da bi se proračunale vrednosti G-mean metrike. Pored različitih algoritama učenja, ispitaćemo i dva načina za odabir parametara sistema: na osnovu G-mean metrike dobijene za test skup i korišćenjem metoda doslednosti.

Optimalne vrednosti parametara, izabrane kroz proces simulacije sistema, prikazane su u tabeli 7.7, a njima odgovarajuće vrednosti G-mean metrike na slici 7.13.

U slučajevima kada se za određivanje parametara koristi metod doslednosti, potrebno je formirati validacioni skup i definisati parametar p_e. Za razliku od test skupa, validacioni ne treba da sadrži negativne ocene, te ga je lako kreirati. S druge strane, parametar p_e se mora odrediti empirijski. Sprovedeni eksperimenti su pokazali da se najveće vrednosti G-mean metrike za *ELM autoenkoder* i *ELM sparse autoenkoder* dobijaju kada ovaj parametar iznosi 0.3 i 0.4, respektivno.

Tabela 7.7. Parametri sistema

Autoenkoder	**Određivanje parametara**	**Broj skrivenih čvorova,** *S*	**Parametar regularizacije,** *R*	**Prag odlučivanja**
ELM autoenkoder	Na osnovu performansi dobijenih za test skup	11	2^{34}	1
ELM autoenkoder	metod doslednosti	25	2^{4}	1
ELM sparse autoenkoder	Na osnovu performansi dobijenih za test skup	8	2^{-18}	1
ELM sparse autoenkoder	Metod doslednosti	25	2	1



Slika 7.13: Zavisnost G-mean metrike od broja korisničkih interakcija za sisteme na bazi autoenkoder mreže.

Kao što se može videti sa slike 7.13, sistemi kod kojih se parametri određuju na osnovu performansi dobijenih za test skup postižu veće vrednosti G-mean metrike u odnosu na

sisteme kod kojih se parametri određuju metodom doslednosti samo u delu krive koje odgovara malom broju korisničkih interakcija - kada sistem još uvek uči korisnička interesovanja. U delu krive koja odgovara situaciji kada je sistem počeo da pruža pouzdane preporuke, nema značajnih razlika u performansama ispitanih sistema. Stoga, ako u obzir uzmemo i činjenicu da je u slučaju kada se estimiranje parametara vrši na osnovu proračunatih performansi neophodno imati barem jednu negativnu interakciju, prilikom implementacije našeg vodiča koristićemo metod doslednosti.

Ukoliko uporedimo sisteme bazirane na *ELM autoenkoderu* i na *ELM sparse autoenkoderu* postignute vrednosti G-mean metrike se značajno ne razlikuju za isti način izbora parametra sistema, te ćemo kroz novu seriju eksperimenata detaljnije istražiti vremena potrebna za njihovo treniranje. Dobijene vrednosti prilikom posmatranja 10 korisničkih interakcija prikazane su u tabeli 7.8.

Tabela 7.8. Relativno vreme treniranja neuralne mreže

Autoenkoder	**Vreme treniranja (%)**
ELM autoenkoder	100
ELM sparse autoenkoder	5.73

Vidi se da sistem baziran na ELM sparse autoenkoderu postiže skoro 20 puta kraće vreme treniranja od drugog kandidata, te imajući u vidu da personalizovani programski vodič mora da na efikasan način koristi resurse korisničkih uređaja, odlučili smo se za korišćenje ovog sistema u posmatranom scenariju. Naučni doprinos ovog istraživanja potvrđen je njegovim objavljivanjem u radu [124].

Pre nego što budemo definisali konačan predlog našeg preporučivača za digitalnu televiziju, kroz seriju eksperimenata ispitaćemo i izbor mera performansi.

7.5 Izbor mera performansi

Na izbor mera performansi koje će se koristiti najveći uticaj ima cilj koji pružalac usluga distribucije medijskih sadržaja želi da postigne ovim sistemom, ali i same karakteristike podataka. Kako nam ovaj cilj nije poznat, istraživanje ćemo ograničiti na uticaj disbalansa klasa i opšte ciljeve koje svaki personalizovani programski vodič za digitalnu televiziju teži da postigne.

Da bismo odredili optimalnu meru koja je pogodna za rad u okruženjima gde postoji problem

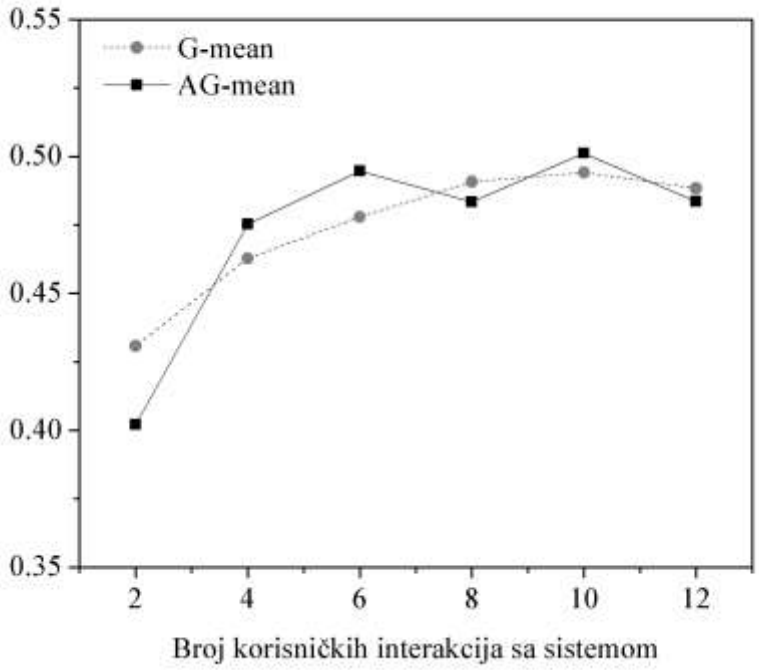
nejednake količine podataka u pojedinačnim klasama, uz pomoć ROC grafika ispitaćemo performanse sistema, opisanog u našem radu [55] i odeljku 7.4, u slučajevima kada se maksimizira vrednost G-mean i AG-mean metrike. Konkretno, koristićemo šemu 1 WELM algoritma učenja, datu izrazom (7.2), jer se ona pokazala kao optimalna u smislu složenosti izbora parametara.

Simulacijom sistema u slučajevima kada se koriste razmatrane mere performansi, odredićemo najpre optimalne vrednosti parametara sistema za svaku od njih, a zatim i predstaviti same vrednosti metrika. Dobijene vrednosti parametara prikazane su u tabeli 7.9.

Tabela 7.9. Vrednosti parametara sistema koje maksimizuju posmatrane mere performansi

Mera performansi	**Broj skrivenih čvorova**	**Parametar regularizacije,** R
G-mean	10	2^2
AG-mean	17	1

Zavisnost vrednosti G-mean i AG-mean metrike od broja korisničkih interakcija dobijena za sisteme koji koriste optimalne parametre za svaku od metrika prikazana je na slici 7.14.

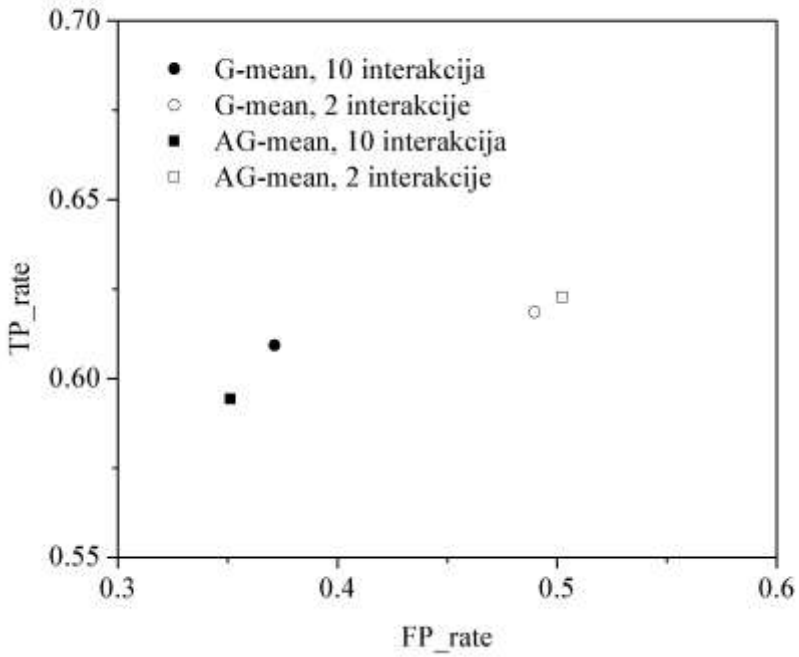


Slika 7.14: Zavisnost G-mean i AG-mean metrike od broja korisničkih interakcija.

Kao što se može videti na slici, bilo da se koristi G-mean metrika, bilo AG-mean metrika, postoji deo krive koji odgovara hladnom startu sistema i deo krive kada sistem počinje da

pruža pouzdane preporuke. U delu krive koji odgovara situaciji kada je sistem naučio korisnička interesovanja i počeo s pružanjem pouzdanih preporuka, kod sistema koji koristi AG-mean metriku mogu se uočiti određene oscilacije. Oscilovanje AG-mean metrike je rezultat oscilacija tačnosti klasifikacije sadržaja koje korisnik voli da gleda, a koje nastaju jer pojedini gledaoci ukoliko propuste svoju omiljenu emisiju u uobičajeno vreme emitovanja, neretko ostaju duže ispred televizora kako bi pogledali reprizu - što komplikuje učenje korisničkih interesovanja. Ovakve oscilacije se ne mogu uočiti kod G-mean metrike jer tačnost klasifikacije sadržaja koje korisnik voli da gleda ima znatno veći uticaj na AG-mean metriku nego na G-mean metriku.

Kako bi detaljnije ispitali performanse sistema za svaku od pojedinačnih klasa, na ROC grafiku posmatraćemo situaciju nakon 2 interakcije, kada sistem i dalje uči o korisničkim interesovanjima, i nakon 10 interakcija kada je sistem počeo da pruža pouzdane preporuke. Rezultati eksperimenata prikazani su na slici 7.15.



Slika 7.15: ROC grafik za sisteme koji maksimizuje G-mean i AG-mean metrike, respektivno.

Nakon 2 korisničke interakcije sistem koji maksimizira vrednost AG-mean metrike ima nešto veće *TP_rate* i *FP_rate* vrednosti nego sistem koji je optimalan u smislu G-mean metrike. Ipak, nakon 10 korisničkih interakcija, odnos između njih se menja, te će sistem koji je optimalan u smislu AG-mean metrike imati nešto manje *TP_rate* i *FP_rate* vrednosti od onih

koje postiže vodič izabran na osnovu G-mean metrike. Ukoliko protumačimo ove vrednosti, zaključujemo da se korišćenjem AG-mean umesto G-mean metrike, prilikom hladnog starta, postiže manja verovatnoća pojave sadržaja koje korisnik ne voli da gleda po cenu da se svi sadržaja koje bi korisniku bili interesantni ne pojave u listi preporuka. Dok se na isti način, za scenario kada je sistem počeo s pružanjem pouzdanih preporuka, postiže nešto veća verovatnoća preporučivanja sadržaja koje korisnik ne voli da gleda, ali i mogućnost preporučivanja sadržaja koje bi korisnik voleo da gleda, a koji ne odgovaraju njegovim dominantnim interesovanjima.

Da rezimiramo, korišćenjem AG-mean metrike moguće je postići različite ciljeve pružalaca usluga distribucije medijskih sadržaja u različitim fazama funkcionisanja sistema. Manja verovatnoća preporučivanja sadržaja koje korisnik ne voli da gleda u fazi hladnog starta može uticati na povećanje poverenja korisnika u preporuke, što je veoma važno kako bi korisnik nastavio da koristi sistem. Iako to dolazi po cenu nepreporučivanja sadržaja koji ne odgovaraju dominantnim interesovanjima, smatramo da je ova osobina manje važna u fazi hladnog starta, ali zato izuzetno važna u drugoj fazi funkcionisanja sistema. U fazi kada sistem počne s pružanjem pouzdanih preporuka, korišćenjem AG-mean metrike povećava se sposobnost pružanja raznovrsnijih preporuka, po cenu povremenog pojavljivanja sadržaja koji korisnik ne želi da gleda u listi preporuka. Smatramo da je uticaj povremenog preporučivanja sadržaja koji se korisniku ne sviđaju na njegovo poverenje u ovoj fazi funkcionisanja mnogo manji nego u inicijalnoj kada tek počinje da se koristi sistem.

Istraživanje opisano u ovom odeljku realizovano je s drugim izborom podataka i objavljeno u našem radu [131]. Za potrebe doktorske disertacije i određivanja konačnog predloga, ponovili smo ga tako da budu uporedivi s rezultatima predstavljenim u odeljku 7.4.

Ostale osobine personalizovanih programskih vodiča, koje se posmatraju prilikom procene učinka sistema, diskutovaćemo u poglavlju koje se bavi analizom performansi našeg konačnog predloga.

7.6. Konačni predlog personalizovanog programskog vodiča

Imajući u vidu sva prethodna istraživanja, predlažemo personalizovani programski vodič za digitalnu televiziju s dva moda rada, koji je implementiran lokalno na korisničkim uređajima.

U slučajevima kada još uvek nisu prikupljene informacije o TV sadržajima koje korisnik ne voli da gleda, sistem će koristiti *ELM sparse autoenkoder* neuralnu mrežu kod koje se

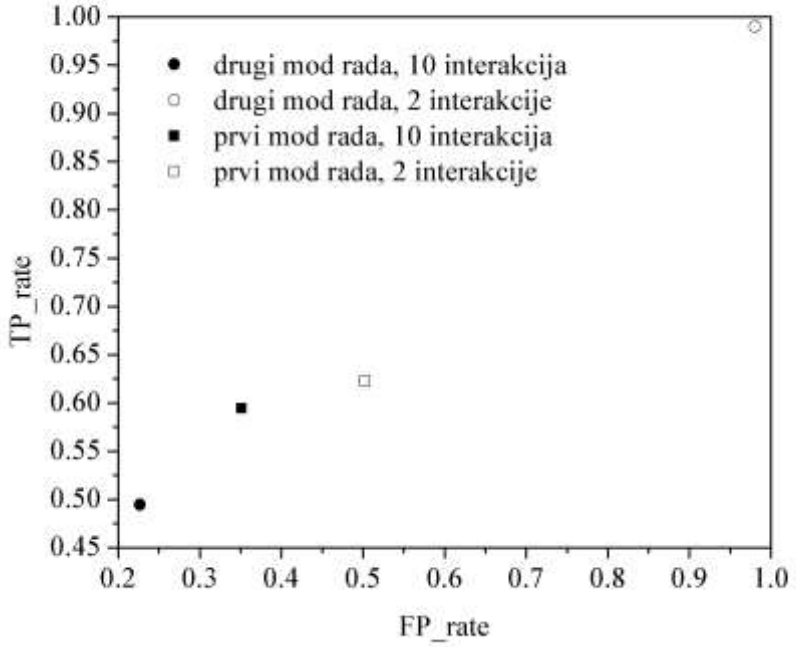
parametri biraju na osnovu metoda doslednosti. Vrednosti parametara korišćene prilikom naše implementacije mogu se pronaći u tabeli 7.7, dok je sama blok šema sistema koji funkcioniše u ovom modu rada prikazana na slici 7.12. Iako je AG-mean metrika pokazala svoje prednosti u odnosu na G-mean, nismo se odlučili za nju jer očekujemo da vodič u ovom modu rada bude samo kada nema dostupnih negativnih interakcija. Kako je za proračun bilo koje od ovih metrika potrebno da postoji barem jedna ovakva ocena, predloženi sistem će u trenutku kada bude mogao da proceni svoje performanse preći u drugi mod rada.

U drugom modu rada vodič će koristiti neuralnu mrežu s jednim skrivenim slojem treniranu WELM algoritmom koji maksimizira AG-mean metriku, uz primenu šeme 1 (izraz 7.2) za borbu protiv disbalansa klasa. Uprošćena šema sistema u ovom modu funkcionisanja u potpunosti odgovara onoj slici 7.3, samo se razlikuje broj skrivenih čvorova – koji se za našu implementaciju može pronaći u tabeli 7.9. Iako uticaj kontekstualnih informacija u smislu dobijene vrednosti AG-mean metrike nismo detektovali, imajući u vidu sprovedene eksperimente i teorijsko istraživanje predstavljeno u poglavlju 4, smatramo da se trebaju koristiti, ali da njihov doprinos kvalitetu preporuka nije moguće izmeriti pomoću ove mere performansi. Korišćenje transformacione matrice, je opcioni deo sistema u ovom modu rada, ali ćemo ga u konkretnoj implementaciji koristiti jer smatramo da može imati uticaja na vreme treniranja u slučajevima kada su TV sadržaji predstavljeni u vektorskom prostoru s velikim brojem dimenzija.

8. Analiza performansi

U ovom poglavlju ispitaćemo performanse predloženog sistema za različite modove rada, diskutovati prednosti i mane našeg pristupa i detaljnije definisati kada će se koji mod koristiti.

Kroz seriju simulacija personalizovanog programskog vodiča, opisanog u prethodnom poglavlju, došli smo do rezultata na osnovu kojih se mogu doneti zaključci o postignutim performansama za svaki mod rada. Kao podatke o korisničkim interakcijama koristili smo one koji su dostupni u okviru ETF skupa podataka. Dobijeni rezultati predstavljeni su na ROC grafiku, prikazanom na slici 8.1.



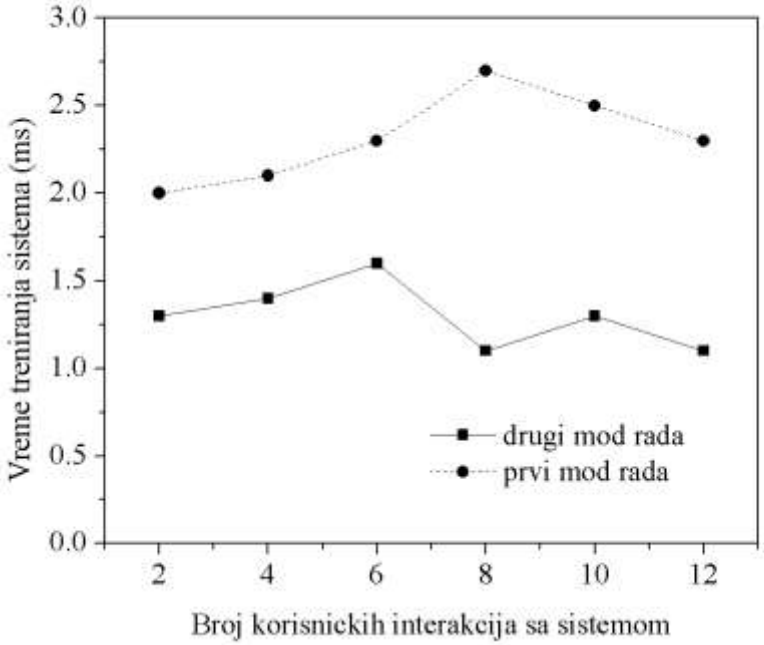
Slika 8.1: ROC grafik za različite modova rada sistema.

Kao što se može videti, nakon 2 interakcije sistem u prvom modu rada veoma je dobar u predikciji sadržaja koje korisnik ne voli da gleda (velika *TP_rate* vrednost), ali je zato veoma loš u predikciji sadržaja koje korisnik želi da gleda (velika *FP_rate* vrednost). Kako su ove vrednosti skoro pa jednake čak se može reći i da se ponaša kao klasifikator koji je estimirao da se nijedan ponuđeni TV sadržaj neće svideti korisniku. Nasuprot ovome, pod istim uslovima u drugom modu rada sistema, baziranom na neuralnoj mreži s jednim skrivenim slojem treniranoj s WELM algoritmom, postižu se znatno uravnoteženije performanse. U odnosu na *ELM sparse autoenkoder*, ovaj mod funkcionisanja dostiže manje *FP_rate* i

TP_rate vrednosti, te je bolji u klasifikaciji sadržaja koje korisnik želi da gleda i nešto lošiji u klasifikaciji sadržaja koje posmatrani ne voli. Kako se broj interakcija povećava, kod oba moda dolazi do smanjenja vrednosti na *x* i *y* osi ROC grafika što rezultuje tačnijom klasifikacijom sadržaja koji se korisniku sviđaju i lošijom klasifikacijom onih koji mu se ne sviđaju. Ukoliko se nakon 10 korisničkih interakcija, sistem i dalje nalazi u prvom modu funkcionisanja postiže se bolja predikcija sadržaja koje korisnik ne želi da gleda, dok ukoliko je u drugom modu sistem će biti bolji u pronalaženju sadržaja koji se sviđaju korisniku.

Imajući u vidu da je opšti cilj personalizovanog programskog vodiča da preporuči TV sadržaje za koje je procenio da će se gledaocu svideti, na osnovu rezultata sa slike 8.1 možemo zaključiti da je pogodno preći u drugi mod rada čim se prikupi barem jedna negativna interakcija. Sistem u prvom modu rada je u mogućnosti da pruži adekvatne preporuke korisniku ali mu je za to potrebno više korisničkih interakcija. Ipak, informacija o tome šta korisnik voli a šta ne u ranoj fazi korišćenja sistema može značajno ubrzati učenje njegovih interesovanja.

Kako bi u potpunosti opravdali predloženi način funkcionisanja vodiča, ispitali smo i vreme treniranje za svaki od slučajeva. Dobijeni rezultati prikazani su na slici 8.2.



Slika 8.2: Poređenje vremena treniranja različitih modova rada.

Može se primetiti da je bez obzira na broj prikupljenih korisničkih interakcija vreme treniranja u prvom modu rada značajno duže nego u slučaju kada se koristi drugi mod funkcionisanja, te se prvobitna odluka pokazala opravdanom.

Iako smo simulacije sistema sproveli u Matlabu na desktop računaru s Intel(R) Core(TM)2 T5250 1.5GHz procesorom i 1.5 GB RAM memorije, razmotrili smo i očekivano vreme treniranja na na Texas Instruments AM3351 Starter Kit uređaju. Ovaj uređaj smo izabrali zbog toga što je po hardverskim karakteristikama sličniji onima uz čiju pomoć korisnik pristupa usluzi digitalne televizije – kao što su STB, mobilni telefon, Smart TV.

Empirijski, kroz veliki broj izvršavanja različitih proračuna, došli smo do procene da vrednosti sa slike 8.2 treba skalirati s faktorom čija vrednost varira od 1.8 do 2.2 kako bi se dobilo vreme treniranja na posmatranom uređaju.

Čak i u pesimističnom slučaju, kada su inicijalne vrednosti vremena treniranja pomnožene s faktorom 2.2, možemo bez ikakvih sumnji se može zaključiti da je predloženi personalizovani programski vodič za digitalnu televiziju pogodan za lokalnu implementaciju na korisničkim uređajima s ograničenim hardverskim resursima.

Što se tiče ostalih osobina sistema kao što su zaštita privatnosti, robusnost sistema, skalabilnost sistema, novina preporuka, i neočekivanost preporuka, na njih u mnogome utiče izabrana tehnika pružanja preporuka, kao i sama implementacija sistema.

Velika prednost izabranog rešenja za zaštitu privatnosti je ta što štiti i one korisnike koji nisu svesni opasnosti koje korišćenje personalizovanog programskog vodiča donosi, dok je najveća mana ovog sistema povećena mogućnost otuđenja podataka u slučaju krađe uređaja. Ipak treba naglasiti da se s mogućnošću krađe uređaja - posebno mobilnih telefona, korisnici svakodnevno susreću i da se na njima mogu nalaziti mnogo osetljiviji podaci od onih sakupljenih za potrebe personalizovanog vodiča.

Za razliku od vodiča baziranih na klijent-server arhitekturi kod kojih se podaci za sve korisnike prikupljaju na jednom mestu, kod predloženog lokalno implementiranog sistema ne postoji problem obrade velike količine podataka, jer lokalno prikupljeni podaci ne mogu dostići te razmere. Optimizacija sistema, u smislu korišćenja ograničenih hardverskih resursa uređaja i učenju korisničkih interesovanja u realnom vremenu, realizuje se kroz izbor odgovarajućeg algoritma učenja neuralne mreže.

Predloženi personalizovani programski vodič se može smatrati veoma robusnim na maliciozne napade koji ciljaju sam proces pružanja preporuka jer je baziran na tehnici filtriranja sadržaja koja je imuna na ovaj tip napada.

S druge strane, zbog same prirode sistema koji preporučuje sadržaje slične onima koji su se svideli korisniku u prošlosti njegovo korišćenje generalno ima negativan uticaj na novinu, neočekivanost i raznovrsnost preporuka. Izborom AG-mean metrike, u slučajevima kada među prikupljenim interakcijama postoje i one o sadržajima koje korisnik ne voli da gleda mogu se u određenoj meri popraviti za sistem koji je izašao iz hladnog starta. Ipak, značajnije poboljšanje ovih osobina može se postići jedino primenom odgovarajućih metoda za formiranje liste preporuka.

Pokrivenost TV sadržaja u smislu da svi korisnici, pa čak i oni za koje sistem ima samo prikupljene podatke o tome šta vole da gledaju, mogu da koriste personalizovani programski vodič postiže se korišćenjem prvog moda rada baziranog na autoenkoder mreži - koji je posebno projektovan za ovakve situacije.

Što se tiče adaptivnosti preporuka, jedna od pretpostavki našeg istraživanja je da su interesovanja korisnika sporo promenljiva, te je predloženi sistem, baziran na neuralnoj mreži, sposoban da detektuje i nauči nova interesovanja korisnika, kao i da prepozna ona koja nisu više aktuelna na osnovu prikupljenih interakcija s korisnikom. Imajući u vidu kratko vreme treniranja predloženog sistema, ukoliko postoji dovoljna količina informacija, on će se vrlo brzo adaptirati na promene u interesovanjima.

Posebna pažnja posvećena je i povećanju poverenja korisnika za vreme hladnog starta sistema, jer se po našem mišljenju, ono tada i formira. Kao i kod novine, neočekivanosti i raznovrsnosti preporuka, poboljšanje ove osobine postiže se korišćenjem AG-mean metrike.

Pouzdanost preporuka i rizik njihovog prihvatanja nisu detaljnije razmatrani, ali se lako mogu uključiti kao poboljšanja predloženog personalizovanog vodiča, te su ostavljena za buduća istraživanja.

8.1 Rezime

U ovom poglavlju smo analizirali performanse koje se postižu korišćenjem predloženog personalizovanog programskog vodiča i diskutovali njegove osobine.

U sledećem ćemo prikazati zaključke proizašle iz teorijskih i eksperimentalnih istraživanja predstavljenih u ovoj doktorskoj disertaciji i predložiti pravce daljeg razvoja.

9. Zaključak

U doktorskoj disertaciji predstavio sam holistički pristup projektovanju personalizovanih programskih vodiča za digitalnu televiziju. Već u fazi projektovanja u obzir su uzete zaštita privatnosti korisnika, mogućnosti rada sistema na uređajima s ograničenim resursima, izbor i način korišćenja kontekstualnih informacija, uticaj nejednake količine podataka u pojedinačnim klasama i izbor mera performansi.

Predloženi sistem, baziran na veštačkim neuralnim mrežama, u mogućnosti je da modelira različite načine na koje korisnici donose odluke o izboru TV sadržaja. Kontekstualne informacije izabrane su tako da pomognu u razumevanju njihovog ponašanja, ali i da ne ometaju uobičajen način gledanja televizije. Posebna pažnja posvećana je izboru algoritma učenja kako bi vodič mogao da se implementira lokalno na korisničkim uređajima s ograničenim resursima (STB, Smart TV, mobilni telefon) bez ometanja njihovih ostalih funkcionalnosti. Ovo treba posebno naglasiti, jer u vreme sprovođenja eksperimenata iz ove oblasti nije postojao sličan sistem, dok u današnje vreme premeštanje bilo celog proračuna ili njegovog dela ka korisničkim uređajima postaje trend. Kako izbor mera performansi direktno zavisi od ciljeva pružaoca usluge distribucije medijskih sadržaja, koji se mogu razlikovati od slučaja do slučaja, pri proceni performansi u obzir smo uzeli samo uticaj preostalih faktora, s posebnim osvrtom na problem disbalans klasa. Ovaj problem koji nastaje usled tendencije korisnika da znatno češće pružaju informacije o sadržajima koje vole da gledaju nego o onima koje ne vole može značajno degradirati performanse. Predloženi sistem funkcioniše u dva moda rada kako bi u potpunosti mogao da prevaziđe ovaj problem i pruži preporuke čak i onim korisnicima za koje ne postoje podaci o TV sadržajima koji im se ne sviđaju.

Ipak, iako je velika pažnja posvećena skalabilnosti sistema i mogućnosti rada na uređajima s ograničenim resursima, prilikom prikupljanja novih podataka korišćeni algoritmi zahtevaju ponovno treniranje neuralne mreže sa svim do tada prikupljenim informacijama. Kako ovo, za veliki broj prikupljenih interakcija, može degradirati vreme treniranja, za buduća istraživanja ostavili smo sekvencijalne algoritme koji pružaju mogućnost treniranja samo s novoprikupljenim podacima.

S druge strane, kako u performanse vodiča, prikazane u disertaciji, nisu uključeni konkretni ciljevi pružaoca usluge distribucije medijskih sadržaja, u budućim eksperimentima neophodno

ih je detaljnije istražiti u realnim uslovima – s jasno definisanim ciljevima i korisnicima koji u realnom vremenu koriste predloženi sistem.

Literatura

[1] "RTS :: Šta donosi digitalizacija, koje su prednosti?" [Online]. Available: http://www.rts.rs/page/rts/sr/Digitalizacija/story/2010/vodici-za-dtv/1789008/sta-donosi-digitalizacija-koje-su-prednosti.html. [Accessed: 18-Nov-2018].

[2] "TV paketi | EON TV | SBB." [Online]. Available: https://sbb.rs/eon-tv/tv-paketi. [Accessed: 18-Nov-2018].

[3] E. T. S. Institute, "Hybrid Broadcast Broadband TV Technical Specification," 2012.

[4] "YouTube by the Numbers (2018): Stats, Demographics and Fun Facts." [Online]. Available: https://www.omnicoreagency.com/youtube-statistics/. [Accessed: 18-Nov-2018].

[5] D. Véras, T. Prota, A. Bispo, R. Prudêncio, and C. Ferraz, "A literature review of recommender systems in the television domain," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 9046–9076, Dec. 2015.

[6] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decis. Support Syst.*, vol. 74, pp. 12–32, Jun. 2015.

[7] L. P. J. (Leo P. J. . Veelenturf, *Analysis and applications of artificial neural networks*. London, U.K.: Prentice Hall, 1995.

[8] T. K. Paradarami, N. D. Bastian, and J. L. Wightman, "A hybrid recommender system using artificial neural networks," *Expert Syst. Appl.*, vol. 83, pp. 300–313, Oct. 2017.

[9] B. Barragáns-Martínez, E. Costa-Montenegro, and J. Juncal-Martínez, "Developing a recommender system in a consumer electronic device," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4216–4228, Jun. 2015.

[10] G. Danezis, J. Domingo-Ferrer, M. Hansen, J.-H. Hoepman, D. Le Métayer, R. Tirtea, and S. Schiffner, *Privacy and data protection by design*, no. December. 2014.

[11] Q. YANG and X. WU, "10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH," *Int. J. Inf. Technol. Decis. Mak.*, vol. 05, no. 04, pp. 597–604, Dec. 2006.

[12] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee, “AIMED- A Personalized TV Recommendation System,” in *Interactive TV: a Shared Experience*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 166–174.

[13] G. Adomavicius and A. Tuzhilin, “Context-Aware Recommender Systems,” in *Recommender Systems Handbook*, Boston, MA: Springer US, 2015, pp. 191–226.

[14] M. Bjelica Elektrotehnički fakultet Beograd Srbija FER, “Preporučitelji sadržaja.”

[15] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[16] M. Pazzani and D. Billsus, “Learning and Revising User Profiles: The Identification of Interesting Web Sites,” *Mach. Learn.*, vol. 27, no. 3, pp. 313–331, 1997.

[17] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, “Content-Based Video Recommendation System Based on Stylistic Visual Features,” *J. Data Semant.*, vol. 5, no. 2, pp. 99–113, Jun. 2016.

[18] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, “Item-based collaborative filtering recommendation algorithms,” in *Proceedings of the tenth international conference on World Wide Web - WWW ’01*, 2001, pp. 285–295.

[19] Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization Techniques for Recommender Systems,” *Computer (Long. Beach. Calif).*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[20] V. Klema and A. Laub, “The singular value decomposition: Its computation and some applications,” *IEEE Trans. Automat. Contr.*, vol. 25, no. 2, pp. 164–176, Apr. 1980.

[21] D. Bokde, S. Girase, and D. Mukhopadhyay, “Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey,” *Procedia Comput. Sci.*, vol. 49, pp. 136–146, Jan. 2015.

[22] M. J. Pazzani, “A Framework for Collaborative, Content-Based and Demographic Filtering,” *Artif. Intell. Rev.*, vol. 13, no. 5/6, pp. 393–408, 1999.

[23] I. S. C. Nicholas, I. S. C. Nicholas, and C. K. Nicholas, “Combining Content and

Collaboration in Text Filtering," *Proc. IJCAI'99 Work. Mach. Learn. Inf. Filter.*, pp. 86--91, 1999.

[24] W. Kogel, "Faster Training of Neural Networks for Recommender Systems," *Masters Theses (All Theses, All Years)*, May 2002.

[25] G. Adomavicius and A. Tuzhilin, "Multidimensional Recommender Systems: A Data Warehousing Approach," Springer, Berlin, Heidelberg, 2001, pp. 180–192.

[26] G. Schröder, G. Schröder, and et al., "Setting Goals and Choosing Metrics for Recommender System Evaluations," in *User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERTI 2), Workshop of 5th ACM Recommender Systems conference (RecSys 2011)*, 2011.

[27] P. Yang, P. Zhao, Y. Liu, and X. Gao, "Robust Cost-Sensitive Learning for Recommendation with Implicit Feedback," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018, pp. 621–629.

[28] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," in *Recommender Systems Handbook*, Boston, MA: Springer US, 2011, pp. 257–297.

[29] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[30] R. Barandela, J. . Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognit.*, vol. 36, no. 3, pp. 849–851, Mar. 2003.

[31] R. Batuwita and V. Palade, "A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems," in *2009 International Conference on Machine Learning and Applications*, 2009, pp. 545–550.

[32] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "A Survey on Graphical Methods for Classification Predictive Performance Evaluation," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1601–1618, Nov. 2011.

[33] Y. Y. Yao, "Measuring retrieval effectiveness based on user preference of documents," *J. Am. Soc. Inf. Sci.*, vol. 46, no. 2, pp. 133–145, Mar. 1995.

[34] M. G. Kendall, “A New Measure of Rank Correlation,” *Biometrika*, vol. 30, no. 1/2, p. 81, Jun. 1938.

[35] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” Jan. 2013.

[36] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, “A Theoretical Analysis of NDCG Type Ranking Measures,” in *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, 2013, pp. 25–54.

[37] D. Braziunas and C. Boutilier, “Local Utility Elicitation in GAI Models,” in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI2005)*, 2005, pp. 42–49.

[38] SWEARINGEN and K., “Beyond algorithms : An HCI perspective on recommender systems,” *ACM SIGIR 2001Workshop Recomm. Syst.*, 2001.

[39] J. L. Herlocker, J. A. Konstan, and J. Riedl, “Explaining collaborative filtering recommendations,” in *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW ’00*, 2000, pp. 241–250.

[40] P. Pu and L. Chen, “Trust building with explanation interfaces,” in *Proceedings of the 11th international conference on Intelligent user interfaces - IUI ’06*, 2006, p. 93.

[41] KONSTAN and J. A., “Lessons on Applying Automated Recommender Systems to Information-Seeking Tasks,” *Proc. Twenty-First Natl. Conf. Artif. Intell. 2006*, pp. 1630–1633, 2006.

[42] M. Bjelica, “Unobtrusive relevance feedback for personalized TV program guides,” *IEEE Trans. Consum. Electron.*, vol. 57, no. 2, pp. 658–663, May 2011.

[43] G. Shani, M. Chickering, and C. Meek, “Mining recommendations from the web,” in *Proceedings of the 2008 ACM conference on Recommender systems - RecSys ’08*, 2008, p. 35.

[44] M. Zhang and N. Hurley, “Avoiding monotony,” in *Proceedings of the 2008 ACM conference on Recommender systems - RecSys ’08*, 2008, p. 123.

[45] L. Chen, W. Wu, and L. He, “Personality and Recommendation Diversity,” in

Emotions and Personality in Personalized Services, Models, Evaluation and Applications, M. Tkalčić et al, Ed. Springer, Cham, 2016, pp. 201–225.

[46] D. Bouneffouf, A. Bouzeghoub, and A. L. Ganarski, "Risk-Aware Recommender Systems," in *Neural Information Processing. ICONIP 2013. Lecture Notes in Computer Science vol 8226*, K. R. M. Lee M., Hirose A., Hou ZG., Ed. Springer, Berlin, Heidelberg, 2013, pp. 57–65.

[47] E. Tan, I. Seaman, H. Leung, and Y.-K. Ng, "Making personalized movie recommendations for children," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services - iiWAS '16*, 2016, pp. 96–105.

[48] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre, "Collaborative recommendation," *ACM Trans. Internet Technol.*, vol. 4, no. 4, pp. 344–377, Nov. 2004.

[49] I. Koychev, I. Koychev, and I. Schwab, "Adaptation to drifting user's interests," *Proc. ECML2000 Work. Mach. Learn. NEW Inf. AGE*, pp. 39--46, 2000.

[50] J. Möller, D. Trilling, N. Helberger, and B. van Es, "Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity," *Information, Commun. Soc.*, vol. 21, no. 7, pp. 959–977, Jul. 2018.

[51] T. George and S. Merugu, "A Scalable Collaborative Filtering Framework Based on Co-Clustering," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 625–628.

[52] J. Masthoff, "Group Recommender Systems: Aggregation, Satisfaction and Group Attributes," in *Recommender Systems Handbook*, Boston, MA: Springer US, 2015, pp. 743–776.

[53] N. F. Awad and M. S. Krishnan, "The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization," *MIS Q.*, vol. 30, no. 1, p. 13, 2006.

[54] D. Krivokapić, Đ. Krivokapić, I. Todorović, S. Komazec, A. Petrovski, and K. Ercegović, "Zaštita Podataka O Ličnosti Vodič Za Organe Vlasti," 2016.

[55] “Zakon o zaštiti podataka o ličnosti,” *Sl. glasnik RS", 68/2012 - odluka US i 107/2012*, 2009. [Online]. Available: https://www.paragraf.rs/propisi/zakon_o_zastiti_podataka_o_licnosti.html. [Accessed: 17-Nov-2018].

[56] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, “Information security in big data: Privacy and data mining,” *IEEE Access*, vol. 2, pp. 1151–1178, 2014.

[57] R. M. Arlein, B. Jai, M. Jakobsson, F. Monrose, and M. K. Reiter, “Privacy-preserving global customization,” in *Proceedings of the 2nd ACM conference on Electronic commerce - EC ’00*, 2000, pp. 176–184.

[58] B. P. Knijnenburg and A. Kobsa, “Making Decisions about Privacy,” *ACM Trans. Interact. Intell. Syst.*, vol. 3, no. 3, pp. 1–23, Oct. 2013.

[59] L. Beckett, “Big Data Brokers: They Know Everything About You and Sell it to the Highest Bidder.” [Online]. Available: https://gizmodo.com/5991070/big-data-brokers-they-know-everything-about-you-and-sell-it-to-the-highest-bidder. [Accessed: 17-Nov-2018].

[60] A. Friedman, B. P. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky, “Privacy Aspects of Recommender Systems,” in *Recommender Systems Handbook*, Boston, MA: Springer US, 2015, pp. 649–688.

[61] U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft, “BlurMe,” in *Proceedings of the sixth ACM conference on Recommender systems - RecSys ’12*, 2012, p. 195.

[62] C. Duhigg, “How Companies Learn Your Secrets - The New York Times.” [Online]. Available: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html. [Accessed: 17-Nov-2018].

[63] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris, “Detecting price and search discrimination on the internet,” in *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI*, 2012, pp. 79–84.

[64] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, “You Might Also Like: Privacy Risks of Collaborative Filtering,” in *2011 IEEE Symposium on Security and Privacy*, 2011, pp. 231–246.

[65] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125.

[66] R. Cissée, R. Cissée, V. Fakultät, I. E. Informatik, and Z. E. Des Grads, “An agent-based approach for privacypreserving recommender systems,” *PROC. 6TH INT. Jt. CONF. Auton. AGENTS MULTIAGENT Syst. (AAMAS’07*, pp. 1--8, 2007.

[67] A. Put, I. Dacosta, M. Milutinovic, B. De Decker, S. Seys, F. Boukayoua, V. Naessens, K. Vanhecke, T. De Pessemier, and L. Martens, “inShopnito: An Advanced yet Privacy-Friendly Mobile Shopping Application,” in *2014 IEEE World Congress on Services*, 2014, pp. 129–136.

[68] B. Heitmann, J. G. Kim, A. Passant, C. Hayes, and H.-G. Kim, “An architecture for privacy-enabled user profile portability on the web of data,” in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems - HetRec ’10*, 2010, pp. 16–23.

[69] N. Lathia, S. Hailes, and L. Capra, “Private distributed collaborative filtering using estimated concordance measures,” in *Proceedings of the 2007 ACM conference on Recommender systems - RecSys ’07*, 2007, p. 1.

[70] D. Vallet, A. Friedman, and S. Berkovsky, “Matrix Factorization without User Data Retention,” Springer, Cham, 2014, pp. 569–580.

[71] V. Ciriani, S. De Capitani Di Vimercati, S. Foresti, and P. Samarati, “k-Anonymity,” *Adv. Inf. Secur.*, pp. 1–36, 2007.

[72] P. Samarati, “Protecting respondents identities in microdata release,” *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.

[73] F. McSherry and I. Mironov, “Differentially private recommender systems,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09*, 2009, p. 627.

[74] K. Nissim, S. Raskhodnikova, and A. Smith, “Smooth sensitivity and sampling in private data analysis,” in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing - STOC ’07*, 2007, p. 75.

[75] P. Paillier, “Public-Key Cryptosystems Based on Composite Degree Residuosity Classes,” in *Advances in Cryptology — EUROCRYPT ’99*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 223–238.

[76] A. Basu, J. Vaidya, H. Kikuchi, and T. Dimitrakos, “Privacy-Preserving Collaborative Filtering on the Cloud and Practical Implementation Experiences,” in *2013 IEEE Sixth International Conference on Cloud Computing*, 2013, pp. 406–413.

[77] J. Canny, “Collaborative filtering with privacy,” in *Proceedings 2002 IEEE Symposium on Security and Privacy*, pp. 45–57.

[78] A. Cavoukian and M. Crompton, *Web seals: review of online privacy programs*. Ontario, 2000.

[79] H. Xu, H.-H. Teo, and B. Tan, “Predicting the Adoption of Location-Based Services: The Role of Trust and Perceived Privacy Risk,” in *ICIS 2005 Proceedings*, 2005.

[80] M. Schunter, V. Herreweghen, and M. Waidner, “Expressive Privacy Promises-How to Improve the Platform for Privacy Preferences (P3P),” 2002.

[81] A. K. Dey and A. K., “Understanding and Using Context,” *Pers. Ubiquitous Comput.*, vol. 5, no. 1, pp. 4–7, Feb. 2001.

[82] B. Fling, *Mobile design and development*. O’Reilly Media, 2009.

[83] C. Palmisano, C. Palmisano, E. Tuzhilin, and M. Gorgoglione, “Using context to improve predictive modeling of customers in personalization applications,” in *KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON*, 2008, p. 1549.

[84] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, “Incorporating contextual information in recommender systems using a multidimensional approach,” *ACM Trans. Inf. Syst.*, vol. 23, no. 1, pp. 103–145, Jan. 2005.

[85] P. Dourish, “What we talk about when we talk about context,” *Pers. Ubiquitous Comput.*, vol. 8, no. 1, pp. 19–30, Feb. 2004.

[86] U. Panniello and M. Gorgoglione, “Incorporating context into recommender systems: an empirical comparison of context-based approaches,” *Electron. Commer. Res.*, vol.

12, no. 1, pp. 1–30, Mar. 2012.

[87] K. Oku, S. Nakajima, J. Miyazaki, and S. Uemura, “Context-Aware SVM for Context-Dependent Information Recommendation,” in *7th International Conference on Mobile Data Management (MDM’06)*, 2006, pp. 109–109.

[88] D. Zillmann, R. T. Hezel, and N. J. Medoff, “The Effect of Affective States on Selective Exposure to Televised Entertainment Fare1,” *J. Appl. Soc. Psychol.*, vol. 10, no. 4, pp. 323–339, Aug. 1980.

[89] U. Panniello, A. Tuzhilin, and M. Gorgoglione, “Comparing context-aware recommender systems in terms of accuracy and diversity,” *User Model. User-adapt. Interact.*, vol. 24, no. 1–2, pp. 35–65, Feb. 2014.

[90] Zhiwen Yu, Xingshe Zhou, Daqing Zhang, Chung-Yau Chin, Xiaohang Wang, and Ji Men, “Supporting Context-Aware Media Recommendations for Smart Phones,” *IEEE Pervasive Comput.*, vol. 5, no. 3, pp. 68–75, Jul. 2006.

[91] M. Aharon, E. Hillel, A. Kagian, R. Lempel, H. Makabee, and R. Nissim, “Watch-It-Next: A Contextual TV Recommendation System,” Springer, Cham, 2015, pp. 180–195.

[92] M. Krstić, “Personalizovani elektronski programski vodič,” Univerzitet u Beogradu, 2012.

[93] L. Deng and D. Yu, “Deep Learning: Methods and Applications,” *Found. Trends® Signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2014.

[94] S. Nagabhushana, *Data warehousing : OLAP and data mining*. New Age International, 2006.

[95] B. Batbayar, “Improving time efficiency of feedforward neural network learning,” RMIT University, 2008.

[96] R. Pagano, P. Cremonesi, M. Larson, B. Hidasi, D. Tikk, A. Karatzoglou, and M. Quadrana, “The Contextual Turn,” in *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys ’16*, 2016, pp. 249–252.

[97] C. Christakou and A. Stafylopatis, “A hybrid movie recommender system based on

neural networks," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, 2005, pp. 500–505.

[98] K. Potdar, T. S., and C. D., "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *Int. J. Comput. Appl.*, vol. 175, no. 4, pp. 7–9, Oct. 2017.

[99] Irie and Miyake, "Capabilities of three-layered perceptrons," in *IEEE International Conference on Neural Networks*, 1988, pp. 641–648 vol.1.

[100] S. Lawrence, C. Lee Giles, and A. C. Tsoi, "Lessons in neural network training: overfitting may be harder than expected," in *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, 1997, pp. 540–545.

[101] T. Isobe, M. Fujiwara, H. Kaneta, T. Morita, and N. Uratani, "Development of TV reception system personalized with viewing habits," *IEEE Trans. Consum. Electonics*, vol. 51, no. 2, pp. 665–674, 2005.

[102] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," in *IEEE International Conference on Neural Networks*, pp. 586–591.

[103] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, Jan. 1993.

[104] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[105] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, Dec. 2006.

[106] J. Nocedal and S. J. Wright, "Line Search Methods," in *Numerical Optimization*, Springer New York, 1999, pp. 30–65.

[107] G.-B. Huang and H. A. Babri, "Upper Bounds on the Number of Hidden Neurons in Feedforward Networks with Arbitrary Bounded Nonlinear Activation Functions," *IEEE Trans. NEURAL NETWORKS*, vol. 9, no. 1, pp. 224–229, 1998.

[108] P. L. Bartlett, “The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network,” *IEEE Trans. Inf. THEORY*, vol. 44, no. 2, p. 525, 1998.

[109] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[110] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[111] M. Bjelica and A. Peric, “Adaptive feedback schemes for personalized content retrieval,” *IEEE Trans. Consum. Electron.*, vol. 57, no. 3, pp. 1251–1257, Aug. 2011.

[112] D. M. J. Tax, “One-class classification Concept-learning in the absence of counter-examples,” Delft University of Technology, 2001.

[113] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[114] Zhi-Hua Zhou and Xu-Ying Liu, “Training cost-sensitive neural networks with methods addressing the class imbalance problem,” *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.

[115] M. Kukar, M. Kukar, and I. Kononenko, “Cost-Sensitive Learning with Neural Networks,” *Proc. 13TH Eur. Conf. Artif. Intell. (ECAI-98*, pp. 445--449, 1998.

[116] Y. Yajima, “One-Class Support Vector Machines for Recommendation Tasks,” Springer, Berlin, Heidelberg, 2006, pp. 230–239.

[117] L. Manevitz and M. Yousef, “One-class document classification via Neural Networks,” *Neurocomputing*, vol. 70, no. 7–9, pp. 1466–1481, Mar. 2007.

[118] D. M. J. Tax and K.-R. Muller, “A consistency-based model selection for one-class classification,” in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004, p. 363–366 Vol.3.

[119] “GroupLens Research Project (online).” [Online]. Available: https://grouplens.org/. [Accessed: 17-Nov-2018].

[120] M. Bjelica, “Towards TV recommender system: experiments with user modeling,” *IEEE Trans. Consum. Electron.*, vol. 56, no. 3, pp. 1763–1769, Aug. 2010.

[121] A. Košir, A. Odi´codi´c, M. Kunaver, M. Tkalčič, and J. F. Tasič, “Database for contextual personalization,” *Elektroteh. Vestn.*, vol. 78, no. 5, pp. 270–274, 2011.

[122] M. Krstic and M. Bjelica, “Context-aware personalized program guide based on neural network,” *IEEE Trans. Consum. Electron.*, vol. 58, no. 4, pp. 1301–1306, Nov. 2012.

[123] M. Krstic and M. Bjelica, “Impact of class imbalance on personalized program guide performance,” *IEEE Trans. Consum. Electron.*, vol. 61, no. 1, pp. 90–95, Feb. 2015.

[124] M. Krstic and M. Bjelica, “Personalized program guide based on one-class classifier,” *IEEE Trans. Consum. Electron.*, vol. 62, no. 2, pp. 175–181, May 2016.

[125] M. Usman, I. Ahmed, M. I. Aslam, S. Khan, and U. A. Shah, “SIT: A Lightweight Encryption Algorithm for Secure Internet of Things,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 1, Apr. 2017.

[126] W. Zong, G.-B. Huang, and Y. Chen, “Weighted extreme learning machine for imbalance learning,” *Neurocomputing*, vol. 101, pp. 229–242, Feb. 2013.

[127] M. Krstić and M. Bjelica, “Personalizovani vodič za izbor multimedijalnih sadržaja,” in *Zbornik radova s konferencije ETRAN 2016*, 2016, p. TE1.2.1. 1-5.

[128] L. Lekamalage, C. Kasun, H. Zhou, G.-B. Huang, and C. M. Vong, “Representational Learning with Extreme Learning Machine for Big Data,” *IEEE Intell. Syst.*, vol. 28, no. 6, pp. 31–34, 2013.

[129] J. Tang, C. Deng, and G.-B. Huang, “Extreme Learning Machine for Multilayer Perceptron,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 4, pp. 809–821, Apr. 2016.

[130] A. Beck and M. Teboulle, “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems,” *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[131] M. Krstic and M. Bjelica, “Performance metrics for personalized program guides,” in *2016 13th Symposium on Neural Networks and Applications (NEUREL)*, 2016, pp. 1–5.

Biografija autora

Marko Krstić je rođen 1.11.1986. godine u Leskovcu. U svom rodnom gradu je završio osnovnu školu "Vasa Pelagić" i prirodno-matematički smer Gimnazije sa odličnim uspehom. Godine 2005. upisuje osnovne studije na Elektrotehničkom fakultetu u Beogradu, koje uspešno završava 2010. godine sa prosečnom ocenom 8.73 na odseku za Telekomunikacije i informacione tehnologije – smer Sistemsko inženjerstvo. Diplomski rad pod nazivom "Carrier Ethernet: arhitektura, standardi i interoperabilnost sa GMPLS tehnologijom" iz predmeta Širokopojasne telekomunikacione mreže je odbranio sa ocenom 10. Godine 2010. upisao je master studije Elektrotehničkog fakulteta u Beogradu na smeru Sistemsko inženjerstvo i radio komunikacije, koje uspešno završava 2012. godine sa prosečnom ocenom 9.67. Master rad pod nazivom "Personalizovani elektronski programski vodič" iz predmeta Personalizacija telekomunikacionih servisa je odbranio sa ocenom 10. Sa istraživanjem ove oblasti nastavlja i na doktorskim studijama na modulu Telekomunikacije koji upisuje 2013. godine. Do sada je na ovu temu objavio 3 rada u međunarodnim časopisima sa SCI liste, 2 rada na međunarodnim konferencijama i 1 rad na domaćoj konferenciji. Rezultate dosadašnjeg istraživanja prezentovao je i široj javnosti 2016. godine na Data Science konferenciji u Beogradu. Istovremeno sa doktorskim studijama, 2013. godine počinje da radi u IT odseku Regulatorne agencije za elektronske komunikacije i poštanske usluge (RATEL), gde je i dan danas zaposljen. Godine 2014. bio je učesnik EYE (*Empowering Young Explorers*) Lab Surfing konferencije u Beogradu, i Blue Sky konferencije u Budimpešti, koje su bile posvećene inovacijama i edukaciji o Horizon 2020 projektima. Pohađao je nekoliko kurseva iz oblasti računarskih mreža i baza podataka, a poseduje Cisco CCNA (*Cisco Certified Network Associate*), Microsoft MCSA (*Microsoft Certified Solutions Associate*) SQL 2012 i EMC *Data Science Associate* sertifikate. Pored istraživačkog rada, svoj doprinos društvu daje i kroz aktivno učešće u ostalim aktivnostima naučne i stručne zajednice. Bio je član tima na bilateranom projektu između Srbije i Slovenije pod nazivom *Electro-Active polyHIPE Polyelectrolytes*. Kao recenzent pomogao je u održavanju *The 6th International Conference on Biomedical Engineering and Biotechnology* (ICBEB 2017) i *7th International Conference on Electronics, Communications and Networks* (CECNet2017). Učestvovao je na *ACM Summer School on Recommender Systems* održanoj 2017. godine u Italiji – letnjoj školi posvećenoj pregledu aktuelnih trendova iz oblasti istraživanja njegove doktorske disertacije. S druge stane prati aktuelnosti i globalne trendove u oblasti sajber bezbednosti, te je bio učesnik

nekoliko međunarodnih sajber vežbi. Govori engleski jezik-srednji nivo i francuski jezik-početni nivo. Član je IEEE (*Institute of Electrical and Electronics Engineers*) udruženja i Udruženja inženjera elektrotehnike Srbije (UDIES).

Izjava o autorstvu

Potpisani Marko Ž. Krstić
broj upisa 5032/2013

Izjavljujem

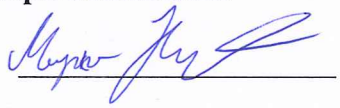
da je doktorska disertacija pod naslovom

Personalizovani programski vodiči za digitalnu televiziju

- rezultat sopstvenog istraživačkog rada,
- da predložena disertacija u celini ni u delovima nije bila predložena za dobijanje bilo koje diplome prema studijskim programima drugih visokoškolskih ustanova,
- da su rezultati korektno navedeni i
- da nisam kršio autorska prava i koristio intelektualnu svojinu drugih lica.

U Beogradu, 11.12.2018.

Potpis doktoranda



Prilog 1

Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Ime i prezime autora: Marko Ž. Krstić

Broj upisa: 5032/2013

Studijski program: Telekomunikacije

Naslov rada: Personalizovani programski vodiči za digitalnu televiziju

Mentor: dr Milan Bjelica, vanredni profesor

Potpisani Marko Ž. Krstić

izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao za objavljivanje na portalu **Digitalnog repozitorijuma Univerziteta u Beogradu**.

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

U Beogradu, 11.12.2018.

Potpis doktoranda

Prilog 2

Izjava o korišćenju

Ovlašćujem Univerzitetsku biblioteku „Svetozar Marković" da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

Personalizovani programski vodiči za digitalnu televiziju

koja je moje autorsko delo.

Disertaciju sa svim prilozima predao sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalni repozitorijum Univerziteta u Beogradu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio.

1. Autorstvo
2. Autorstvo – nekomercijalno
3. Autorstvo – nekomercijalno – bez prerade
4. Autorstvo – nekomercijalno – deliti pod istim uslovima
5. Autorstvo – bez prerade
6. Autorstvo – deliti pod istim uslovima

(Molimo da zaokružite samo jednu od šest ponuđenih licenci, kratak opis licenci dat je na poleđeni lista).

U Beogradu, 11.12.2018.

Potpis doktoranda

1. Autorstvo – Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence, čak i u komercijalne svrhe. Ovo je najslobodnija od svih licenci.

2. Autorstvo – nekomercijalno. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca ne dozvoljava komercijalnu upotrebu dela.

3. Autorstvo – nekomercijalno – bez prerade. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, bez promena, preoblikovanja ili upotrebe dela u svom delu, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca ne dozvoljava komercijalnu upotrebu dela. U odnosu na sve ostale licence, ovom licencom se ograničava najveći obim prava korišćenja dela.

4. Autorstvo – nekomercijalno – deliti pod istim uslovima. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence i ako se prerada distribuira pod istom ili sličnom licencom. Ova licenca ne dozvoljava komercijalnu upotrebu dela i prerada.

5. Autorstvo – bez prerade. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, bez promena, preoblikovanja ili upotrebe dela u svom delu, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca dozvoljava komercijalnu upotrebu dela.

6. Autorstvo – deliti pod istim uslovima. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence i ako se prerada distribuira pod istom ili sličnom licencom. Ova licenca dozvoljava komercijalnu upotrebu dela i prerada. Slična je softverskim licencama, odnosno licencama otvorenog koda.