ELEKTROTEHNIŠKI VESTNIK 78(5): 270–274, 2011
EXISTING SEPARATE ENGLISH EDITION

Podatkovna baza za kontekstualno personalizacijo

Andrej Košir[1], **Ante Odić**[1], **Matevž Kunaver**[1], **Marko Tkalčič**[1], **Jurij F. Tasič**[1]

[1]*Faculty of Electrical Engineering, University of Ljubljana,*
Tržaška 25, 1000 Ljubljana, Slovenia

† *E-mail: andrej.kosir@fe.uni-lj.si*

Povzetek. V zadnjih letih opažamo, da se je razvoj personaliziranih aplikacij osredotočil na uporabo kontekstualnih informacij, torej informacij o situaciji, v kateri uporabnik dostopa do ponujenih vsebin. Težava, na katero naletijo raziskovalci na tem področju, pa je v tem, da ne obstaja javno dostopna podatkovna baza, ki bi vsebovala potrebne podatke za reševanje odprtih problemov s področja kontekstualnih opisov. Glavni razlogi za tako stanje so številne težave, ki nastopijo pri zbiranju kontekstualnih informacij o uporabniku. V prispevku predstavljamo javno dostopno podatkovno bazo, ki vsebuje podatke primerne za raziskave s področja kontekstualne personalizacije. Podatki v bazi so bili pridobljeni s sodelovanjem uporabnikov, ki so si ogledali izbrani film in takoj zatem v sistem vnesli kontekstualne podatke o situaciji, v kateri so bili med ogledom filma. Poleg opisa podatkov prispevek vsebuje tudi osnovno statistiko in izbrane lastnosti potencialno kontekstualnih spremenljivk. V trenutku oddaje je podatkovna baza vsebovala 12 kontekstualnih spremenljivk, več kot 90 uporabnikov, 950 vsebin in 1600 ocen vsebin. Podatke smo zbirali s pomočjo namenske spletne aplikacije, ki je javno dostopna in preko katere še vedno poteka aktiven zajem podatkov. Podatke o filmskih vsebinah dodatno razširimo z metapodatki, pridobljenimi iz javno dostopnih podatkovnih baz.

Ključne besede: Kontekstualna personalizacija, testna množica, načrt poskusa

Database for contextual personalization

In recent years, research into user centric and personalized applications has focused on the utilization of contextual information about the situation in which the user is consuming the content item. However, there is no database suitable for the investigation of specific open issues of contextual information description and utilization available today. The reason for this are several known difficulties with user related contextual information acquisition. A description of a publicly available database for personalization and user adaptation including contextual information is given in the paper. The data was acquired from users watching movies and then providing contextual information about the event in addition to submitting ratings about the movie. Beside describing raw data, the paper outlines basic statistics and selected properties of potentially contextual variables. At the time of submission, the database included 12 contextual variables, more than 90 users, 950 items and 1600 ratings. Data was acquired using a dedicated web application which is still publicly available and the data acquisition is still in progress. Content items (movies) can be enhanced by content item metadata using publicly available databases.

1 UVOD

S stališča uporabnika predstavlja veliko število ter tudi hitra rast števila uporabniku dostopnih multimedijskih vsebin in storitev velik problem. Delo in upravljanje z modernimi komunikacijskimi sistemi ter vsebinami, ki jih nudijo, je še vedno neprijetno oziroma v nekaterih primerih za širšo množico uporabnikov celo neizvedljivo. Osnovni pristop k reševanju tega problema je personalizacija in prilagajanje uporabnikom. Večina pristopov s tega področja temelji na napovedovanju uporabnikovih dejanj, najpogosteje ocen, ki jih bo uporabnik dodelil določeni vsebini.

V zadnjih letih se uveljavlja kontekstualna personalizacija, saj kontekst vpliva na način, kako uporabnik dostopa do izbrane vsebine in kakšne odločitve sprejme v zvezi z njo. Kontekstualni podatki lahko vsebujejo tako informacije o situaciji kot tudi informacije o uporabnikovem stanju, ko dostopa do vsebine [1]. Kontekst je tako lahko podatek o času, vremenu, socialnem statusu, razpoloženju itd. [2], [3]. Vendar pa je povezava med kontekstom in dejanskimi uporabnikovimi odločitvami zelo kompleksna in kot takšna tudi zahtevna za modeliranje. Prav tako pa je težavno zbrati kontekstualne podatke, saj proces zbiranja le-teh pogosto moti proces odločanja in lahko tako vpliva ali celo uniči zbrane kontekstualne podatke.

Za čim boljši zajem kontekstualnih podatkov je potrebno proces zajema približati uporabniku. Primerna osnova za to so filmske vsebine, ki jih uporabniki (zlasti mlajša generacija) pogosto gledajo na osebnih računalnikih. Prednost naše podatkovne baze je v tem, da so kontekstualni podatki, ki jih vsebuje, zajeti med samim dostopanjem do vsebine in so zato bolj zanesljivi. Poleg tega smo zajeli 12 različnih tipov (potencialnih) kontekstualnih podatkov, kar omogoča obširne raziskave

Prejet 9. december, 2011
Odobren 16. december, 2011

več odprtih problemov s področja kontekstualne personalizacije. A priori analiza statistične moči kaže, da predstavljena podatkovna baza zadošča osnovnim zahtevam glede velikosti (števila uporabnikov, ocen itd.).

V prispevku predstavljamo podatkovno bazo za razvoj aplikacij na področju kontekstualne personalizacije. Podali bomo osnovne statistične podatke o izbranih spremenljivkah. Proces zajema podatkov je še vedno v teku, saj je spletna aplikacija še vedno dostopna. Menimo, da je za razvoj pomembno dejstvo, da so podatki, vsebovani v predstavljeni, bazi zajeti med procesom dostopa do vsebine in so zato bolj natančni in relevantni.

2 PODATKOVNA BAZA ZA KONTEKSTUALNO PERSONALIZACIJO

Dostop do zanesljive podatkovne baze za kontekstualno personalizacijo je postal kritičnega pomena za razvoj postopkov na področju personalizacije. Kontekst ima velik vpliv na proces uporabnikovega odločanja in podatkovna baza, ki jo predstavljamo v tem prispevku, nam omogoča študijo tega vpliva. Poleg same podatkovne baze bomo predstavili tudi ovire in zahteve, ki jih je potrebno upoštevati pri postavitvi take baze. V sledečih podpoglavjih bomo opisali splošna navodila, ki jih je potrebno upoštevati pri zajemu kontekstualnih podatkov. V naslednjem poglavju pa bomo podrobneje prestavili našo podatkovno bazo in postopke, ki smo jih uporabili med njenim nastankom.

V okviru tega prispevka uporabljamo pojem ”zajem podatkov o uporabniku” za celotni proces zajema in vnosa podatkov v podatkovno bazo. Pojem ”aplikacija za zajem podatkov” pa opisuje dodatne funkcionalnosti (gumbi in tekstovna polja), ki smo jih dodali platformi, na kateri je uporabnik dostopal do vsebin (osebni računalnik), z namenom, da uporabniku omogočimo vnos kontekstualnih informacij.

2.1 Kako izbrati in motivirati uporabnike za zbiranje kontekstualnih informacij?

Kot smo že omenili, je pri zajemu kontekstualnih informacij zelo pomembno, da ima sam zajem čim manjši vpliv na proces uporabnikovega odločanja. To pomeni, da mora biti proces zajema vgrajen v okolje, v katerem uporabnik dostopa do vsebin in sprejema odločitve.

Pomembno je tudi, da uporabnik sodeluje zaradi pravih razlogov. Naše mnenje je, da je najboljši motiv pomoč pri razvoju personaliziranih aplikacij zase ter za druge uporabnike [4].

Če povzamemo - med procesom zajema kontekstualnih informacij naj se uporabnik čim bolj drži svojih navad in okoliščin, v katerih običajno dostopa do vsebin. Če uporabnik na primer gleda film na svojem osebnem računalniku, naj se kontekstualni podatki zajemajo tako, da uporabnika ne motijo oziroma da se jih po možnosti niti ne zaveda. Pri izgradnji naše podatkovne baze smo zato izbrali uporabnike, ki so vešči dela z računalniki in že sami po sebi uporabljajo računalnik za dostop do multimedijskih vsebin.

2.2 Katere vsebine ponuditi?

Prav tako kot ne želimo vplivati na uporabnika in na okoliščine, v katerih dostopa do vsebin, moramo poskrbeti za to, da tudi ponudba vsebin ostane nespremenjena. Najboljši način, s katerim to zagotovimo je, da se ponudba vsebin ter storitev, ki te vsebine ponujajo, ne spremenijo, kadar poteka zajem podatkov o uporabnikih. Zato je aplikacija za zajem kontekstualnih informacij običajno samostojna in neodvisna od storitve za dostop do vsebin.

2.3 Kako zajemati podatke?

Da lahko dosežemo neinvaziven zajem podatkov, moramo podatke zbirati preko komunikacijskih naprav in storitev, ki se uporabljajo v realnem okolju. Dodatne funkcionalnosti, ki so potrebne za zajem podatkov, morajo biti enostavne, uporabniku prijazne in predvsem za uporabnika čim bolj nevidne. Idealna rešitev je avtonomni proces zajema podatkov, katerega poteka se uporabnik niti ne zaveda.

Če se želimo čimbolj približati idealni rešitvi, moramo aplikacijo za zajem podatkov vgraditi v napravo, s katero uporabnik dostopa do vsebin, na tak način, da le-ta ne moti uporabnika in mu omogoča, da napravo uporablja tako, kot je navajen. Uporabnikovi vedenjski vzorci se ne smejo spremeniti tudi v situaciji, ko v sistem vnaša kontekstualne informacije v drugačni socialni situaciji (ko je na primer pri prijatelju ali z družino).

2.4 Katere dodatne informacije potrebujemo, če želimo, da bo podatkovna baza koristna za raziskave?

Če želimo podatkovno bazo uporabiti za raziskave na področju kontekstualne personalizacije, potrebujemo podatke o uporabnikih, vsebinah in kontekstu. Če želimo podatke tudi statistično ovrednotiti, potrebujemo še splošne podatke o uporabnikih (spol, starost) ter podatke o uporabnikovih vedenjskih vzorcih pri dostopanju do vsebin (pogostost uporabe....). Vsebine je potrebno opremiti z metapodatki ali pa vsaj vgraditi mehanizem, ki bi kasneje omogočal dostop do metapodatkov, kot so naslov filma, žanr, igralci itd. Primer filmskih metapodatkov se nahaja na spletni strani IMDB [5].

2.5 Kateri načrti poskusov naj bodo podprti?

Večina postopkov, s katerimi lahko natančno ocenimo uspešnost in natančnost personalizacijskih postopkov, temelji na statističnih metodah. Primeri predstavitve testnih rezultatov so matrika razvrščanja, ROC krivulja, metoda za primerjavo učinkov in statistično testiranje hipotez. Če želimo te metode uporabiti za vrednotenje sistema, je potrebno poskus pravilno zasnovati. Predvsem mora biti poskus zasnovan tako, da ga lahko izvedemo na izbrani podatkovni bazi. Zasnova je

odvisna predvsem od zajetih spremenljivk ter njihovih lastnosti. Vsako izmed spremenljivk lahko uvrstimo med kategorične, ordinalne ali numerične spremenljivke. Tipi spremenljivk nas omejujejo pri izbiri metod vrednotenja. Pomembna je tudi velikost podatkovne baze. Z uporabo apriori analize moči testov [6] lahko za vsako zasnovo poskusa ugotovimo, koliko podatkov je potrebnih za načrtovano interpretacijo rezultatov. Število potrebnih podatkov se lahko nanaša na število uporabnikov, vsebin, ocen ali polnost podatkovne baze (število ocen na vsebino).

Nadaljnje podrobnosti o zasnovi poskusa so odvisne od problema, ki ga želi raziskava analizirati. Očitno je, da ni možno vedno zadostiti vsem zahtevam. Zato je toliko bolj pomembno, da nam podatkovne baze, ki so na voljo, omogočijo, da vnaprej preverimo, če je določena zahteva izpolnjena ali ne.

3 KONTEKSTUALNA PODATKOVNA BAZA (LDOS-COMODA)

V tem poglavju bomo podrobneje opisali LDOS-CoMoDa podatkovno množico. Na kratko bomo opisali postopek zajema podatkov, kateri podatki so nam na voljo ter njihove lastnosti. Opisali bomo tudi statistične lastnosti podatkov z namenom, da omogočimo čim lažjo uporabo podatkovne baze. Vsebine v naši podatkovni bazi so filmi, naprava za dostop do vsebin pa je osebni računalnik, na katerem se nahaja tudi spletna aplikacija za zajem kontekstualnih informacij. Podrobnejši podatki so na voljo v naslednjih podpoglavjih.

3.1 Uporabniki, vsebine, kontekstualne informacije ter metapodatki

LDOS-CoMoDa podatkovna baza je bila zasnovana tako, da bi čim bolj zadostila zahtevam, ki smo jih opisali v prejšnjem poglavju. Vsebuje 30 spremenljivk izmed katerih jih je 12 kontekstualne narave. Ostale spremenljivke vsebujejo splošne podatke o uporabniku (starost, spol, mesto, država) ali pa metapodatke o filmskih vsebinah (direktor, država izvora, jezik, leto izdaje, žanr1, žanr2, žanr3, igralec1, igralec2, igralec3, proračun).

Ker so kontekstualne spremenljivke za nas še posebej pomembne, jih podrobneje opisujemo v tabeli 1. Vse kontekstualne spremenljivke so kategoričnega ali numeričnega tipa. Nekatere so kategorične že zaradi same narave svojih vrednosti (npr. vreme), ostalim pa je bil tip izbran med zasnovo podatkovne baze. To je bilo potrebno določiti na samem začetku, da bi omejili število razredov, s katerimi bo sistem operiral med procesom zajema podatkov ter da bi s primerno izbiro tipov tudi lahko poenostavili analizo podatkov v procesu reševanja problemov s področja kontekstualne personalizacije.

Osnove statistike LDOS-CoMoDa podatkovne baze v trenutku oddaje prispevka (15.12.2011) so podane v tabeli 2.

Ime spremen.	|Rg|	MVR	Opis
time	4	0.017	jutro, popoldne, večer, noč
daytype	3	0.015	delovni dan, vikend, počitnice
season	4	0.017	pomlad, poletje, jesen, zima
location	3	0.016	doma, javni prostor, pri prijateljih
weather	5	0.021	sončno / jasno, deževno, nevihtno, snežno, oblačno
social	7	0.013	sam, partner, prijatelji, sodelavci, starši, javnost, družina
endEmo	7	0	žalosten, vesel, prestrašen, presenečeni, jezni, zgrožen, nevtralen
dominantEmo	7	0	žalosten, vesel, prestrašen, presenečeni, jezni, zgrožen, nevtralen
mood	3	0	pozitivno, nevtralno, negativno
physical	2	0.022	zdrav, bolan
decision	2	0.021	lastna izbira, izbira drugih
interaction	2	0.020	prvič, že večkrat

Tabela 1: Kontekstualne spremenljivke in njihove osnovne lastnosti, imena navajamo v originalni obliki. |Rg| predstavlja število ordinalnih ali kategoričnih razredov posamezne spremenljivke, MVR pa delež manjkajočih vrednosti

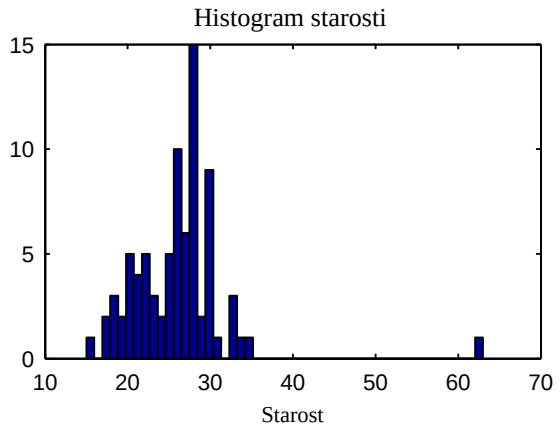
Tabela 2: Osnovni podatki podatkovne baze

število uporabnikov	95
število vsebin	961
število ocen	1665
povprečna starost	27.0
število držav	6
število mest	18
največ ocen podanih s strani enega uporabnika	220
najmanj ocen podanih s strani enega uporabnika	1

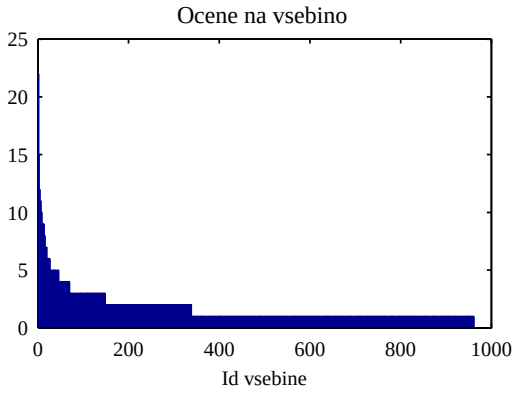
Ko pričakujemo posplošitev rezultatov statističnih postopkov na populacijo, je pomembna reprezentativnost analiziranega vzorca. Znano je, da je starost uporabnika zelo pomemben parameter v procesu napovedovanja njegovega vedenja in njegove uporabe modernih komunikacijskih naprav. Histogram starosti uporabnikov je podan na sliki 1. Distribucija uporabnikov in ocen je prikazana na slikah 2 in 3.

Polnost podatkovne baze je prikazana na sliki 4. Svetlost posameznih točk je odvisna od števila ocen, dodeljenih izbranemu podsklopu vsebin (en stolpec za vsak podsklop vsebin) s strani vsakega posameznega uporabnika (vsak uporabnik je predstavljen s svojo vrstico). Iz analize je razvidna visoka raznolikost med uporabniki in vsebinami. Ker so se uporabniške identifikacijske številke generirale zaporedno, lahko opazimo, da so uporabniki, ki so z ocenjevanjem vsebin začeli kasneje v podatkovno bazo prispevali manj ocen, kot tisti, ki so

sodelovali od samega začetka zajema podatkov.



Slika 1: Histogram starosti uporabnikov. Večina se jih nahaja med 18 in 35 letom, kar reprezentativnost baze omejuje na ta starostni interval.

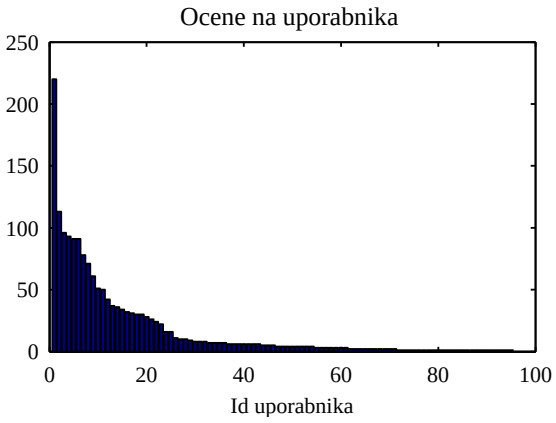


Slika 2: Število ocen na vsebino. Večino vsebin je prejelo 2 do 3 ocene.

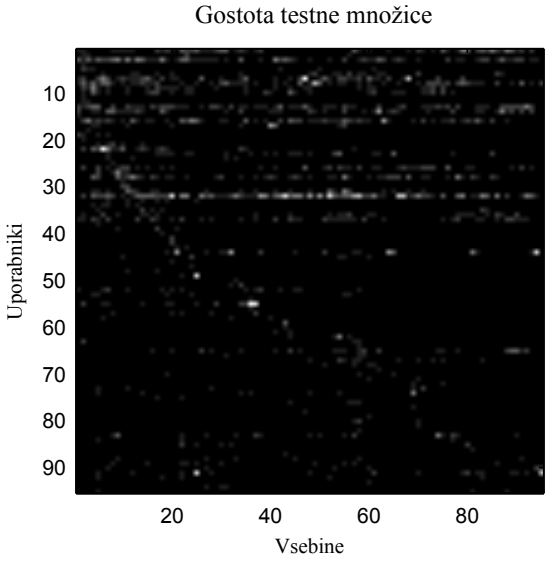
3.2 Zajem podatkov

Proces zajemanja kontekstualnih informacij o uporabniku je zelo občutljiv na kontekstualne motnje. Zato je pravilna interpretacija zajetih podatkov težaven in kompleksen proces. Kot smo prikazali v poglavju 2, lahko že sam proces zajema podatkov moti uporabnika in tako spremeni zajete kontekstualne informacije.

Podatke v podatkovni bazi smo zajemali s pomočjo posebej za ta namen zasnovane uporabniku prijazne spletne aplikacije. Pri tem se je pomembno zavedati, da smo uporabnikom naročili, da vnesejo ocene in kontekstualne informacije takoj zatem, ko so dostopali do vsebin (filmov). Tako vnesene kontekstualne informacije so veliko bolj natančne kot tiste, ki jih uporabnik vnaša po spominu za vsebino, do katere je dostopal pred nekaj dnevi ali meseci. Uporabnike smo dodatno motivirali s tem, da smo jim pojasnili, da z vnosom ocen pomagajo pri razvoju novih postopkov in si izboljšujejo svoj



Slika 3: Število ocen na uporabnika. Med uporabniki opazimo velike razlike glede števila podanih ocen, kjer je veliko število uporabnikov podalo manj kot 10 ocen, medtem ko je skupina uporabnikov s 50 in več podanimi ocenami prav tako sorazmerno velika.



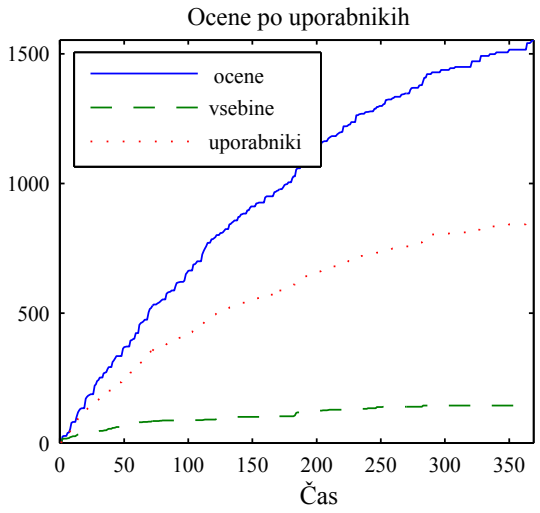
Slika 4: Gostota testne množice. Svetlost točk je premosorazmerna s številom podanih ocen.

uporabniški model [4]. Poleg zajema podatkov spletna aplikacija nudi tudi sledenje ogledanih filmov, iskalnik po nekaterih kontekstualnih podatkih (čas dneva, socialni status...) ter skupinski priporočilni sistem.

Slika 5 prikazuje rast števila uporabnikov s časom. Opazimo, da hitrosti rasti s časom upadajo, kar je pričakovano, lahko pa zaznamo določene hitrejše lokalne porasti števila uporabnikov. Spletna aplikacija še vedno deluje in je na voljo na http://212.235.187.145/spletnastran/raziskave/um/emotions/login.php.

3.3 Dostop do LDOS-CoMoDa podatkovne baze

Raziskovalci, zainteresirani za uporabo podatkovne baze LDOS-CoMoDa , bodo na zahtevo po elektron-



Slika 5: Potek zajema testnih podatkov v času.

ski pošti prejeli z geslom zaščiteno povezavo, potem ko bodo vzpostavili kontakt preko spletne pošte na naslovu ldos-comoda@ldos.si. Poleg podatkovne baze bo raziskovalcem posredovana tudi posodobljena verzija osnovnih podatkov o bazi ter navodila za dostop do podatkov.

4 Zaključek

Z več kot 90 uporabniki, 900 vsebinami ter 1600 ocenami nudi LDOS-CoMoDa podatkovna baza okolje, primerno za raziskave več odprtih problemov na področju kontekstualne personalizacije [7], [8], [9]. Najpomembnejši del podatkovne baze so kontekstualne spremenljivke, ki opisujejo uporabnikova čustva itd, in nudijo dokaj natančen opis uporabnikovega dejanskega konteksta med dostopanjem do vsebine. Podatkovna baza je javno dostopna na podlagi predhodne zahteve preko elektronske pošte.

Glavna prednost naše podatkovne baze je v tem, da vsebuje kontekstualne informacije iz faze uporabnikovega dostopa do vsebin, za katere mislimo, da so najbolj natančne. Vsebuje 12 tipov potencialno kontekstualnih informacij. Na podlagi apriori analize moči testov smo ugotovili, da je velikost in polnost podatkovne baze zadostna, da jo lahko uporabimo v raziskavah na področju več odprtih problemov kontekstualne personalizacije.

Kljub vsemu pa ima podatkovna množica tudi nekaj slabosti, kar je hkrati tudi razlog, da je proces zajema kontekstualnih informacij še vedno aktiven. Apriori analiza statistične moči testov [6] je pokazala, da je ob tipični velikosti učinka (ang. effect size) potrebno število uporabnikov okoli 1400. Naša podatkovna baza ta prag preseže, vendar je pri določenih uporabniških poskusih bazo potrebno razdeliti na podskupine, ki pa praga ne presežejo več. Zato bo potrebno zagotoviti še več uporabnikov in njihovih ocen ter tako izboljšati velikost in polnost podakovne baze.

References

[1] A. Dey, G. Abowd, Towards a better understanding of context and context-awareness, Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing (1999) 304–307.

[2] L. Baltrunas, B. Ludwig, S. Peer, F. Ricci, Context relevance assessment and exploitation in mobile recommender systems, Personal and Ubiquitous Computing (2011) 1–20doi:10.1007/s00779-011-0417-x.

[3] F. Díez, J. E. Chavarriaga, P. G. Campos, A. Bellogín, Movie Recommendations based in explicit and implicit features extracted from the Filmtipset dataset, in: Proceedings of the Workshop on Context-Aware Movie Recommendation, 2010, pp. 45—-52.

[4] J. Herlocker, J. Konstan, L. Terveen, J. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems 22 (1) (2004) 5–53. doi:/10.1145/963770.963772.

[5] The internet movie database (imdb) @ONLINE (Dec. 2011). URL http://www.imdb.com/

[6] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum, 1988.

[7] A. Odić, M. Kunaver, J. Tasič, A. Košir, Open issues with contextual information in existing recommender system databases, in: ERK 2010 Proceedings, 2010.

[8] G. Adomavicius, R. Sankaranarayanan, S. Sen, A. Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Transactions on Information Systems (TOIS) 23 (1) (2005) 103–145.

[9] Z. Yujie, W. Licai, Some Challenges for Context-aware Recommender Systems, in: Computer Science and Education (ICCSE), 2010 5th International Conference on, 2010, pp. 362–365.

Andrej Košir je izredni profesor na Fakulteti za elektrotehniko Univerze v Ljubljani. Njegov raziskovalni interes vključuje operacijske raziskave v telekomunikacijah, uporabniško modeliranje in procesiranje socialnih signalov.

Ante Odić je mladi raziskovalec na Fakulteti za elektrotehniko Univerze v Ljubljani. V okviru doktorskega študija raziskuje uporabo kontekstualnih informacij v personaliziraih storitvah.

Matevž Kunaver je raziskovalec in asistent na Fakulteti za elektrotehniko Univerze v Ljubljani. Njegove raziskave vključujejo skupinske in hibridne priporočilne sisteme za različne aplikacije v telekomunikacijah.

Marko Tkalčič je raziskovalec na Fakulteti za elektrotehniko Univerze v Ljubljani. Njegov raziskovalni interes vključuje uporabo emotivnih in osebnostnih parametrov pri modeliranju uporabnikov in vsebin v telekomunikacijskih aplikacijah.

Jurij F. Tasič je redni profesor na Fakulteti za elektrotehniko Univerze v Ljubljani. Raziskovalni interes obsega napredne algoritme v komunikacijskih sistemih, obdelavo večdimenzionalnih signalov in vzporedne algoritme.