



VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
GRAFINIŲ SISTEMŲ KATEDRA

Tadas Savickas

**PANAŠUMO MODELIŲ EFEKTYVUMAS SOCIALINIULOSE TINKLUOSE
THE EFFICIENCY OF SIMILARITY BASED MODELS IN SOCIAL
NETWORKS**

Baigiamasis magistro darbas

Informacinių technologijų studijų programa, valstybinis kodas 621E14004

Informatikos inžinerijos mokslo kryptis

Vilnius, 2010

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
GRAFINIŲ SISTEMŲ KATEDRA

TVIRTINU
Katedros vedėjas

(Parašas)

Romualdas Baušys

(Vardas, pavardė)

(Data)

Tadas Savickas

PANAŠUMO MODELIŲ EFEKTYVUMAS SOCIALINIULOSE TINKLUOSE
THE EFFICIENCY OF SIMILARITY BASED MODELS IN SOCIAL
NETWORKS

Baigiamasis magistro darbas

Informacinių technologijų studijų programa, valstybinis kodas 621E14004
Informatikos inžinerijos mokslo kryptis

Vadovas

(Moksl. laipsnis, vardas, pavardė)

(Parašas)

(Data)

Konsultantas

(Moksl. laipsnis, vardas, pavardė)

(Parašas)

(Data)

Konsultantas

(Moksl. laipsnis, vardas, pavardė)

(Parašas)

(Data)

Vilnius, 2010

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
GRAFINIŲ SISTEMŲ KATEDRA

TVIRTINU

Technologijos mokslo sritis
Informatikos inžinerijos mokslo kryptis
Informatikos inžinerijos studijų kryptis
Informacinių technologijų studijų programa, valstybinis
kodas 621E14004

Katedros vedėjas

(Parašas)

Romualdas Baušys

(Vardas, pavardė)

(Data)

**BAIGIAMOJO MAGISTRO DARBO
UŽDUOTIS**

2010 - - Nr.

Vilnius

Studentui (ei) Tadui Savickui

(Vardas, pavardė)

Baigiamojo darbo tema: Panašumo modelių efektyvumas socialiniuose tinkluose

patvirtinta 2010 m. dekanas potvarkiu Nr.

Baigiamojo darbo užbaigimo terminas 2010 m. gegužės 31 d.

BAIGIAMOJO DARBO UŽDUOTIS:

Išanalizuoti rekomendacinės sistemos veikimo principus ir pasiūlyti metodą rekomendacinės sistemos efektyvumui ir rekomendacijų tikslumui pagerinti. Pasirinktam rekomendacinės sistemos metodui tobulinti naudoti realią duomenų bazę. Atlikti eksperimentus ir palyginti gautus rezultatus, išanalizuoti atskirų vartotojų grupių poveikį rezultatams. Pateikti išvadas ir pasiūlymus.

Baigiamojo darbo rengimo konsultantai:

.....
(Moksl. laipsnis, vardas, pavardė)

Vadovas

.....
(Parašas)

.....
(Moksl. laipsnis, vardas, pavardė)

Užduotį gavau

.....
(Parašas)

.....
(Vardas, pavardė)

.....
(Data)

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
GRAFINIŲ SISTEMŲ KATEDRA

Technologijos mokslo sritis
Informatikos inžinerijos mokslo kryptis
Informatikos inžinerijos studijų kryptis
Informacinių technologijų studijų programa, valstybinis
kodas 621E14004

PAŽYMA
APIE BAIGIAMĄJĮ MAGISTRO DARBĄ
.....Nr.
Vilnius

Studentas (ė).....
(Vardas, pavardė)

Studento (ės) studijų svertinis įvertinimų vidurkis..... balo.

Baigiamojo darbo tema:

Baigiamasis darbas peržiūrėtas ir studentui (ei)
leidžiama ginti šį baigiamąjį darbą magistro laipsnio suteikimo komisijoje.

Katedros vedėjas
(Parašas)

doc. dr. Romualdas Baušys
(Moksl. laipsnis, vardas, pavardė)

Vadovo atsiliepimas

APIE BAIGIAMĄJĮ MAGISTRO DARBĄ
.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Vadovas
(Parašas)

.....
(Moksl. laipsnis, vardas, pavardė)

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Grafinių sistemų katedra

ISBN ISSN
Egz. sk. 2
Data 2010-06-01

Informacinių technologijų studijų programos baigiamasis magistro darbas

Pavadinimas: Panašumo modelių efektyvumas socialiniuose tinkluose

Autorius: Tadas Savickas Vadovas: Dr. Lina Vasiliauskienė

Kalba

☒ Lietuvių

☐ Anglų

Anotacija

Baigiamajame magistro darbe yra nagrinėjamos pagrindinės rekomendacinių sistemų socialiniuose tinkluose problemos: rekomendacijų tikslumas, pasyvūs vartotojai, neaiškios rekomendacijos. Iškeliama pagrindiniai tikslai, kuriais siekiama išanalizuoti rekomendacinių sistemų veikimo principus ir pasiūlyti metodą rekomendacinės sistemos efektyvumui ir rekomendacijų tikslumui gerinti. Analizuojama užsienio literatūra, atliekamas eksperimentas su realia duomenų baze. Analizuoti rezultatai vertinami skirtingų vartotojų grupių (pasyvūs, aktyvūs, tarpiniai vartotojai), analizuojamas kiekvienos grupės poveikis bendram rezultatui. Aprašytos aiškios rekomendacijos ir pateikti jų pavyzdžiai. Rezultatai vertinti trimis skirtingais kriterijais: PAP, PAVP ir ĮA. Pirmi du rodo skaičiavimo nuokrypius, trečiasis – įverčių apimtį. Atlikti penki bandymai skirtingomis sąlygomis ir pateikti kiekvieno bandymo bei bendri rezultatai. Išsikelti tikslai yra įgyvendinti, nes analizuojant gautus rezultatus, tiek bendras rekomendacijų tikslumas, tiek pasyvių vartotojų pagėrėjo. Bendras pagėrėjo 4% ir 11%, pasyvių vartotojų – 11% ir 17%. Darbe aprašomas modelis gali būti nesunkiai pritaikomas rekomendacinei sistemai. Šis baigiamasis darbas gali būti naudojamas kaip pagrindas ar literatūros šaltinis tolesniems tyrimams Lietuvoje.

Darbo apimtis – 53 p. teksto be priedų, 24 pav., 10 lent., 31 literatūros šaltinis.

Prasminiai žodžiai: rekomendacinė sistema, socialinis tinklas, vartotojų grupės, nuokrypis, įverčių apimtis, raida, BF metodas, rekomendacija.

Vilnius Gediminas Technical University
VGTU faculty of fundamental Science faculty
Department of Graphical systems

ISBN ISSN
Copies No. 2
Date 2010-06-01

Information Technology study program master thesis

Title: The efficiency of similarity based models in social networks

Author: Tadas Savickas Academic supervisor: Dr. Lina Vasiliauskienė

Language

☐ Lithuanian

☒ English

Annotation

The thesis analyses the main problems of recommender systems in social networks: the accuracy of recommendations, cold start users, uncertain recommendations. The main goals are raised, which are used to analyze the principles of processes in recommender systems, and to offer a novel method to improve the efficiency and the accuracy of recommendations. The foreign articles are discussed and the new method is implemented on an existing data set. The results are evaluated according to the different groups of users (cold start users, heavy raters and intermediate users) and the effect of each group to the main result is analyzed. The transparent recommendations are explained with the examples. The results are evaluated using three different metrics: MAE, MAUE and RC. The first two determine the deviation of the calculations the third determines the coverage of the ratings. Five experiments were made with different conditions and the results of each are presented along with the general results. The held purposes were accomplished because the accuracy of recommendations increased for all users and for cold start users as well. The benefit of the accuracy for all users is 4% and 11%, for cold start users 11% and 17%. The model described in the thesis can be easily incorporated to the recommender system. This thesis can be used as the basis of future work of recommender systems in Lithuania.

Work size – 53 p. text, 24 illustrations, 10 tables, 31 bibliographic sources.

Keywords: recommender system, social network, users groups, deviation, volume of estimate, evolution, CF method, recommendation.

PAVEIKSLŲ SĄRAŠAS

1 pav. Socialinio tinklalapio MovieLens namų puslapis.....	21
2 pav. Bendro filtravimo rekomendacinės sistemos architektūra.....	22
3 pav. Programos veikimo schema	25
4 pav. Aiški rekomendacija	25
5 pav. Standartinio BF modelio ir siūlomo modelio palyginamoji schema	30
6 pav. Filmų duomenų bazės dalis	33
7 pav. Pirmo eksperimento PAP kriterijaus rezultatai.....	35
8 pav. Pirmo eksperimento PAVP kriterijaus rezultatai.....	35
9 pav. Pirmo eksperimento ĮA kriterijaus rezultatai.....	35
10 pav. Antro eksperimento PAP kriterijaus rezultatai	36
11 pav. Antro eksperimento PAVP kriterijaus rezultatai	37
12 pav. Antro eksperimento ĮA kriterijaus rezultatai	37
13 pav. Trečio eksperimento PAP kriterijaus rezultatai	38
14 pav. Trečio eksperimento PAVP kriterijaus rezultatai	38
15 pav. Trečio eksperimento ĮA kriterijaus rezultatai	39
16 pav. Ketvirto eksperimento PAP kriterijaus rezultatai	40
17 pav. Ketvirto eksperimento PAVP kriterijaus rezultatai	40
18 pav. Ketvirto eksperimento ĮA kriterijaus rezultatai	40
19 pav. Penkto eksperimento PAP kriterijaus rezultatai	41
20 pav. Penkto eksperimento PAVP kriterijaus rezultatai	42
21 pav. Penkto eksperimento ĮA kriterijaus rezultatai	42
22 pav. Visų eksperimentų PAP kriterijaus vidurkis.....	43
23 pav. Visų eksperimentų PAVP kriterijaus vidurkis.....	44
24 pav. Visų eksperimentų ĮA kriterijaus vidurkis.....	44

LENTELIŲ SĄRAŠAS

1 lentelė. Rekomendacijų pateikimo metodai.....	19
2 lentelė. Mišrių metodų sudarymo būdai	20
3 lentelė. Panašumo reikšmės pasirinktam vartotojui.	26
4 lentelė. Panašumo reikšmės tarp vartotojų pasirinktai kategorijai.	27
5 lentelė. Pirmo eksperimento rezultatai.	34
6 lentelė. Antro eksperimento rezultatai.....	36
7 lentelė. Trečio eksperimento rezultatai.....	38
8 lentelė. Ketvirtro eksperimento rezultatai.....	39
9 lentelė. Penkto eksperimento rezultatai.....	41
10 lentelė. Visų eksperimentų rezultatų vidurkis.	43

SANTRUMPOS

BF – bendro filtravimo metodas

BRS – bendras rekomendacijų skaičius.

I_A – įverčių apimtis

I_D – įverčių dalis

NFS – neįvertintų filmų skaičius.

PAP – pagrindinė absoliuti paklaida

PAVP – pagrindinė absoliuti vartotojo paklaida

PKK – Pearsono koreliacijos koeficientas

RS – rekomendacinė sistema

RSVG – rekomendacijų skaičius vartotojų grupei.

ST – socialinis tinklas

VIFS – vienodų įvertintų filmų slenkstis

VPS – vartotojų panašumo slenkstis

TURINYS

1. ĮVADAS	12
2. LITERATŪROS APŽVALGA.....	15
2.1. Atliktų tyrimų rekomendacinėse sistemose analizė	15
2.2. RS socialiniuose tinkluose	16
2.2.1. Pagrindinės RS charakteristikos	18
2.3. BF panašumo modelis	21
2.3.1. Panašumo kriterijai ir įverčių spėjimas.....	22
3. BF MODELIO TOBULINIMAS.....	24
3.1. Pagrindinė rekomendacinė sistema	24
3.2. Siūlomas modelis su kategorijomis.....	25
3.2.1. Kategorijos.....	26
3.2.2. Aišrios rekomendacijos.....	27
3.2.3. Vartotojų skirstymas	28
3.3. Siūlomo modelio schema ir palyginimas su standartiniu BF modeliu.....	28
4. LYGINAMASIS MODELIŲ EKSPERIMENTAS.....	31
4.1. Vertinimo kriterijai.....	31
4.2. Duomenų rinkinys	32
4.3. Eksperimentų rezultatai.....	33
4.3.1. Pirmas eksperimentas	34
4.3.2. Antras eksperimentas	36
4.3.3. Trečias eksperimentas.....	37
4.3.4. Ketvirtas eksperimentas	39
4.3.5. Penktas eksperimentas	41
4.3.6. Visų eksperimentų apibendrinimas.....	43
5. IŠVADOS IR PASIŪLYMAI	45
5.1. IŠVADOS.....	45
5.2. PASIŪLYMAI	46
LITERATŪROS SĄRAŠAS.....	47
PUBLIKUOTAS STRAIPSNIS	50
PRIEDAI	54
1 priedas. Vartotojų panašumo nustatymas	54
2 priedas. Rekomendacijos pateikimas.....	55
3 priedas. Vartotojų įverčių nuskaitymas ir skirstymas į grupes.....	56

1. ĮVADAS

Problemos aktualumas. Dėl tobulėjančių kompiuterinės technikos ir interneto technologijų interneto naudojimas labai sparčiai plečiasi. Vartotojai vis daugiau užduočių atlieka interaktyvioje erdvėje, taip sutaupydami daug laiko ir energijos. Taip pat vartotojai su kitais bendraminčiais internete gali bendrauti visą parą neišeidami iš namų. Dėl šios priežasties paskutiniu metu labai sparčiai plečiasi socialinių tinklų (toliau – ST) naudojimas ir retas interneto vartotojas neturi savo paskyros kokiam nors socialiniame tinkle. Kadangi ST suburia labai didelį žmonių skaičių, tai tampa puikiai vieta gamintojams reklamuoti save, savo produktus, taip pat jais prekiauti. Todėl ST būtų galima išskirti į kelis tipus: komerciniai (Amazon.com, Ebay.com), bendravimo (Facebook.com, Twitter.com), rekomendaciniai (MovieLens.org, Epinions.com, Last.fm).

Šiame darbe bus tiriami rekomendaciniai ST. Šie tinklai orientuojasi į vieną ar keletą tam tikrų produktų, bet ne su tikslu parduoti, o rekomenduoti vartotojui galimai patinkančius produktus: knygas, filmus, muziką ir pan. Šis darbas ir tyrimas bus paremtas MovieLens socialiniu tinklu ir realia šio ST duomenų baze. Šiame ST vartotojams yra rekomenduojami filmai. Kadangi ST apima labai didelį kiekį įvairios informacijos ir produktų, iškyla pagrindinė **problema** – kaip vartotojui pasirinkti teisingą produktą, kuris galimai būtų įdomus ir artimas vartotojo skonui. Rekomendaciniai ST naudoja rekomendacines sistemas (toliau – RS), kurios palengvina vartotojo pasirinkimą. Tačiau šioje vietoje iškyla kita **problema** – RS efektyvumas ir tikslumas bei pasitikėjimas šia RS, nes jeigu sistema pateiks netikslias rekomendacijas vartotojams, tai niekas nenorės naudotis tokia RS.

Problemos ištyrimas ir naujumas. RS naudojimą socialiniuose tinkluose tiria daug autorių ir ši tema yra plačiai paplitusi. Atliekami įvairūs tyrimai, rengiami konkursai. Didelė dalis tyrėjų yra įvairių užsienio universitetų studentai. Dėl to yra daug straipsnių, kurie yra pristatomi įvairiose konferencijose anglų kalba. Vertinant Lietuvos mastu ši darbo tema yra nauja, nes apie socialiuose tinkluose taikomas RS Lietuvoje nėra tyrinėta, taip pat nėra šiai tyrimo sričiai reikalingos medžiagos ir literatūros lietuvių kalba. Lietuvis, tiriantis šią sritį, yra Gediminas Adomavičius, tačiau jis dirba Minesotos universitete.

Daugelis autorių tyrinėja RS efektyvumą ir rekomendacijų tikslumą, siekdami RS spėjinius padaryti kuo tikslesnius, palyginti su realių vartotojų įverčiais. Bandoma sukurti įvairius metodus skirtus RS tikslumui pagerinti, taip pat atliekama vartotojų profilių analizė lyginant profilius ir analizuojant, kokią įtaką panašūs profiliai turi atitinkamiems vartotojų sprendimams.

Analizuojant šiam tiriamajam darbui reikalingą ir susijusią informaciją, buvo pastebėta, kad ankstesniuose šios srities tyrimuose buvo labai mažai kreipiama dėmesio arba visai neatsižvelgiama į atskiras vartotojų grupes, t.y. pasyvūs, aktyvūs ir visi likę vartotojai.

Darbo tikslas – išanalizuoti RS veikimo principus ir pasiūlyti metodą rekomendacinės sistemos efektyvumui ir rekomendacijų tikslumui pagerinti.

Šiam tikslui įvykdyti iškeliami tokie pagrindiniai uždaviniai:

- 1) apžvelgti naudojamas RS ir susijusią literatūrą,
- 2) išanalizuoti rekomendacinės sistemos veikimo principą,
- 3) surasti atvirojo kodo rekomendacinę sistemą ir ją patobulinti pasiūlant metodą,
- 4) atlikti eksperimentus ir jiems panaudoti realius duomenis,
- 5) išanalizuoti atskirų vartotojų grupių įtaką RS tikslumui,
- 6) pagerinti RS rekomendacijų tikslumą,
- 7) pateikti išvadas ir pasiūlymus.

Iškeltiems tikslams pasiekti yra pasirinktas MovieLens ST, kuriame yra naudojama bendro filtravimo (toliau – BF) metodas ir reali, nemokama duomenų bazė. Ši ST sukūrė „GroupLens“ tyrėjų grupė iš Minesotos universiteto. BF metodas dažniausiai yra naudojamas rekomendacinėse sistemose, todėl jis yra kaip pagrindas naujiems metodams kurti ir RS tobulinti. Šis metodas remiasi panašių vartotojų principu, t.y. lyginama visų vartotojų įverčių raida. RS atrinkus panašiausius vartotojus, jų įverčiai panaudojami apskaičiuoti rekomendacijai, kuri pateikiama tam vartotojui, kuris nematė atitinkamo filmo ar neįvertino atitinkamo produkto.

Magistro tiriamajame darbe taip pat yra naudojamas BF metodas, paremtas panašiais vartotojais kaip pagrindas tyrimui. Pagrindinis sprendimas šiam metodui tobulinti yra jo naudojimas ne visai duomenų bazei, o išskirstytai pagal kategorijas, šiuo atveju filmus išskirstant pagal žanrą. Taip pat atskiroms vartotojų grupėms bus naudojamas skirtingas panašumo slenkstis, kuris turi įtakos galutiniam rekomendacijos tikslumui.

Praktinė šio darbo vertė yra rekomendacinių sistemų, naudojančios standartinį BF metodą, efektyvumo tobulinimas. Taip pat šis tiriamasis darbas gali būti kaip literatūra ir pagrindas kitiems tolesniems šios srities tyrimams ir RS analizei Lietuvoje. Pasiūlytas metodas gali būti taikomas ne tik filmams, bet įvairioms produktų kategorijoms (knygos, muzika, automobiliai, paslaugos). Metodas gali būti pritaikomas socialiniam tinklalapiui.

Pagrindinės hipotezės. 1. RS efektyvumas galis skirtis atitinkamose vartotojų grupėse ir tam tikra vartotojų grupė gali lemti daugiau ar mažiau galutinį rezultatą. **2.** Atskiroms vartotojų grupėms turėtų būti taikomi skirtingi panašumo slenksčiai.

Tyrimo objektas – socialiniuose tinkluose naudojamų rekomendacinių sistemų tikslumo ir efektyvumo tyrimas.

Tyrimo metodai. Straipsnių, dokumentų, kitos literatūros analizė, metodo pritaikymas, statistinė duomenų analizė.

Darbo struktūra. Darbą sudaro įvadas, 3 dalys, išvados, anotacija (lietuvių ir anglų kalbomis), literatūros sąrašas, publikuotas straipsnis ir priedai.

Straipsnis, parašytas šio baigiamojo darbo pagrindu, iškeltos problemos, uždaviniai ir pagrindiniai gauti rezultatai buvo pristatyti 13-ojoje Lietuvos jaunųjų mokslininkų konferencijoje „Mokslas – Lietuvos ateitis“ Vilniaus Gedimino technikos universitete.

2. LITERATŪROS APŽVALGA

2.1. Atliktų tyrimų rekomendacinėse sistemose analizė

Apie RS tobulinimą ir rekomendacijų tikslumo gerinimą yra parašyta daug straipsnių ir publikacijų. Juose yra pristatomi įvairūs sugalvoti RS modeliai, kuriais stengiamasi padaryti RS patikimesnę. Tyrėjus, nagrinėjančius RS, galima išskirti į dvi atskiras grupes: tyrėjai, tiriantys panašumo modelius, ir tyrėjai, nagrinėjantys pasitikėjimu paremtus modelius. RS tyrėjai iki šiol daugiausia tiria algoritmus, kuriais stengiamasi rekomendacijas padaryti kuo tikslesnes ir artimesnes vartotojo skoniui (Herlocker 1999) (McLaughlin 2004).

Šis darbas yra priskiriamas prie pirmosios tyrėjų grupės. Rekomendacijų skaičiavimui yra nagrinėjamos tokios strategijos kaip *produktais paremta BF* (Sarwar 2001), *Bayesiano tinklų* (Breese 1998) ar *aplinkybių analizės* (Canny 2002), tačiau *vartotojas-vartotojas BF* strategija tampa vis artimesnė ir geriausiai atspindi realiaame pasaulyje gaunamas rekomendacijas. Darbe yra nagrinėjamas *vartotojas-vartotojas BF* panašumo modelis (Resnick 1994) (Shardanand 1995), kuris paremtas panašiais vartotojais, juos lyginant pagal filmų įverčių raidą. Darbui reikalinga literatūra nagrinėjama taip pat remiantis pirmosios tyrėjų grupės parašytais straipsniais ir pateiktomis publikacijomis. Autorių yra gana daug, tačiau vienus iš pagrindinių tyrėjų galima laikyti „GroupLens“ grupę¹ iš Minesotos universiteto Jungtinėse Amerikos Valstijose. Jie sukūrė ST pavadinimu „MovieLens“². Šis ST apima didelę filmų duomenų bazę ir jame yra taikoma rekomendacinė sistema, kuri registruotiems ST vartotojams rekomenduoja filmus.

Pirmajame nagrinėjamame darbe „Accounting for Taste: Using Profile Similarity to improve Recommender Systems“ (Bonhard 2006), autoriai teigia, kad RS efektyvumą galima pagerinti parodant daugiau informacijos apie vartotoją, iš kurio atėjo rekomendacija. Jie atliko interaktyvų eksperimentą, kuriuo norėjo parodyti, kad vartotojo, kuris rekomenduoja, giminiškumas (broliai, seserys, tėvai, pusbroliai ir pan.), profilių panašumas ir įvertinti tie patys produktai turi įtakos vartotojo, kuris gauna rekomendaciją, galutiniam sprendimui. Gauti rezultatai patvirtino autorių padarytą prielaidą. Vartotojai priima geresnius sprendimus tada, kai žino daugiau informacijos apie vartotoją, iš kurio gauna rekomendaciją. Tačiau rezultatai, gauti atliekant eksperimentą, nėra lygiaverčiai, nes autoriai negalėjo imituoti giminystės ryšio tarp vartotojų, kadangi eksperimento dalyviai buvo parinkti atsitiktine tvarka ir tarp jų nebuvo giminystės ryšio.

Kitame darbe “New Recommendation Techniques for Multicriteria Ratings Systems” (Adomavičius 2007) autoriai pasiūlė produktus vertinti ne vienu aspektu, o keturiais. Jie taip pat naudojo BF panašumo metodą kaip pagrindą tyrimams. Eksperimentui naudojo filmų duomenų bazę. Autoriai padarė prielaidą, kad vartotojams filmas gali patikti arba nepatikti keliais aspektais, o

¹ <http://www.grouplens.org/>

² <http://movielens.umn.edu/login>

galutinis įvertis turėtų būti atitinkamas pagal konkrečių aspektų įverčius. Todėl atliekant eksperimentą filmai turėjo keturis kriterijus, kuriuos vertino vartotojai: filmo istorija, vaidyba, režisūra, vaizdo efektai. Įvertinus filmą šiais skirtingais aspektais, buvo apskaičiuojamas galutinis įvertis. Gauti eksperimento rezultatai su realia duomenų baze parodė, kad naudojant šį metodą galima sėkmingai pagerinti rekomendacijų tikslumą.

Vienas iš naujausių straipsnių šia tema yra “Exploiting User Similarity based on Rated-Item Pools for Improved User-based Collaborative Filtering” (Shi 2009). Šiame straipsnyje tyrėjai nagrinėja vartotojų įverčius. Jie kiekvieno vartotojo įverčių raidą padalina į tris lygias dalis. Pavyzdžiui, vartotojas yra įvertinęs 10 filmų dešimties balų vertinimo skalėje. Įverčiai yra 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 atskiriems filmams. Pagal autorių siūlomą metodą šis vartotojas turės tris duomenų rinkinius: a) patinkantys filmai - filmai įvertinti 10, 9, 8, 7, b) neįdomūs filmai – filmai įvertinti 1, 2, 3, c) kiti filmai – filmai įvertinti 5, 6, 7. Tokiu būdu yra sukuriami trys atskiri duomenų rinkiniai kiekvienam vartotojui. Skaičiuojant galutinį panašumą tarp atitinkamų vartotojų atskiruose duomenų rinkiniuose naudojama BF panašumo formulė. Pradžioje yra apskaičiuojamas išskirstytų įverčių rinkinių panašumas, tada apskaičiuojamas panašumas lyginant neišskirstytą įverčių raidą ir paskutiniu etapu gautos panašumo reikšmės yra padauginamos iš atitinkamų koeficientų ir gauti rezultatai susumuojami. Koeficientų suma yra lygi vienetui. Tačiau šio metodo trūkumas, nes autoriai įverčius skirsto į lygias dalis. Tai reiškia, jeigu vartotojas bus įvertinęs visus filmus, pavyzdžiui, 9, tai metodas bet koku atveju paskirstys į patinkančius, nepatinkančius ir vidutinius filmus, nors visi filmai turi aukštą įvertį – 9. Autoriai šį trūkumą pastebėjo, tačiau išanalizavę naudojamus eksperimentui duomenis jie padarė prielaidą, kad įverčiai pasiskirsto teisingai į tris lygias dalis be didesnių nuokrypių. Gauti eksperimento rezultatai, taikant skirtingas eksperimentui sąlygas, parodė, kad jų metodas pagerina RS rekomendacijų tikslumą, palyginti su paprastu BF panašumo metodu.

Taip pat siūloma BF metodo rekomendacijų tikslumą gerinti tiriant spėjinių klaidas (Ding 2008). Tačiau šio metodo trūkumas tas, kad reikalinga daug resursų klaidų analizei viso tyrimo metu atlikti.

Alternatyvūs metodai siekia analizuoti panašumo metodą dar išsamiau, skaičiuojant panašumą ne tik tarp tiesioginių, galimai panašių vartotojų, bet ir panašumą su panašaus vartotojo panašiais vartotojais (Zhang 2007). Deja, tokios strategijos panaudojimas yra ganėtinai brangus, nes skaičiavimas panašumo tarp panašaus vartotojui panašių vartotojų yra pakankamai sudėtingas ir eksponentiškai greitai augantis didėjant vartotojų skaičiui.

2.2. RS socialiniuose tinkluose

ST yra apibrėžiami labai įvairiai, vienas pagrindinių apibrėžimų – tai interaktyvi platforma, kur žmonės bendrauja ir įvairiai komunikuoja tarpusavyje internetu. Žmonės kuria įvairias

bendruomenės, kurios atspindi vartotojų tam tikrus interesus ir pomėgius. Yra daug bendravimo per socialinius tinklus tipų. Tinklalapiai skirstomi į du tipus³ (Wikipedia): akcentuojama žmonių tarpusavio bendravimas ir žmonių mėgstami užsiėmimai, hobiai. Pirmasis tipas yra skirtas žmonių socialiniam bendravimui, t.y. tokie ST Facebook, Bebo ir Myspace. Antrasis tipas orientuojasi į žmonių hobį. Daugelis žmonių mėgsta fotografuoti ir domisi fotografija, todėl yra tokie ST kaip Flickr ar Photobucket. Vartotojai šiuose ST gali keistis nuotraukomis, jas vertinti ir diskutuoti šia tema. Taip pat yra daug panašių ST, kur vartotojai gali peržiūrėti ir vertinti įvairius dalykus: filmus, muziką ir gauti rekomendacijas atsižvelgdami į vartotojo skonį.

Atsižvelgiant į ankščiau pateiktą ST apibrėžimą (Bouyd 2007) (Palau 2004), pirmoji panaši sistema buvo sukurta 1997 m. SixDegrees.com leido vartotojams susikurti savo asmeninius profilius, draugų sąrašus ir po metų atsirado galimybė peržiūrėti draugus, esančius draugų sąrašė. Žinoma šios galimybės egzistavo ir anksčiau, bet atskirai. Pavyzdžiui, asmeninius profilius susikurti buvo galima, bet be draugų sąrašo. Vėliau SixDegrees.com atsirado galimybė siųsti asmenines žinutes draugams. Nuo 1997 m. iki 2001 m. daug panašių ST pradėjo naudoti įvairius asmeninių profilių ir jiems priklausančių draugų kombinacijas. AsianAvenue, BlackPlanet ir MiGente leido vartotojams susikurti įvairius profilių tipus – asmeninius, profesinius, pasimatymų. 2000 m. švedų internetinė bendruomenė LunaStorm įtraukė draugų sąrašams, svečių knygoms ir tinklaraščiams skirtus puslapius. Kitas svarbus ST buvo paleistas 2001 m. ir pavadintas Ryze.com. Šis ST apėmė tik komercinius tinklus, pirmiausia vartotojus tik iš San Francisko komercijos ir technologijų bendruomenės (Bouyd 2007). Šiomis dienomis vienas populiariausių verslo ST yra LinkedIn⁴, kuris vienija apie 65 milijonus vartotojų, kurie keičiasi mintimis, idėjomis, galimybėmis. Taip pat šiuo metu vieni iš populiariausių socialinių tinklų, apimančių milijonus vartotojų, yra Facebook⁵, Friendster⁶ ir MySpace⁷.

RS yra plačiai naudojamos interaktyvioje erdvėje tam, kad būtų galima pasiūlyti tam tikrus dalykus vartotojams, kuriems galimai tie dalykai turėtų patikti. Šios sistemos tobulinamos todėl, kad vartotojai galėtų sutaupyti savo laiką ieškodami pagal jų poreikius reikalingų produktų. Nes šiuolaikinėje interaktyvioje erdvėje yra galybė produktų ir surasti sau tinkamiausią yra pakankamai sudėtinga. Komercinis tinklas Amazon⁸ vartotojams rekomenduoja produktus, atsižvelgdamas į vartotojų jau įsigytus produktus bei paieškos raidą. Tuo tarpu ST MovieLens leidžia vartotojams vertinti filmus didelėje filmų duomenų bazėje ir pateikia rekomendacijas pagal vartotojų įverčių raidą.

³ http://en.wikipedia.org/wiki/Social_web

⁴ <http://www.linkedin.com/>

⁵ <http://www.facebook.com/>

⁶ <http://www.friendster.com/>

⁷ <http://www.myspace.com/>

⁸ <http://www.amazon.com/>

RS raida yra geras būdas išvengti pagrindinės problemos – informacijos pertekliaus. Esant internete dideliui kiekiui produktų (knygos, filmai, muzika, automobiliai ir pan.) yra įmanoma vartotojams parekomenduoti visus produktų tipus. Sėkminga RS gali komercinei internetinei įmonei padidinti finansines įplaukas, nes produktas bus išigytas ten, kur vartotojui bus pasiūlyta reikalingas daiktas, labiausiai tinkantis pagal poreikius, ir vartotojas nesugaiš daug laiko paieškoms.

Paprastai rekomendacijos yra gaunamos skaičiuojant įverčių panašumą tarp skirtingų vartotojų. Tačiau parekomenduoti produkto neužtenka, reikia, kad vartotojas gautų informaciją apie tai, iš kur rekomendacija atėjo ir koks tai produktas. Daugelis žmonių intuityviai nusprendžia, kuriomis rekomendacijomis pasitikėti, kai jos yra gaunamos iš pažįstamų, giminių ar draugų. Todėl RS pagrindinis tikslas atkartoti šį sprendimų priėmimo procesą. Dėl šios priežasties yra labai svarbu suprasti, kaip žmonės priima sprendimus ir kurios rekomendacijos charakteristikos jiems yra svarbiausios.

Straipsnyje (Bonhard 2006) autoriai įrodinėja, kad vartotojais paremtų BF sistemų rekomendacijos yra pakankamai artimos rekomendacijoms realiame pasaulyje. Tai yra todėl, kad rekomendacijos yra paremtos panašiais vartotojais.

2.2.1. Pagrindinės RS charakteristikos

Rekomendacines sistemas galima būtų skirstyti pagal rekomendacijų pateikimo metodus (Adomavičius 2007). Labiausiai paplitę metodai yra šie (Balabanovic 1997):

- Turiniu paremtas metodas. Šis metodas analizuoja atitinkamo produkto turinį (metaduomenis, ypatumus, požymius, tekstą) ir pagal jį bandoma surasti susijusius produktus.
- Bendro filtravimo metodas. Metodas vartotojui rekomenduoja produktus, kuriems rekomendacija yra apskaičiuojama iš atrinktų panašių vartotojų, lygindamas produktų įverčių raidą.
- Mišrus metodas. Tai yra dviejų ar daugiau metodų kombinacijos (Adomavičius 2005).

Taip pat RS galima klasifikuoti pagal techniką, kuri taikoma algoritmams (Breese 1998):

- Atmintimi paremta technika. Paprastai ši technika remiasi tuo, kad analizuoja ir pateikia rekomendacijas pagal anksčiau pasirinktus vartotojo produktus.
- Modeliu paremta technika. Sistema tiria vartotojo aktyvumą ir analizuoja vartotojo praeityje atliktus veiksmus, kad galėtų susidaryti nuomonę apie vartotoją ir išmokti pateikti rekomendacijas pagal vartotojo skonį.

Rekomendacijų pateikimo metodus galima skirstyti dar smulkiau (1 lentelė) (Burke 2004). Taip pat galima sudaryti įvairius mišrius metodus. Būdai, pagal kuriuos gali būti daromi mišrūs

metodai aprašyti 2 lentelėje (Burke 2004). Hibridinės rekomendacinės sistemos jungia du ar daugiau rekomendavimo metodų siekiant gauti geresnius rezultatus.

1 lentelė. Rekomendacijų pateikimo metodai

METODAI	REIKALINGA INFORMACIJA	ĮVESTIES DUOMENYS	PROCESAI
BENDRO FILTRAVIMO	Vartotojų įverčiai produktams.	Vartotojų įverčiai produktams.	Surasti panašius vartotojus vartotojui V pagal įverčių raidą ir apskaičiuoti rekomendacijas produktui.
TURINIŲ PAREMTAS	Produktų požymiai.	Vartotojų įverčiai produktams.	Sukurti klasifikatorių, kuris atitinka vartotojo nuostatas ir naudoti jį produktams.
DEMOGRAFINIS	Demografinė informacija apie vartotoją ir jo įverčius produktams.	Demografinė informacija apie vartotoją.	Pagal vartotojo demografinę informaciją surasti panašius vartotojus ir jų įverčius naudoti skaičiavimams.
PRAKTIŠKUMO PAREMTAS	Produktų savybės.	Produktų savybės charakterizuojančios vartotojų nuostatas.	Nustatyti reikalavimus produktams ir produktus pagal atitinkančius reikalavimus suskirstyti į klases.
ŽINIOMIS PAREMTAS	Produktų savybės. Informacija parodanti kaip produktai atitinka vartotojų nuostatas.	Vartotojų nuostatų ar interesų aprašymas.	Išreikšti nuostatų atitikimą tarp produkto ir vartotojo.

Pirmasis *bendro filtravimo* metodas, kuris jau buvo minėtas, naudoja vartotojų įverčių raidą, pagal juos suranda panašius vartotojus ir naudodamas jų įverčius apskaičiuoja rekomendacijas neįvertintiems produktams ir pateikia vartotojui. Pagrindinis šio metodo privalumas, kad jis yra nepriklausomas nuo duomenų bazės tipo ir gali būti naudojamas bet kokiai produktų grupei (Zanker 2007).

Turinių paremta rekomendacija yra informacijos filtravimo tyrimo rezultatas (Belkin 1992). *Turinių paremta* rekomendavimo technika yra naudojama analizuoti vartotojo profilio nuostatomis, atitinkančioms produktų savybes, ir pagal jas surasti panašius produktus ir pateikti atitinkamas rekomendacijas.

Demografinė technika paremtos RS skirsto vartotojų asmeninius atributus (demografinę vartotojo informaciją) į tam tikras kategorijas, pagal kurias yra pateikiamos rekomendacijos. Šios technikos privalumas yra tas, kad ji nereikalauja vartotojų įverčių raidos, kad galėtų rekomenduoti produktą. Priešingai nei pirmosios dvi rekomendavimo technikos.

Paskutinės dvi rekomendavimo technikos analizuoja vartotojų nuostatas ir jas atitinkančių produktų parametrus. *Praktiškumo* technika (Huang 2008) pateikia pasiūlymus skaičiuodama kiekvieno produkto praktiškumo vertę tam tikram vartotojui pagal jo kriterijus. Tačiau pagrindinė

problema, kaip sukurti naudingumo funkciją kiekvienam vartotojui. *Žiniomis* paremta sistema (Burke, 1997) renka žinias, kaip konkretus produkto elementas gali tenkinti konkrečius vartotojų poreikius, ir pateikia rekomendacijas, kuriose atsispindi produkto savybių atitikimas vartotojo poreikiams.

2 lentelė. Mišrių metodų sudarymo būdai

SUDARYMO BŪDAI	APRAŠYMAS
ĮVERČIŲ	Rekomendacijos pateikimui kelių rekomendavimo technikų įverčiai sujungiami į vieną.
KEITIMO	Sistema pasirenka reikalingą rekomendavimo metodą priklausomai nuo esamos situacijos.
MAIŠYTAS	Rekomendacijos pateikiamos keliais skirtingais rekomendavimo metodais.
SAVYBIŲ KOMBINAVIMO	Skirtingų rekomendacijų duomenų šaltinių savybės sujungiamos į vieną rekomendavimo algoritmą.
PAKOPINIS	Pirmas rekomendavimo metodas nustato pradinį duomenį - kandidatą, kitas metodas iš šių kandidatų atrinka tinkamiausius.
SAVYBIŲ PAPILDYMO	Pirmos rekomendavimo technikos išvestis naudojama kitos rekomendavimo technikos įvesčiai.
META-LYGIO	Rekomendavimo technikos išmoktas modelis naudojamas kaip įvestis kitai technikai.

Labiausiai paplitęs ir dažniausiai rekomendacinėms sistemoms naudojamas yra bendro filtravimo metodas. Šis metodas šiame tiriamajame darbe yra tyrimo pagrindas. Tyrimui paimta duomenų bazė iš MovieLens socialinio tinklo (1 pav.), kuriame rekomendacijos vartotojams pateikiamos naudojant taip pat šį bendro filtravimo metodą.

Rekomendacinės sistemos paskutiniu metu yra gana plačiai tiriamos, todėl yra sukurta daug metodų ir technikų. Atliekama daug tyrimų šioje srityje. Tačiau pagrindinės idėjos ir priežastys, kurios šį tiriamąjį darbą padaro išskirtinį yra:

- Lietuvoje ši sritis nenagrinėjama.
- Socialiniai tinklai paskutiniu metu labai plinta.
- Į internetą yra perkeliama vis daugiau produktų ir paslaugų.
- Tyrimas atliekamas atsižvelgiant į svarbius kriterijus, kurie analizuojamuose užsienio tyrėjų eksperimentuose buvo nenagrinėjami arba mažai nagrinėjami:
 - panašumo slenksčio įtaka rezultatams,
 - vienuodų įvertintų produktų slenksčio poveikis rezultatams,
 - vartotojų grupių išskirstymas,
 - atskirų vartotojų grupių paplitimas duomenų bazėje,
 - konkrečių vartotojų grupių įtaka rekomendacinių sistemų pateikiamoms rekomendacijoms.

- Siūlomas metodas apima dviejų populiariausių rekomendacijų skaičiavimo metodų (bendro filtravimo ir turiniu paremto) pagrindines savybes.



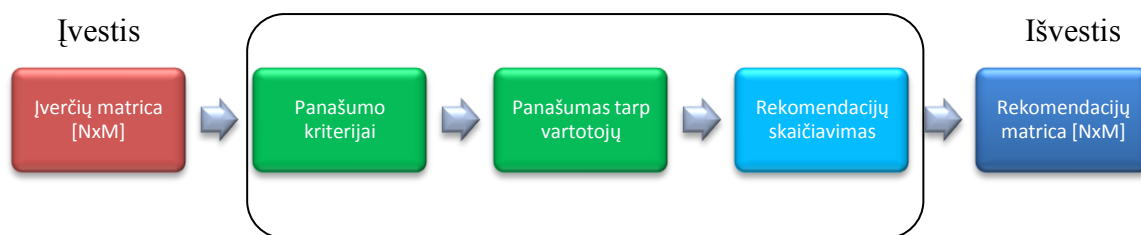
1 pav. Socialinio tinklalapio MovieLens namų puslapis

2.3. BF panašumo modelis

Realiame gyvenime renkantis naujus daiktus ar paslaugas, daugelyje situacijų sulaukiama patarimų ir pasiūlymų iš savo draugų, kuriais galima pasitikėti ir priimti jų siūlomą informaciją kaip naudingą. Mes girdime daug patarimų iš draugų apie muziką, filmus, knygas, automobilius, įvairias paslaugas ir pan. Todėl RS tyrėjai atkreipia dėmesį į šią situaciją ir bando tobulinti RS sistemas taip, kad jos galėtų rekomenduoti produktus vartotojams remiantis kitų vartotojų informacija.

BF metodas yra vienas sėkmingiausiai paplitusių ir plačiausiai naudojamas interaktyvioje erdvėje. Jį naudoja internetinės kompanijos ir elektroninės parduotuvės: Amazon (Linden 2003), Ebay, Google News (Das 2007) ir kt. BF metodas filtruoja visus vartotojus ir suranda panašiausius atitinkamus vartotojus. Jiems yra pateikiamos rekomendacijos apskaičiuotos naudojant panašiausių vartotojų įverčių raidą. Sudėtingiausia šio metodo vieta yra panašumo tarp vartotojų apskaičiavimas lyginant visų vartotojų įverčių raidą.

Prieš aprašant kaip apskaičiuoti panašumo reikšmes, būtina aptarti paprastos rekomendacinės bendro filtravimo sistemos architektūrą. Iš pradžių turime įvesties matricą. Kitais dviem etapais skaičiuojame panašumą tarp vartotojų ir spėjinių rekomenduojamam konkrečiam vartotojui. Pabaigoje turime išvesties matricą. 2 pav. matome visą BF RS procesą.



2 pav. Bendro filtravimo rekomendacinės sistemos architektūra

Įvestis. Sistema turi turėti pradinį duomenį tam, kad galėtų apskaičiuoti rekomendacijas. Paprastai įvesties duomenys yra visi įverčiai, kuriuos yra priskyre visi vartotojai atitinkamiems produktams. Tai vadinama įverčių matrica $N \times M$. Eilutėse N yra visi vartotojai registruoti sistemoje, stulpeliuose M yra visi vartotojų įvertinti produktai. Matricoje esanti informacija yra visos įverčių reikšmės atitinkamiems produktams iš atitinkamų vartotojų.

Pirmas etapas. Tikslas yra kiekvienam atskiram vartotojui sugeneruoti panašių vartotojų sąrašą. Šiam etapui atlikti sistemai yra reikalingas algoritmas, apskaičiuojantis panašumą tarp vartotojų pagal pasirinktus *panašumo kriterijus*. Algoritmas gali būti pasirinktas pagal įverčių raidą, paskyrų panašumą, vartotojų giminiškumą ir pan. Detaliau šis algoritmas bus analizuojamas 2.3.1 poskyryje. *Panašumas tarp vartotojų* etapu sukuriamą matricą, kuri susideda iš panašumo reikšmių tarp visų vartotojų.

Antras etapas. *Rekomendacijų skaičiavimo* tikslas yra naudojant panašių vartotojų duomenis (pavyzdžiui, įverčių raidą) ir panašumo reikšmes tarp vartotojų (iš *panašumo tarp vartotojų* matricos, kuri sudaroma pirmame etape) apskaičiuoti, kokia reikšmė vartotojas gali įvertinti dar nematytą produktą. Populiariausias yra Resnicko algoritmas, kuris bus išanalizuotas 2.3.1 poskyryje.

Išvestis. Paskutinis etapas yra sukurti išvesties matricą, kuri susideda iš vartotojų N , produktų M , kurių šie vartotojai dar neįvertino, ir sistemos apskaičiuotų spėjinių šiems produktams. Šie spėjiniai yra skaičiuojami antrame etape. Pirmiausia šie spėjiniai naudojami naujai sistemai įvertinti (sistemos įvertinimas atliekamas naudojant realią duomenų bazę ir tam tikrus įverčių vienetų, kurie aprašyti 4.1 skyriuje). Jeigu įvertinus sistemą spėjiniai yra pakankamai tikslūs, tada tokį metodą galima taikyti interaktyvioje erdvėje su tikrais vartotojais rekomenduojant jiems produktus.

2.3.1. Panašumo kriterijai ir įverčių spėjimas

Paprastai BF sistemos remiasi panašumo prielaida, kuri sako, kad panašūs profiliai (panašūs savo įverčių raida) yra geri partneriai rekomendacijoms gauti. RS, kuri ieško panašių partnerių, turi apskaičiuoti panašumą tarp šių vartotojų. Vartotojai, kurie yra panašūs į vartotoją, kuriam pateikiamos rekomendacijos, yra laikomi kaimynais šiam vartotojui. Kaimynų suradimas (Vozalis 2003) susideda iš dviejų etapų:

- Skaičiuojamas panašumas tarp vartotojų naudojant panašumo kriterijus.
- Atrenkami panašiausi vartotojai naudojant tam tikrą panašumo slenkstį.

Panašumas tarp dviejų vartotojų skaičiuojamas naudojant Pearsono koreliacijos koeficientą (toliau – PKK) (Rashid 2006) (Breese 1998):

$$panašumas(g, r) = \frac{\sum_{p \in produktai} (g(p) - \bar{g})(r(i) - \bar{r})}{\sqrt{\sum_{p \in produktai} (g(p) - \bar{g})^2 \sum_{p \in produktai} (r(i) - \bar{r})^2}}. \quad (1)$$

Šios formulės paaiškinimas nėra sudėtingas, kaip gali pasirodyti iš pirmo žvilgsnio. Visų pirma išskirkime formulėje naudojamus du vartotojus:

- g – tai vartotojas, kuris gauna rekomendaciją ir priima galutinį sprendimą, ar ta rekomendacija jam yra naudinga ar nenaudinga.
- r – tai vartotojas, kuris yra panašus vartotojui g , yra atrinktas kaip kaimynas ir sistema jį naudoja kaip rekomenduojantį vartotoją.

Svarbu pažymėti, kad sumos ženklas formulėje nurodo, kad panašumas yra skaičiuojamas naudojant visus produktus p , kuriuos abu vartotojai g ir r yra įvertinę. Įverčiai, kuriuos šie vartotojai suteikė konkrečiam produktui p yra $g(p)$ ir $r(t)$. Visų įverčių, kuriuos vartotojai yra suteikę skirtingiems produktams, vidurkis yra $\bar{g}$ ir $\bar{r}$. Šios formulės skaitiklis yra vartotojų įvertintų produktų p įverčių nuokrypio nuo įverčių vidurkio, sandaugų suma. Formulės vardiklis normalizuoja panašumo rodiklį. Panašumo rodiklis tarp vartotojų yra nuo 0 iki 1. Kuo rodiklis yra artimesnis 0, tuo vartotojai yra skirtingesni ir atvirkščiai, kuo panašumo rodiklis artimesnis vienetui tuo vartotojai yra panašesni.

Daugelis eksperimentų parodė, kad kiti panašumo vertinimo rodikliai RS neatspindi panašumo tarp vartotojų taip gerai kaip PKK. Dėl šios priežasties šio tiriamojo darbo eksperimentams panašumo skaičiavimui tarp vartotojų buvo pasirinktas PKK.

Taigi, kai sistema apskaičiuoja panašumą tarp visų vartotojų g ir r , tada sistemai reikia sugeneruoti rekomendacijas atitinkamiems produktams p . Šiam tikslui įgyvendinti sistema naudoja Resnicko spėjimo formulę:

$$g(i) = \bar{g} + \frac{\sum_{r \in R(i)} (r(i) - \bar{r}) panašumas(g, r)}{\sum_{r \in R(i)} |panašumas(g, r)|}. \quad (2)$$

Šia formule apskaičiuojama rekomendacija $g(i)$ produktui i , kurio vartotojas g dar neįvertino, o atrinkti panašūs vartotojai r iš sąrašo R produktą i įvertino tam tikrais balais. Formulės skaitiklis yra panašumo kriterijaus tarp vartotojų g ir r ir vartotojo r produkto i įverčio nuokrypio nuo įverčių vidurkio, sandaugų suma tarp visų panašių vartotojų iš sąrašo R . Pagrindinis svarbiausias rodiklis, kuris labiausiai lemia galutinį rezultatą, yra panašumo tarp vartotojų reikšmė. Kuo labiau vartotojai panašūs, tuo labiau didėja rekomendacijos tikslumas.

3. BF MODELIO TOBULINIMAS

Ankstesniuose skyriuose buvo analizuojamos rekomendacinės sistemos bei socialinis tinklas bendruoju principu. Buvo pristatytos technologijos kaip sistema skaičiuoja spėjinius, apibrėžti panašumo kriterijai (Resnicko spėjimo formulė). Šiame skyriuje bus nagrinėjama standartinė BF rekomendacinė sistema, pristatomas siūlomasis modelis, lyginamos modelių veikimo schemas, aprašomas vartotojų pasiskirstymas ir aiškos rekomendacijos.

3.1. Pagrindinė rekomendacinė sistema

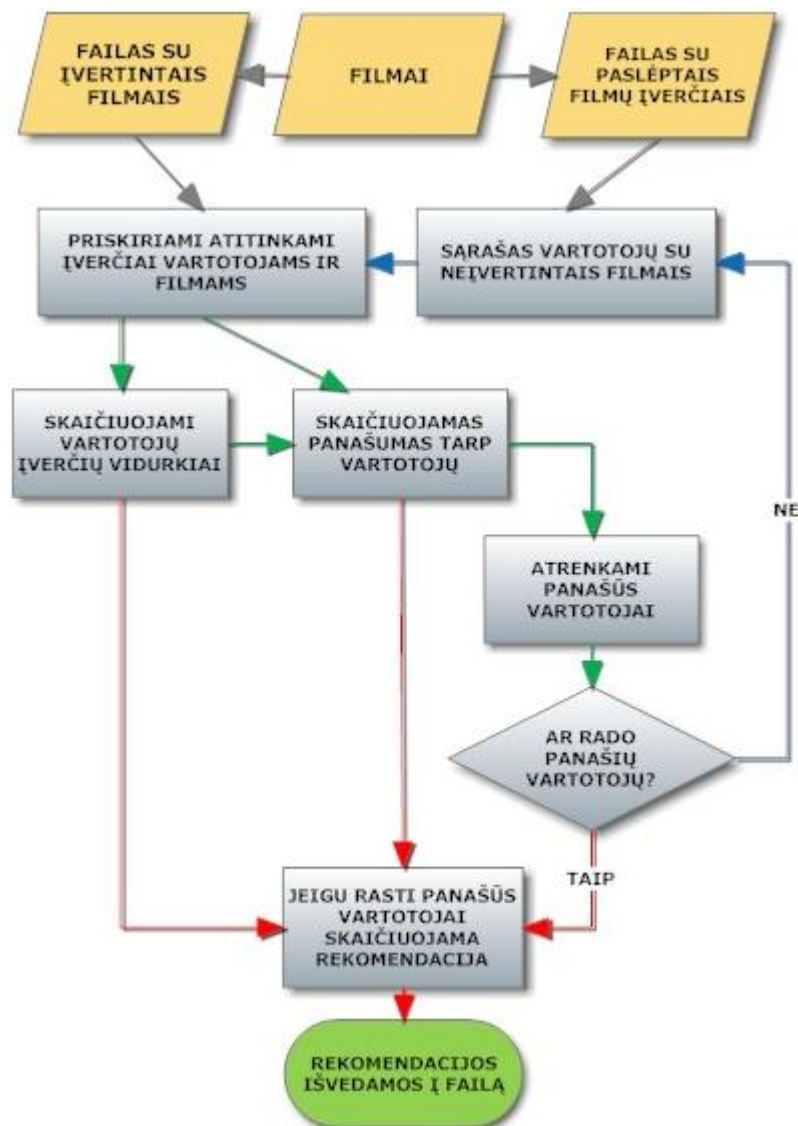
Pagrindinis šio tiriamojo darbo tikslas pagerinti standartinės rekomendacinės sistemos skaičiavimo efektyvumą, pasiūlant naują metodą standartinio metodo patobulinimui. Šiam darbui yra pasirinkta atvirojo kodo rekomendacinė sistema (JWeb), kurią sukūrė Jasonas Weberis 2009 metais ir patalpino savo internetiniame puslapyje⁹. Ši standartinė RS buvo sukurta Netflix konkursui. Pagrindiniai šios RS komponentai:

- Bendras filtravimo metodas.
- Pearsono koreliacijos koeficientas.
- Resnicko spėjimo formulė.

Ši RS naudoja duomenų rinkinį, kurį sudaro vartotojai, produktai ir įverčiai, paimtą iš Netflix konkurso. Surasdama panašius vartotojus ši RS gali spėti galimus vartotojų įverčius. Produktų sąrašą sudaro tik filmai. Šiai programai buvo atrinkti 4000 vartotojų ir 3000 filmų iš Netflix duomenų bazės. Kiekvienas vartotojas yra įvertinęs tik dalį filmų. Kūrėjai atsitiktinai parinko 1600 vartotojų ir 1200 filmų iš testavimui skirtą sąrašą. Likę vartotojai ir filmai buvo naudojami sistemos treniravimui. Sistemos tikslas yra atspėti įverčius, kuriuos atrinkti vartotojai duos neįvertintiems filmams. Vėliau šie atspėti įverčiai gali būti lyginami su realiais įverčiais taip analizuojant sistemos veikimo efektyvumą.

Ši RS veikimui reikalingi trys įvesties failai: filmų, neįvertintų filmų (pasirinkti filmai sistemos testavimui) ir įvertintų filmų (vartotojų įverčiai filmams). Šiuos failus sistema naudoja susikurti testavimo ir sistemos treniravimo sąrašams. Sistema šiuos sąrašus naudoja skaičiuoti vartotojų įverčių vidurkiams ir panašumui tarp vartotojų. Jeigu vartotojai atitinka panašumo kriterijus, jie yra naudojami apskaičiuoti rekomendacijai naudojant Pearsono koreliacijos koeficientą. Gauti rezultatai išvedami į failą, kuris vėliau gali būti naudojamas sistemos efektyvumo vertinimui. Ši programa parašyta C# kalba.

⁹ <http://www.jweb.us>



3 pav. Programos veikimo schema

3.2. Siūlomas modelis su kategorijomis

Šiame poskyryje bus aprašomas siūlomas rekomendacinėse sistemose naudojamo BF panašumo modelio patobulinimas ir pagrindinės idėjos ir siūlymai, kurie eksperimentinėje dalyje bus taikomi ir tiriama praktiškai su realia duomenų baze.

Dažniausiai standartinė RS naudoja panašumo arba pasitikėjimo reikšmes. Šios reikšmės priskiriamos konkrečiam vartotojui ir apima visus aspektus. Tačiau realybėje mes pažinodami draugus ar artimuosius, kurių klausiamo patarimo arba rekomendacijos vienu ar kitu klausimu, mes žinome, kokios jų žinios ir požiūris į tam tikrus skirtingus dalykus. Dėl šios priežasties mes galime pasirinkti žmones, iš kurių norime gauti rekomendacijas, susijusias su konkrečiu dalyku. Kaip pavyzdį galima paimti filmus. Jeigu žinome, kad draugas mėgsta komedijas, bet nemėgsta trilerių, tai mes neklausime patarimo, koki trilerį jis patartų pažiūrėti. Arba jeigu pažįstamas mėgsta skaityti knygas, bet nesidomi muzika, tai mes neprašysime, kad jis parekomenduotų paklaudyti tam tikro atlikėjo dainų.

3.2.1. Kategorijos

Vertinant vartotojų skonį atskirais aspektais darome prielaidą, kad yra naudinga produktą išskirstyti į atskiras kategorijas arba atskirus produktų tipus, o panašumo reikšmės tarp vartotojų skaičiuoti skirtingose kategorijose. Todėl tam tikram vartotojui skirtingose kategorijose gali būti panašūs visai kiti vartotojai ir panašumo reikšmės bus skirtingos. Duomenų rinkinys gali būti pavaizduojamas kaip lentelė kiekvienam atskiram vartotojui (1 lentelė) arba atskirai kategorijai (2 lentelė). Priklausomai nuo kategorijos iš kurios vartotojas nori gauti rekomendaciją, jam bus surasti atitinkami panašūs vartotojai. Tam, kad sistema atrinktų panašiausius vartotojus, sistemai turi būti nustatytas tam tikras panašumo slenkstis.

Kaip pavyzdį paimkime vartotoją vardu Jonas, o panašumo slenkstį sistemai nustatykime 0,5. Sistema jam turi parekomenduoti knygą ir muzikos kūrinių, kurių Jonas dar neįvertino. Jono duomenis turime 1 lentelėje. Taigi sistemai parekomenduoti Jonui knygą atrinks Giedrių ir Algirdą, o parekomenduoti muzikos kūriniui atrinks Vytautą ir Algirdą. Čia galime išžvelgti kategorijų privalumą, nes jeigu Jono su visais kitais vartotojais panašumo rodiklis būtų bendras visoms kategorijoms, tada jam būtų neteisingai parinkti panašūs vartotojai. Šiuo atveju Vytautas netinka rekomenduoti Jonui knygos, o Giedrius netinka rekomenduoti muzikos kūriniui.

3 lentelė. Panašumo reikšmės pasirinktam vartotojui

	Muzika	Filmai	Knygos	Automobiliai
Giedrius	0.45	0.32	0.91	0.58
Vytautas	0.69	0.14	0.48	0.92
Algirdas	0.89	0.90	0.60	0.13

Kitas būdas pavaizduoti panašumo duomenis tarp vartotojų yra išskiriant atskirą kategoriją. Tokiu atveju turime vartotojų matricą (2 lentelė) su panašumo tarp jų reikšmėmis. Kaip matome iš pateiktos pavyzdinės duomenų lentelės, panašumas tarp vartotojų yra toks pats, nepriklausomai ar skaičiuosime Vyginto panašumą Antanui, ar Antano panašumą Vygintui. Taip yra dėl PKK (2.3.1 skyrius, 1 formulė), kur yra imami tik bendrai vartotojų įvertinti produktai ir kiti vartotojų įvertinti produktai neturi reikšmės šiam rezultatui. Tačiau kitokia situacija yra su tomis pačiomis panašumo reikšmėmis skaičiuojant spėjinį neįvertintam produktui, čia (2.3.1 skyrius, 2 formulė) jau turi reikšmę atskiro vartotojų visų įvertintų produktų įverčiai ir jų vidurkis.

4 lentelė. Panašumo reikšmės tarp vartotojų pasirinktai kategorijai

	Antanas	Vygintas	Ona	Ramunė
Antanas		0.56	0.87	0.45
Vygintas	0.56		0.74	0.82
Ona	0.87	0.74		0.11
Ramunė	0.45	0.82	0.11	

3.2.2. Aiškos rekomendacijos

Antra idėja šiam metodui patobulinti yra „aiškos rekomendacijos“ (2 pav.). Šis patobulinimas gali būti skirtas ne konkrečiam metodui, bet tiesiog rekomendacinei sistemai. Nes vartotojai labiau pasitiki rekomendacine sistema, kuri pateikia aiškas, suprantamas ir su pakankamu kiekiu informacijos rekomendacijas.

Vartotojai teikia pirmenybę toms rekomendacijoms, kurios yra pakankamai aiškos ir yra matoma iš kokių vartotojų atėjo rekomendacija. Tokiu atveju vartotojas gali pasirinkti vartotojus – favoritus, kuriais jis gali pasitikėti gaudamas kitas rekomendacijas. Taip pat vartotojas jausis patogiai naudodamasis tokia RS. Žinoma, jeigu vartotojas naudosis jam patinkančia RS, jis taip pat ją rekomenduos ir savo draugams ar pažįstamiems. Tokiu būdu žmonių ST tokioje RS plėsis daug sparčiau. Į gautą rekomendaciją įdėtos nuorodos į labiausiai panašius ar patikimus vartotojus gali būti vienas iš būdų parodyti rekomendacijos turinį. Taip pat būtų naudinga parodyti, kaip kiti vartotojai, iš kurių ateina rekomendacija, įvertino tam tikrą produktą.

REKOMENDACIJA	REKOMENDUOJANTYS/Įvertinimai	
„INFILTRUOTI“ Kategorija <Filmai> RS pasiūlytas įvertis: 8	<u>Tomas</u>	10
	<u>Tadas</u>	8
	<u>Laima</u>	6
	<u>Rita</u>	7
„KAZINO“ Kategorija <Knygos> RS pasiūlytas įvertis: 9	<u>Marius</u>	10
	<u>Ramunė</u>	8
	<u>Rita</u>	8
	<u>Tautvydas</u>	9

4 pav. Aiški rekomendacija

RS patikimumui pagerinti siūloma RS papildyti aiškiomis rekomendacijomis, kurios pavyzdys pateiktas 4 paveiksle. Pavyzdyje galima matyti, kad vartotojui yra pasiūlyta filmas ir knyga su atitinkamai siūlomais įverčiais, kurie priklauso nuo atrinktų panašių, rekomenduojančių vartotojų įverčių.

3.2.3. Vartotojų skirstymas

Daugelis tyrėjų ir RS algoritmų naudoja visus vartotojus eksperimentui neskirstydami jų į grupes. Todėl jų eksperimentų rezultatai nėra visiškai tikslūs. Tam, kad pagerinti RS iš pradžių reikia atkreipti dėmesį į didžiąją dalį vartotojų sudarančią grupę – šalti vartotojai arba kitaip pasyvūs vartotojai. Pasyvūs vartotojai neturi įvertinę daug skirtingų produktų ir jų įverčių raida yra pakankamai trumpa. Čia slypi ir pagrindinė problema. Dėl šios priežasties RS suranda labai mažai panašių vartotojų arba nesuranda jų visai. Dėl to yra sunku pasyviems vartotojams pateikti rekomendacijas, o jeigu yra pateikiamos tai dažniausiai yra netikslūs. Kadangi pasyvus vartotojas gauna blogas rekomendacijas arba negauna jų visai, tai toks vartotojas gali nustoti naudotis šia RS.

Kitas tipas, į kurį galima išskirti vartotojus - yra aktyvūs vartotojai. Šie vartotojai, priešingai nei pasyvūs, yra įvertinę pakankamai daug filmų ir turi ilgą įverčių raidą. Aktyvūs vartotojai taip pat gali turėti poveikį galutiniams eksperimentų rezultatams. Vartotojų įverčių raida apima didžiąją dalį duomenų bazėje esančių įverčių, todėl sistema gali jiems apskaičiuoti tikslesnius spėjinius. Jeigu eksperimentui atsitiktinai parinktume vartotojus, kurių didžioji dalis būtų pasyvūs vartotojai, tada RS spėjamų įverčių nuokrypis nuo realių reikšmių bus mažesnis, atitinkamai bus priešinga situacija, jei didžioji dalis vartotojų duomenų bazėje bus pasyvūs. Todėl yra būtina ištirti atskirų vartotojų grupių poveikį RS tikslumui.

Todėl darbe yra išskiriamos trys vartotojų kategorijos:

- Pasyvūs – turintys trumpą įverčių raidą.
- Aktyvūs – turintys įvertinę didelę dalį produktų.
- Tarpiniai – visi kiti vartotojai, esantys tarp aktyvių ir pasyvių.

Konkretus vartotojų išskirstymas ir dalijimas į grupes bei tų grupių dydžio ir kriterijų nustatymas darbe atliekamuose bandymuose aprašomas 4.1 ir 4.2 skyriuose.

3.3. Siūlomo modelio schema ir palyginimas su standartiniu BF modeliu

Standartinis BF modelis ir siūlomas modelis tarpusavyje turi keletą panašumų, bei nemažai skirtumų. Pagrindinis panašumas yra tas, kad abu modeliai naudoja tas pačias formules apskaičiuoti panašumui tarp vartotojų ir apskaičiuoti galutinei rekomendacijai tam tikram vartotojui, tačiau visas kitas sistemos veikimo procesas yra skirtingas. 4 paveikslėlyje pateikiama abiejų modelių bendra schema. Atskiros schemas dalys pažymėtos skirtingomis spalvomis. Oranžinė ir geltona spalva žymi duomenis (vartotojus, vartotojų ID, filmų ID, įverčius, neįvertintus filmus) reikalingus abiem modeliams. Abu modeliai išskirti skirtingomis spalvomis: siūlomo modelio sistemos visi etapai žymimi žalia spalva, o standartinio modelio – mėlyna. Bendri abiejų modelių etapai žymimi alyvine spalva.

Pirmame etape abu modeliai naudoja tą patį duomenų failą, kuriame yra aprašomi visi neįvertinti vartotojų filmai. Yra imamas pirmas įrašas, kuris sudarytas iš vartotojo unikalaus

numerio ir filmu unikalaus numerio. Toliau modeliai atlieka skirtingus veiksmus. Standartinis modelis skaičiuoja visų vartotojų įverčių vidurkius, bei ieško panašių vartotojų, tam vartotojui, kuris yra įrašė paimtame iš neįvertintų filmų sąrašo, lygindamas vartotojo iš sąrašo su likusiais vartotojais įverčių raidą.

Siūlomas modelis šiame sistemos veikimo etape paima visus įvertintus filmus su įverčiais ir filmus įvertinusiems vartotojais ir pagal filmų žanrus suskirsto visus filmus į 7 kategorijas. Tada paimtą iš sąrašo neįvertintą filmą su vartotoju, kuris jo neįvertino, sistema atrenka, kuriai kategorijai neįvertintas filmas priklauso. Kitu etapu sistema atsirenka tik tai konkrečiai kategorijai priklausančius filmus ir visus vartotojus su tų filmų įverčiais. Sistema naudodama tik tos kategorijos įverčius, lygina vartotojų įverčių raidą su vartotoju iš neįvertintų filmų sąrašo ir suranda panašius jam vartotojus bei apskaičiuoja tik toje kategorijoje esančių įverčių vidutinės reikšmės kiekvienam vartotojui. Kitas etapas yra vienodas abiemis metodams, kuriame yra nustatomi vartotojų panašumo kriterijai:

- Vartotojų panašumo slenkstis (VPS).
- Vienodų įvertintų filmų slenkstis (VIFS).

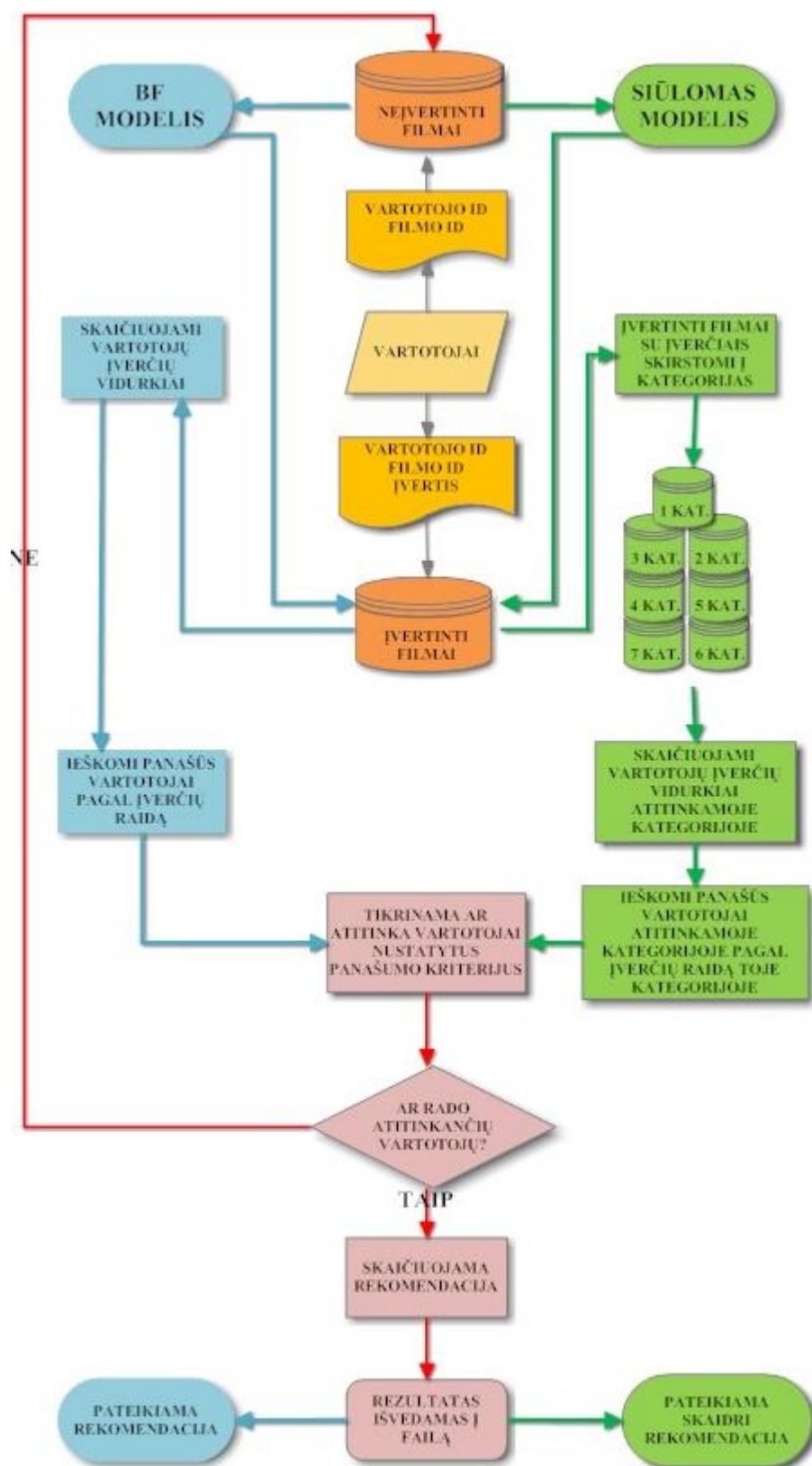
Nustatant VPS yra atrenkami panašiausi vartotojai. VPS reikšmės gali būti nuo 0 iki 1 (2.3.1 skyrius). Reikia pastebėti, kad ši reikšmė labai stipriai lemia įverčių apimtį, todėl ji turi būti nustatoma individualiai kiekvienai duomenų bazei. Su aukštu VPS sistema atrinks pačius panašiausius vartotojus, kurie įtakos tikslesnius spėjinius. Tačiau tikėtina, kad jeigu duomenų bazėje bus daug vartotojų ir produktų, o atitinkamai nedidelis įverčių skaičius, sistema sunkiai suras panašius vartotojus pagal aukštą VPS, dėl to labai stipriai sumažės rekomendacijų kiekis vartotojams ir nukentės sistemos efektyvumas.

Kitas taip pat svarbus slenkstis, nustatomas vartotojams, yra VIFS. Pastebėta, kad šis slenkstis nebuvo naudotas nė viename modelyje, kurie buvo nagrinėjami analizuojant su darbu susijusią literatūrą. VIFS buvo panaudotas standartinio BF metodo atviro kodo rekomendacinėje sistemoje, kuri yra naudojama kaip pagrindas šiam tiriamajam darbui. Todėl šio slenksčio įtaka taip pat yra tiriama ir analizuojamas jo poveikis RS efektyvumui (4.3 skyrius). Yra daroma prielaida, kad VIFS didinimas gali lemti rekomendacijų tikslumo gerėjimą, tačiau sumažinti rekomendacijų kiekį, taip pat kaip ir naudojant VPS. Todėl jis turi būti taikomas atsižvelgiant į duomenų bazės dydį: vartotojų, produktų ir įverčių skaičių.

Jeigu vartotojas atitinka sistemai užduotus kriterijus, tada jis yra naudojamas rekomendacijai skaičiuoti. Tai yra bendras etapas, kuriuo yra skaičiuojama konkreti rekomendacija kiekvienam vartotojui, abiemis metodams. Jame yra naudojama Resnicko formulė (2.3.1 skyrius).

Paskutiniame RS etape yra pateikiamos rekomendacijos vartotojams. Standartiniame metode yra tiesiog pateikiamas sistemos spėjiny vartotojui kaip rekomendacija, kurioje vartotojas mato

įvertį, kuriuo jis galimai įvertins nematytą filmą. Tuo tarpu siūlomame modelyje yra siūloma pateikti vartotojui skaidrią rekomendaciją (3.2.2 skyrius), kurioje vartotojas matytų ne tik galimą savo įvertį, tačiau kokių vartotojų ši rekomendaciją atėjo, bei kaip rekomenduojamą filmą įvertino jam panašūs vartotojai.



5 pav. Standartinio BF modelio ir siūlomo modelio palyginamoji schema

4. LYGINAMASIS MODELIŲ EKSPERIMENTAS

4.1. Vertinimo kriterijai

Įvertinti rekomendacinės sistemos vertę yra naudojama nemažai kriterijų. Vieni iš pagrindinių kriterijų yra tie, kurie vertina rekomendacinės sistemos skaičiavimo paklaidą ir įverčių apimtį. Tačiau daugeliu atveju tyrėjai nevertina RS sistemos naudingumo atskirose vartotojų grupėse. Dėl šios priežasties jų rezultatai gali būti ne visiškai tikslūs atsižvelgiant į skirtingas vartotojų grupes. Kad pagerintumėme RS efektyvumą, pradžioje turime atkreipti dėmesį į didžiąją vartotojų dalį – pasyvius vartotojus, todėl, kad šie vartotojai gauna mažiausią dalį rekomendacijų ir jos yra pakankamai netikslios. Kita svarbi grupė vartotojų – aktyvūs vartotojai. Jie gauna didžiausią dalį rekomendacijų ir dažniausiai nulemia galutinį RS sistemos tikslumą. Šie vartotojai apima didžiausią įverčių dalį ir dažniausiai jiems yra apskaičiuojami tikslesni spėjiniai nei kitoms vartotojų grupėms. Jeigu atsitiktinai paimsime produktų įverčių rinkinį, kurio rezultatus lyginsime su RS spėjimais, o didžioji dalis įverčių bus aktyvių vartotojų, tada sistemos tikslumas bus pakankamai aukštas, tačiau atskiroms grupėms jis skirsis. Šis teiginys gali būti taikomas ir iš pasyvių vartotojų pusės.

Įvertinti šiame darbe lyginamų RS modelių efektyvumui yra naudojami trys vertinimo kriterijai, kurie parodys lyginamų metodų tikslumą ir įverčių apimtį skirtingoms vartotojų grupėms.

Pagrindinė absoliuti paklaida (PAP) – vertinimo kriterijus, kuris yra vidurkis absoliutaus nuokrypio tarp RS spėjinių ir tikrųjų įverčių (Papagelis 2004). Tai yra metodas, kai sistema paslepia tikrąjį įvertį ir bando jį atspėti naudodama pasirinktą RS metodą (BF panašumo arba siūlomą metodą su kategorijomis). Toliau yra apskaičiuojamas įverčio skirtumas tarp tikrojo ir spėjamo įverčio. Pabaigoje visiems šiems skirtumams yra apskaičiuojamas vidurkis – PAP (skyrius 4.3). Trumpai, PAP matuoja RS tikslumą. Tikslumas yra svarbus vartotojams, kuriems pateikiamos rekomendacijos. Daugelis RS su aktyviais vartotojais dirba pakankamai gerai, tačiau prastai su pasyviais. To priežastis yra, kad pasyviems vartotojams visų spėjinių paklaida yra didesnė, o aktyvių – mažesnė. Kadangi aktyvių vartotojų įverčių yra dauguma ir jiems apskaičiuojama dauguma spėjinių, todėl sistemos PAP yra pakankamai geras, o didelė paklaida pasyviems vartotojams išnyksta dėl įverčių mažumos. Todėl yra naudojamas papildomas vertinimo kriterijus – PAVP, tam, kad PAP būtų lygiaverčiai skirtingose vartotojų grupėse.

Pagrindinė absoliuti vartotojo paklaida (PAVP) – pradžioje apskaičiuojamas PAP kiekvienam atskiram vartotojui ir atskirų vartotojų PAP apskaičiuojamas vidurkis. Šiuo būdu kiekvienas vartotojas turi vienodą svorį PAVP skaičiavime (Massa 2004). Tai yra labai svarbu esant dideliame kiekiui pasyvių vartotojų (skyrius 4.3).

Įverčių apimtis (IA) – tam tikras procentas reikšmių produktams, kuriuos sistema gali parekomenduoti bent vieną kartą (Massa 2004). Šis vertinimo kriterijus parodo RS galimybę parekomenduoti kiek įmanoma daugiau produktų vartotojams. Pavyzdžiui, jeigu turime duomenų bazę su 1000 skirtingų neįvertintų produktų ir pasyviems vartotojams sistema parekomendavo 50 skirtingų produktų, tada IA pasyviems vartotojams bus 5% (skyrius 4.3). Svarbu yra pažymėti, kad skaičiuojant IA, pasiskirstymas skirtingų neįvertintų produktų yra netolygus tarp atskirų vartotojų grupių. Pasiskirstymui paaiškinti paimkime pavyzdį. Turime duomenų bazėje 1000 filmų, kuriems sistema turi apskaičiuoti spėjinius ir parekomenduoti vartotojams. Sistema sugebėjo pateikti 800 rekomendacijų iš jų 100 pasyviems vartotojams, 500 – aktyviems ir 200 – tarpiniams. Tokiu atveju **įverčių dalis (ID)** vartotojų grupėms atitinkamai bus – 12,5%, 62,5% ir 25%, o sistemos IA atskiroms grupėms atitinkamai bus – 10%, 50% ir 20 %.

Šie skaičiavimai atliekami pagal šias formules:

$$ID = \frac{RSVG}{BRS} \cdot 100, \quad (3)$$

$$IA = \frac{RSVG}{NFS} \cdot 100. \quad (4)$$

RSVG – rekomendacijų skaičius vartotojų grupei. Šiuo atveju būtų atitinkamai 100, 500 ir 200.

BRS – bendras rekomendacijų skaičius. Pavyzdyje tai yra 800.

NFS – neįvertintų filmų skaičius. Pavyzdyje tai yra 1000.

4.2. Duomenų rinkinys

BF modelio ir siūlomo modelio testavimui ir rezultatų vertinimui buvo pasirinkta reali filmų duomenų bazė (GroupLens). Ši duomenų bazė yra sudaryta GroupLens tyrėjų grupės iš jų sukurto MovieLens socialinio tinklo. Šis ST skirtas filmams.

Eksperimentams naudojama duomenų bazę sudaro:

- 1339 filmai,
- 943 vartotojai,
- 77639 įverčiai.

Pagal siūlomą modelį visi filmai buvo išskirstyti į 7 kategorijas:

- 1 kategorija – veiksmo, veiksmo moksliniai,
- 2 kategorija – nuotykių, animaciniai, nuotykiu vaikams,
- 3 kategorija – komedijos, animacinės komedijos vaikams,
- 4 kategorija – kriminalai,
- 5 kategorija – dramos, karinės dramos,
- 6 kategorija – romantiniai,
- 7 kategorija – trileriai, mistiniai trileriai, siaubo trileriai.

FILMO ID	FILMO PAVADINIMAS	1 GRUPĖ	2 GRUPĖ	3 GRUPĖ	4 GRUPĖ	5 GRUPĖ	6 GRUPĖ	7 GRUPĖ
1	Toy Story (1995)	0	0	1	0	0	0	0
2	GoldenEye (1995)	1	1	0	0	0	0	1
3	Four Rooms (1995)	0	0	0	0	0	0	1
4	Get Shorty (1995)	1	0	1	0	1	0	0
5	Copycat (1995)	0	0	0	1	1	0	1
6	Shanghai Triad (Yao a yao dao wai po qiao) (1995)	0	0	0	0	1	0	0
8	Babe (1995)	0	0	1	0	1	0	0
9	Dead Man Walking (1995)	0	0	0	0	1	0	0
10	Richard III (1995)	0	0	0	0	1	0	0
11	Seven (Se7en) (1995)	0	0	0	1	0	0	1
12	Usual Suspects, The (1995)	0	0	0	1	0	0	1
13	Mighty Aphrodite (1995)	0	0	1	0	0	0	0
14	Postino, II (1994)	0	0	0	0	1	1	0
15	Mr. Holland's Opus (1995)	0	0	0	0	1	0	0
16	French Twist (Gazon maudit) (1995)	0	0	1	0	0	1	0

6 pav. Filmų duomenų bazės dalis

Siekiant išsiaiškinti kokią įtaką atskiros vartotojų grupės daro galutiniams rezultatams, ir kokią naudą joms suteikia siūlomas modelis, vartotojai buvo išskirstyti į tris kategorijas:

- Pasyvūs – 348,
- Aktyvūs – 239,
- Tarpiniai – 356.

Vartotojų grupės buvo siekiama paskirstyti apytiksliai po lygiai. Pasyviems vartotojams buvo nustatyta 30 filmų įverčio riba, aktyviems vartotojams – 90 filmų įverčio riba. Tai reiškia, kad jeigu vartotojas yra įvertinęs mažiau nei 30 filmų jis yra pasyvus, o jeigu vartotojas yra įvertinęs daugiau nei 90 filmų jis yra aktyvus. Visi likę vartotojai yra tie kurie patenka tarp šių dviejų ribų. Filmai yra vertinami nuo 0 iki 5 balų.

Visi 77639 įverčiai yra padalinti į dvi dalis. Pirmoji dalis įverčių yra skirta sistemos testavimui, kita dalis – sistemos treniravimui. Pirmąją dalį sudaro 15000 įverčių, tačiau jie yra paslėpiami ir sistemai yra nežinomi, juos RS bando atspėti. Tam, kad sistema galėtų įvertinti filmus ir pateikti rekomendacijas, ji turi turėti duomenų failą su jau įvertintais filmais. Taigi duomenų failui yra palikti visi likę 52639 įverčiai, kuriuos sistema žinodama susidaro nuomonę apie kiekvieną vartotoją, turi vartotojų įverčių raidą, pagal kurią yra lyginami vartotojai, ir ieško panašių vartotojų, bei pagal juos pateikia rekomendacijas vartotojams esantiems duomenų testavimo faile.

4.3. Eksperimentų rezultatai

Šiame skyriuje yra pateikiami penkių eksperimentų rezultatai, bei pabaigoje pateikiamas visų eksperimentų vidutinis rezultatas (skyrius 4.3.6). Siekiant išsiaiškinti siūlomo metodo efektyvumą yra nustatytos 5 skirtingos sąlygos:

- 1 sąlyga: vartotojų panašumo slenkstis (VPS) yra 0,2 ir minimalus vienodų įvertintų filmų skaičius (VIFS) – 5 (skyrius 4.3.1).
- 2 sąlyga: VPS yra 0,2 ir VIFS – 10 (skyrius 4.3.2).
- 3 sąlyga: VPS yra 0,2 ir VIFS – 20 (skyrius 4.3.3).
- 4 sąlyga: VPS yra 0,3 ir VIFS – 5 (skyrius 4.3.4).
- 5 sąlyga: VPS yra 0,4 ir VIFS – 5 (skyrius 4.3.5).

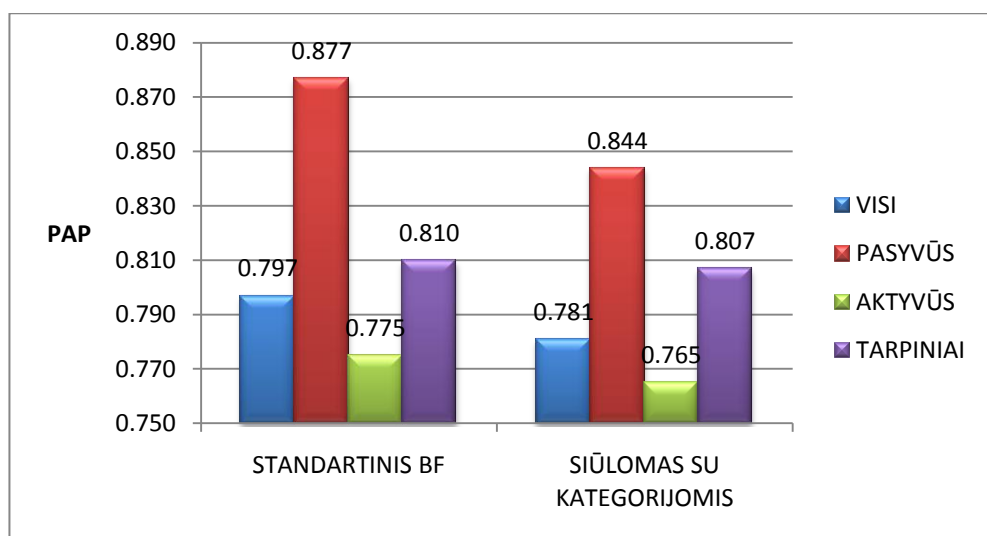
Bandymų rezultatai yra pateikiami dviem būdais: lentele ir grafikais, kuriuose išskirti visi trys vertinimo kriterijai. Lentelėje PAP ir PAVP pagerėjimas/pablogėjimas žymimas žalia/raudona spalva su minuso/pliuso ženklu priekyje, tai reiškia, kad paklaida sumažėjo/padidėjo tam tikru procentu. Įverčių apimties pagerėjimas/pablogėjimas žymimas žalia/raudona spalva su pluso/minuso ženklu. Šiuo atveju minusas reiškia pablogėjimą, nes įverčių apimtis sumažėjo, o tai yra laikoma kaip rezultato suprastėjimas.

4.3.1. Pirmas eksperimentas

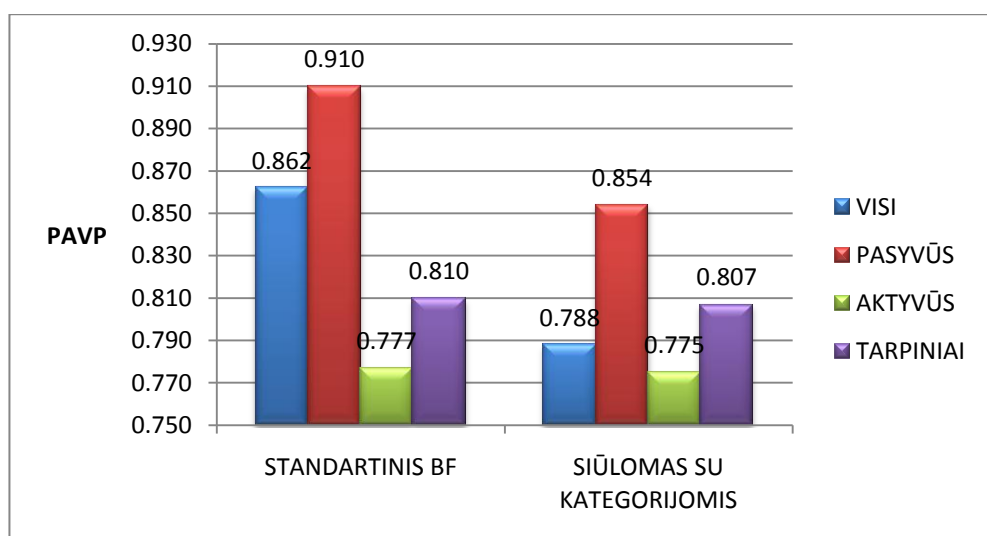
Pirmajame bandyme yra naudojamas žemiausias panašumo slenkstis – 0,2. Jis gana stipriai įtakoja įverčių apimtį, nes sistema didžiąją dalį vartotojų atrenka kaip panašius vartotojus. Dėl šios priežasties ĮA yra beveik 100%. Šiame bandyme siūlomas modelis šiek tiek pablogina ĮA, tačiau visais atvejais sumažina skaičiavimo paklaidą visoms vartotojų grupėms. Iš rezultatų matome, kad didžiausias pagerėjimas yra pasyviems vartotojams. Tačiau nors pasyvių vartotojų beveik daugiausia jiems yra apskaičiuojama mažiausia dalis rekomendacijų su didžiausia paklaida tarp trijų vartotojų grupių. Kaip buvo minėta ankstesniuose skyriuose, tokį rezultatą įtakoja trumpa įverčių raida ir sumažėjęs kiekis panašių vartotojų.

5 lentelė. Pirmo eksperimento rezultatai.

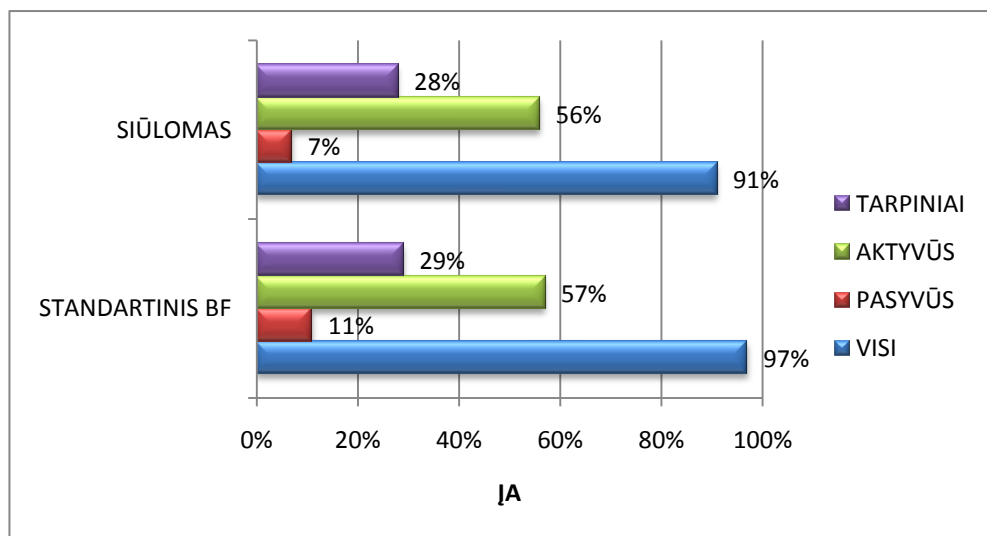
METODAS/ VARTOTOJŲ GRUPĖS			KRITERIJAI/ REZULTATAI			
STANDARTINIS BF	PAP	ĮVERČIŲ DALIS	PAVP	GRUPĖS SVORIS	ĮA	
	Visi	0.797	100 %	0.862	100 %	97.26 %
	Pasyvūs	0.877	11.31 %	0.910	36.90 %	11 %
	Aktyvūs	0.775	58.56 %	0.777	25.34 %	56.95 %
	Tarpiniai	0.810	30.13 %	0.810	37.76 %	29.31 %
	SIŪLOMAS SU KATEGORIJOMIS	PAP	PAP POKYTIS	PAVP	PAVP POKYTIS	ĮA POKYTIS
Visi	0.781	- 2 %	0.788	- 8.6 %	- 6.62 %	
Pasyvūs	0.844	- 3.76 %	0.854	- 6.2 %	- 4.8 %	
Aktyvūs	0.765	- 1.3 %	0.775	- 0.3 %	- 0.3 %	
Tarpiniai	0.807	- 0.4 %	0.807	- 0.4 %	- 1.52 %	



7 pav. Pirmo eksperimenta PAP kriterija rezultāti



8 pav. Pirmo eksperimenta PAVP kriterija rezultāti



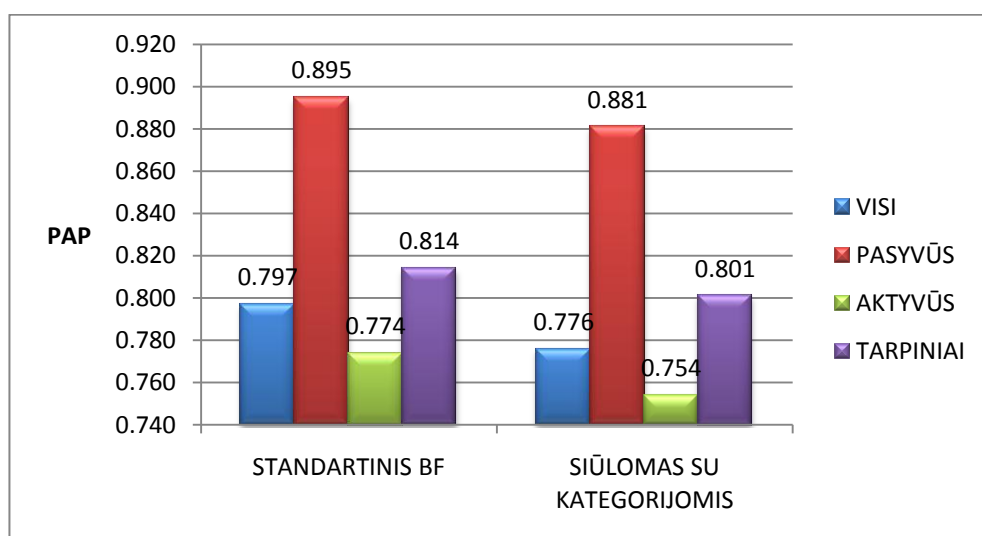
9 pav. Pirmo eksperimenta JA kriterija rezultāti

4.3.2. Antras eksperimentas

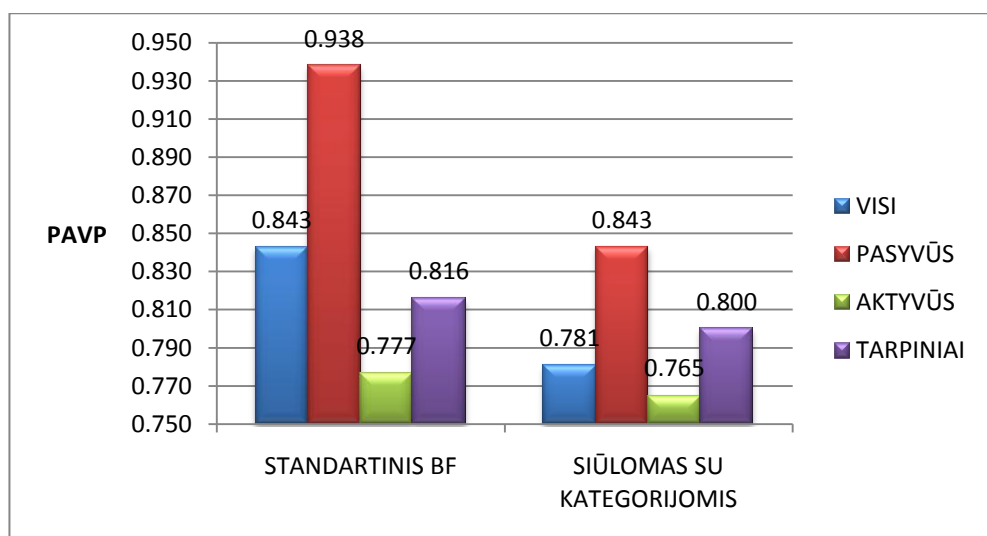
Antrame ir trečiame (skyrius 4.3.3) bandymuose taip pat yra paliekamas tas pats panašumo slenkstis 0,2, tačiau didinamas minimalus vienodų įvertintų filmų slenkstis iki 10 ir 20. Šiais bandymais norima išsiaiškinti ar vienodų įvertintų filmų slenksčio didinimas pagerina tikslumą ir žinoma, ar siūlomas modelis pagerina tikslumą lyginant su standartiniu. Šiame bandyme slenkstis yra nustatytas 10 filmų. Rezultatai rodo, kad tikslumas, nežymiai, bet pagerėja didinant šį slenkstį, o su siūlomu modeliu tikslumas pagerėja labai akivaizdžiai, ypač pasyviems vartotojams, lyginant su standartiniu BF metodu. Tačiau ĮA sumažėja, bet tai yra normalu, nes filmų išskirstymas į kategorijas trumpina kiekvieno vartotojų įverčių raidą, o slenksčio didinimas sumažina panašių vartotojų skaičių.

6 lentelė. Antro eksperimento rezultatai.

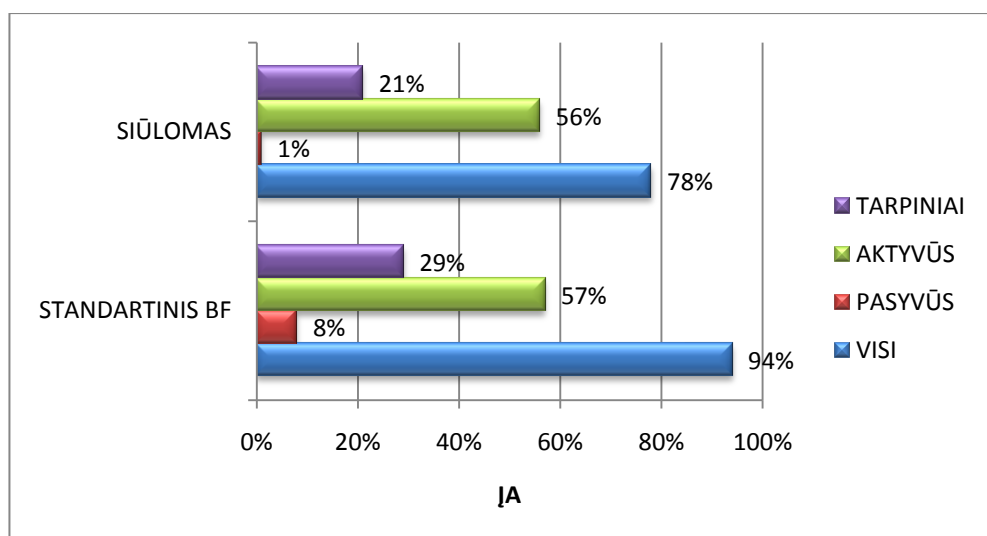
METODAS/ VARTOTOJŲ GRUPĖS	KRITERIJAI/ REZULTATAI				
STANDARTINIS BF	PAP	ĮVERČIŲ DALIS	PAVP	GRUPĖS SVORIS	ĮA
Visi	0.797	100 %	0.843	100 %	94.06 %
Pasyvūs	0.895	8.51 %	0.938	36.90 %	8.01 %
Aktyvūs	0.774	60.53 %	0.777	25.34 %	56.93 %
Tarpiniai	0.814	30.96 %	0.816	37.76 %	29.12 %
SIŪLOMAS SU KATEGORIJOMIS	PAP	PAP POKYTIS	PAVP	PAVP POKYTIS	ĮA POKYTIS
Visi	0.776	- 2.63 %	0.781	- 16.74 %	- 16.19 %
Pasyvūs	0.881	- 1.56 %	0.843	- 10.13 %	- 6.7 %
Aktyvūs	0.754	- 2.58 %	0.765	- 1.54 %	- 1.56 %
Tarpiniai	0.801	- 1.6 %	0.800	- 1.96 %	- 7.93 %



10 pav. Antro eksperimento PAP kriterijaus rezultatai



11 pav. Antro eksperimento PAVP kriterijaus rezultatai



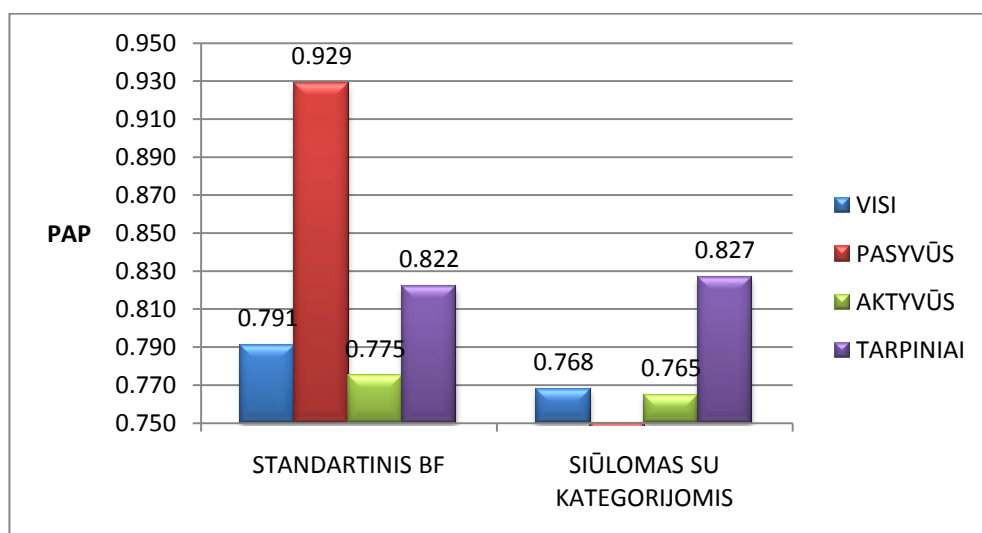
12 pav. Antro eksperimento IA kriterijaus rezultatai

4.3.3. Trečias eksperimentas

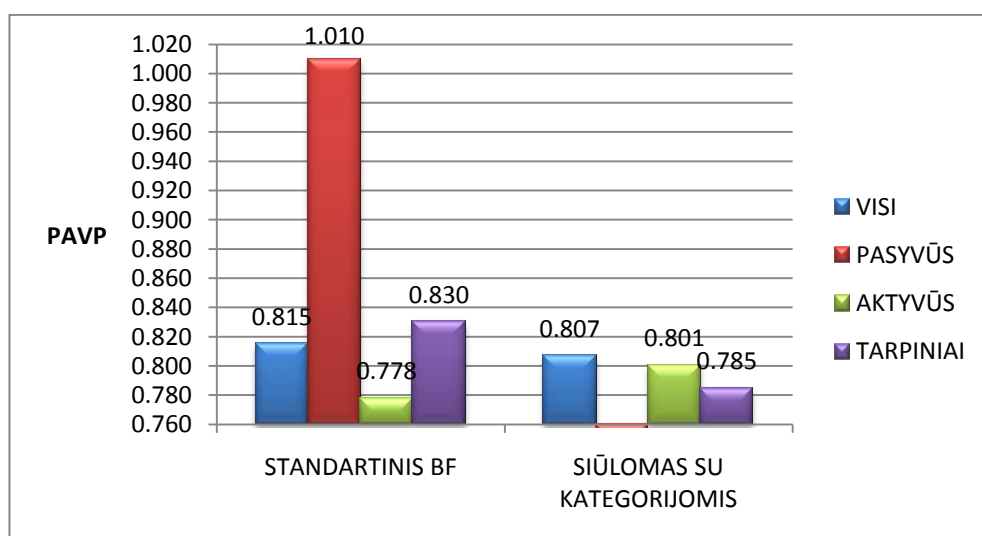
Šiame bandyme kaip ir antrajame yra didinama vienodų įvertintų filmų riba. Ji šiuo atveju yra 20 filmų. Pagal šia naudojamą duomenų bazę tai yra per didelė riba siūlomam modeliui. Rezultatai rodo gana didelį IA sumažėjimą, bei kai kurioms grupėms paklaidos padidėjimą, o pasyviems vartotojams nebuvo apskaičiuota nei viena rekomendacija, tačiau bendras paklaidos sumažėjimas nedidelis, bet išliko. Taigi net ir labai nepalankiomis sąlygomis siūlomas modelis rekomendacijų tikslumą pagerina. Šiais pirmais trimis bandymais išsiaiškinta, kad didinant vienodų įvertintų filmų ribą tikslumas, nežymiai, bet gerėja, tačiau mažėja IA. Taip pat siūlomas modelis visais trimis atvejais pagerino rekomendacijų tikslumą, bet sumažino IA dėl priežasčių paminėtų 4.3.2 skyriuje.

7 lentelė. Trečio eksperimento rezultatai.

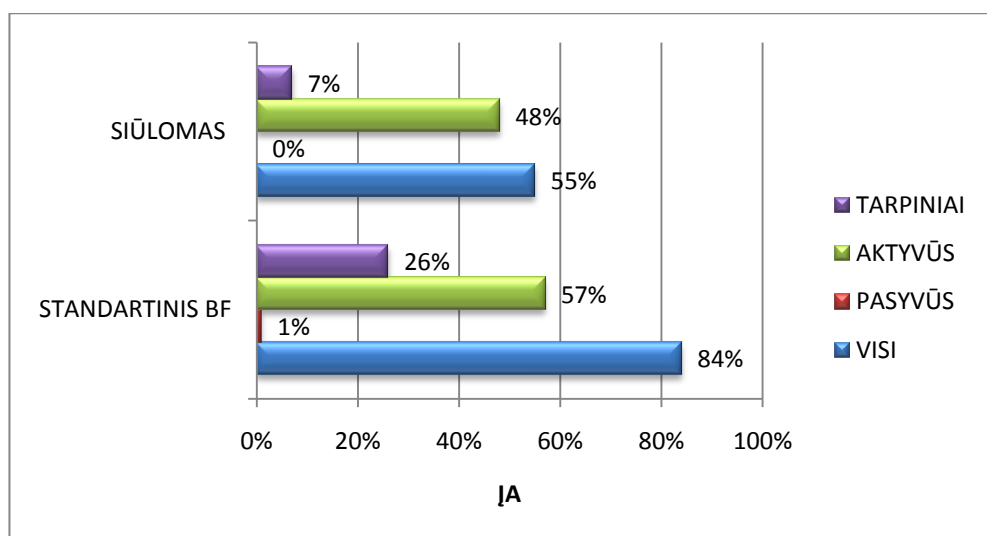
METODAS/ VARTOTOJŲ GRUPĖS	KRITERIJAI/ REZULTATAI				
<i>STANDARTINIS BF</i>	<i>PAP</i>	<i>IVERČIŲ DALIS</i>	<i>PAVP</i>	<i>GRUPĖS SVORIS</i>	<i>IA</i>
Visi	0.791	100 %	0.815	100 %	83.77 %
Pasyvūs	0.929	0.73 %	1.006	36.90 %	0.61 %
Aktyvūs	0.775	67.69 %	0.778	25.34 %	56.70 %
Tarpiniai	0.822	31.58 %	0.830	37.76 %	26.45 %
<i>SIŪLOMAS SU KATEGORIJOMIS</i>	<i>PAP</i>	<i>PAP POKYTIS</i>	<i>PAVP</i>	<i>PAVP POKYTIS</i>	<i>IA POKYTIS</i>
Visi	0.768	- 2.91 %	0.807	- 1 %	- 28.75 %
Pasyvūs	-	0 %	-	0 %	0 %
Aktyvūs	0.765	- 1.3 %	0.801	+ 2.96 %	- 8.98 %
Tarpiniai	0.827	+ 0.6 %	0.785	- 5.42 %	- 19.5 %



13 pav. Trečio eksperimento PAP kriterijaus rezultatai



14 pav. Trečio eksperimento PAVP kriterijaus rezultatai



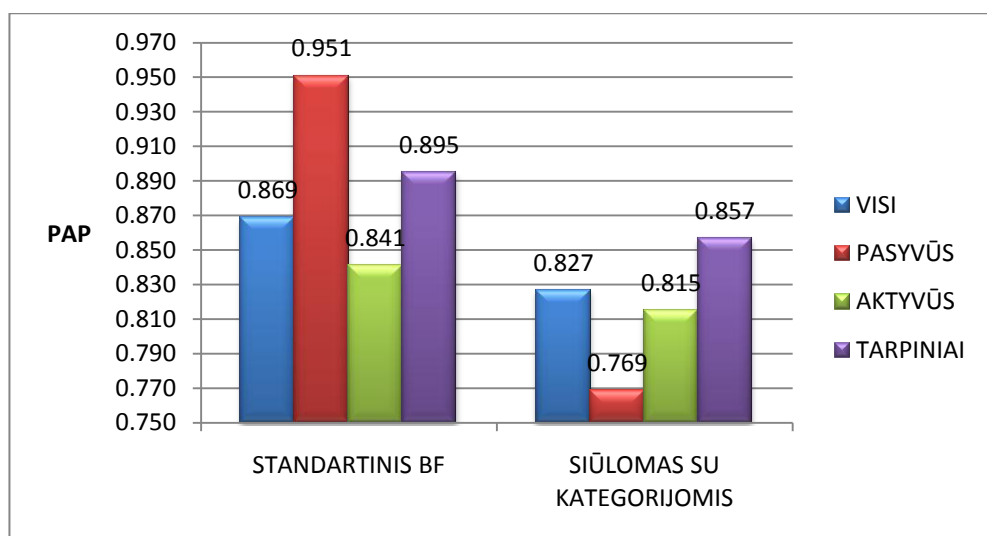
15 pav. Trečio eksperimento JA kriterijaus rezultatai

4.3.4. Ketvirtas eksperimentas

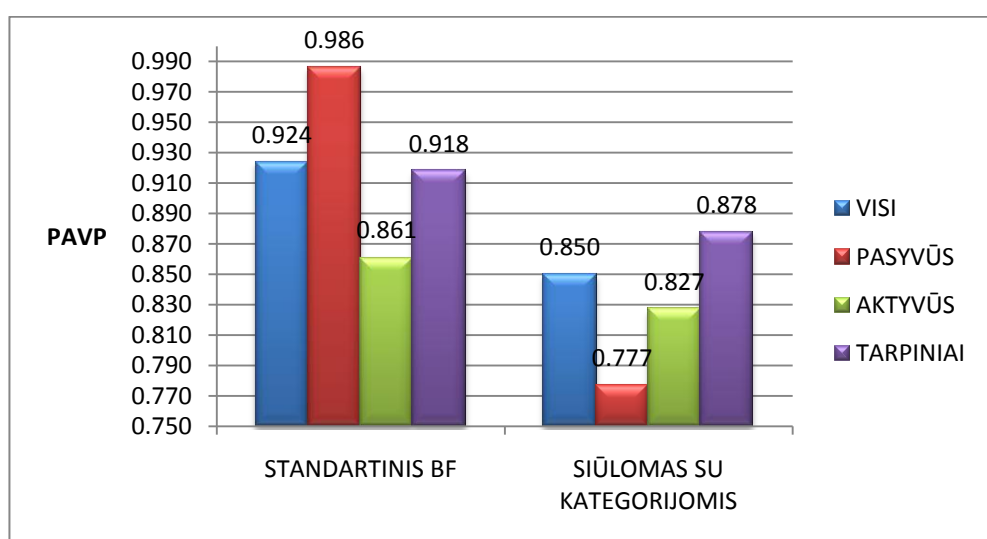
Ketvirtuoju ir penktuoju bandymais siekiama išsiaiškinti ar didinant panašumo ribą tarp vartotojų pagerėja/pablogėja rekomendacijų tikslumas ir kaip keičiasi JA. Šiuo atveju yra taikomas 0,3 panašumo slenkstis, tai reiškia, kad vartotojai, kurių panašumo rodiklis yra 0,3 ar mažesnis atkrenta ir nėra naudojami rekomendacijų skaičiavimui. Rezultatai rodo, kad siūlomas modelis visais atvejais pagerino rekomendacijų tikslumą gana ženkliai, bei padidino bendrą įverčių apimtį. Tačiau žiūrint bendrai į JA, ji stipriai yra pamažėjus lyginant su JA rezultatais gautais kai panašumo slenkstis buvo 0,2. Galima daryti išvadą, kad šiai duomenų bazei panašumo slenksčio didinimas blogina rezultatus. Tai atsitiko dėl to, kad šioje duomenų bazėje vartotojai įvertinę ganėtinai nedaug filmų, todėl šios ribos didinimas sumažina panašių vartotojų skaičių. Tačiau ši sąlyga yra taikoma ir kitoms duomenų bazėms nesvarbu ar ji didesnė ar mažesnė, o slenksčio parinkimas yra individualus kiekvienam duomenų rinkiniui.

8 lentelė. Ketvirto eksperimento rezultatai.

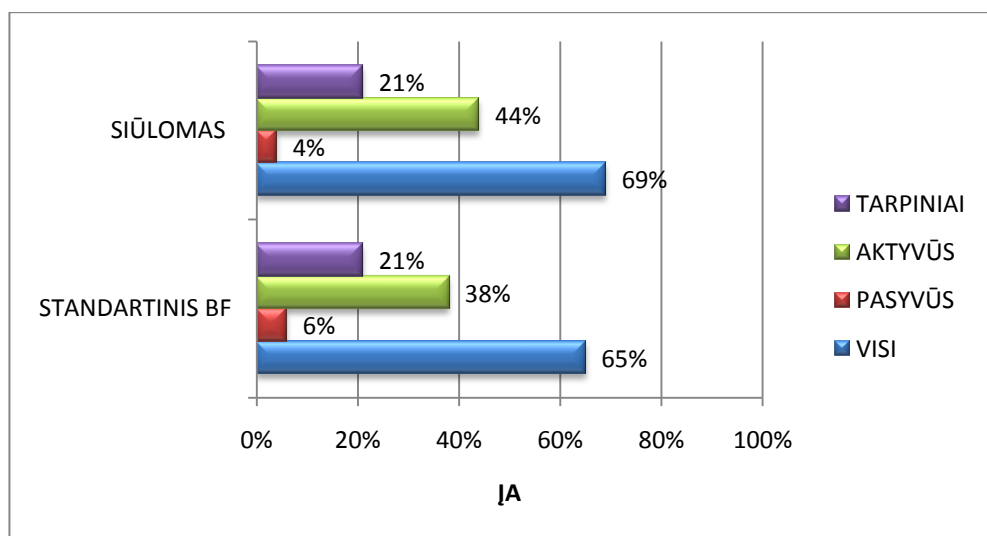
METODAS/ VARTOTOJŲ GRUPĖS	KRITERIJAI/ REZULTATAI				
STANDARTINIS BF	PAP	ĮVERČIŲ DALIS	PAVP	GRUPĖS SVORIS	JA
Visi	0.869	100 %	0.924	100 %	64.95 %
Pasyvūs	0.951	9.69 %	0.986	36.90 %	6.29 %
Aktyvūs	0.841	58.65 %	0.861	25.34 %	38.09 %
Tarpiniai	0.895	31.66 %	0.918	37.76 %	20.57 %
SIŪLOMAS SU KATEGORIJOMIS	PAP	PAP POKYTIS	PAVP	PAVP POKYTIS	JA POKYTIS
Visi	0.827	- 4.8 %	0.850	- 8 %	+3.9 %
Pasyvūs	0.769	- 19.1 %	0.777	- 21.2 %	-2.67 %
Aktyvūs	0.815	- 3.1 %	0.827	- 3.95 %	+6.52 %
Tarpiniai	0.857	- 4.25 %	0.878	- 4.36 %	+0.05 %



16 pav. Ketvirto eksperimento PAP kriterijaus rezultatai



17 pav. Ketvirto eksperimento PAVP kriterijaus rezultatai



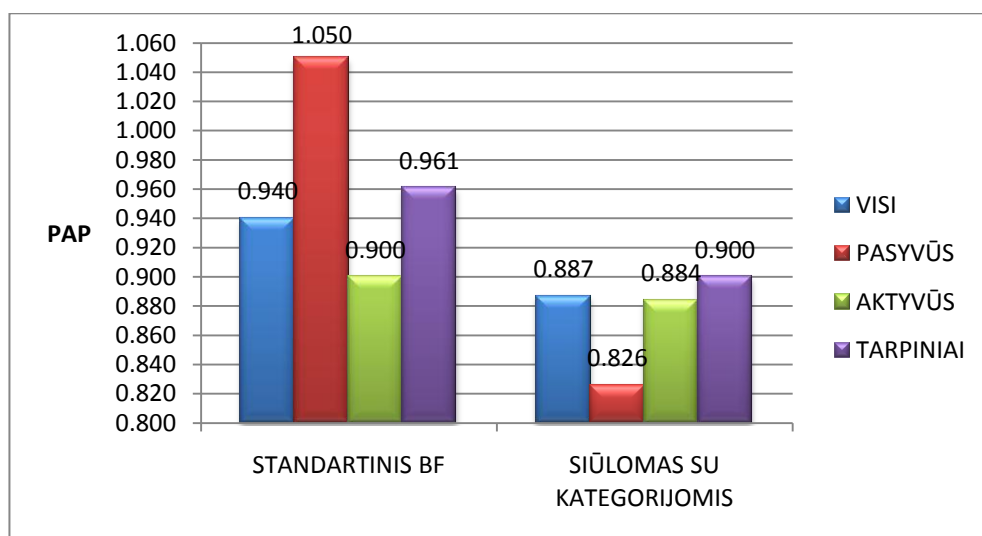
18 pav. Ketvirto eksperimento JA kriterijaus rezultatai

4.3.5. Penktas eksperimentas

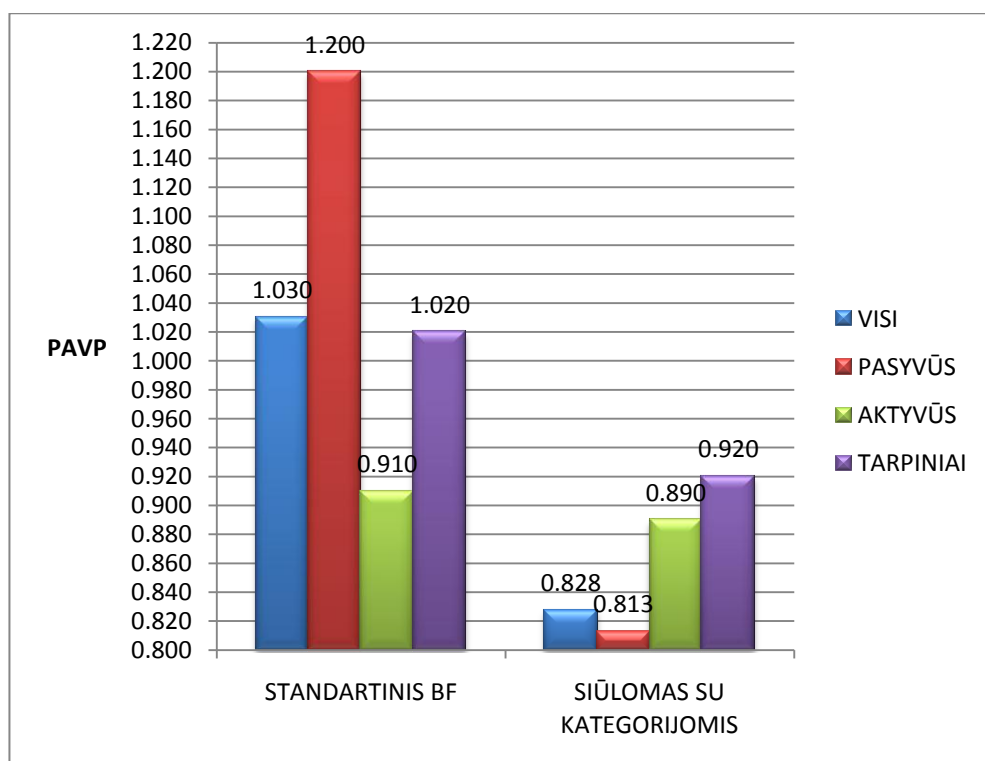
Paskutiniuoju eksperimentu buvo norima sudaryti nepalankias sąlygas abiem metodams ir pažiūrėti kaip siūlomas metodas pagerina/pablogina rekomendacijų tikslumą, bei kaip pakeičia I_A. Iš rezultatų galima matyti, kad siūlomas metodas akivaizdžiai pagerino bendrą rekomendacijų tikslumą, bei tikslumą atskiroms vartotojų grupėms. Taip pat labai pagerino I_A, kuris standartiniame metode sumažėjo net iki 17%. Siūlomame modelyje I_A pagėrėjimą įtakoja tai, kad vienas filmas gali priklausyti kelioms skirtingoms kategorijoms, todėl vartotojas įvertinęs filmą, kuris priklauso skirtingoms kategorijoms, gali būti atrinktas kelis kartus kaip panašus vartotojas, priklausomai nuo to kelioms kategorijoms priklauso jo įvertintas filmas.

9 lentelė. Penkto eksperimento rezultatai.

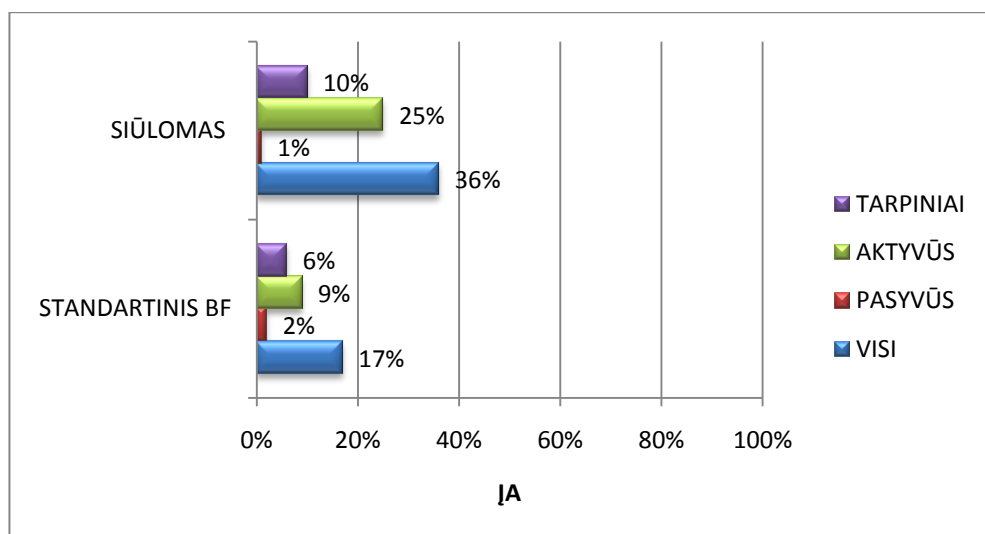
METODAS/ VARTOTOJŲ GRUPĖS	KRITERIJAI/ REZULTATAI				
<i>STANDARTINIS BF</i>	<i>PAP</i>	<i>IVERČIŲ DALIS</i>	<i>PAVP</i>	<i>GRUPĖS SVORIS</i>	<i>I_A</i>
Visi	0.940	100 %	1.027	100 %	16.73 %
Pasyvūs	1.046	12.79 %	1.196	36.90 %	2.14 %
Aktyvūs	0.900	51.75 %	0.910	25.34 %	8.66 %
Tarpiniai	0.961	35.46 %	1.019	37.76 %	5.93 %
<i>SIŪLOMAS SU KATEGORIJOMIS</i>	<i>PAP</i>	<i>PAP POKYTIS</i>	<i>PAVP</i>	<i>PAVP POKYTIS</i>	<i>I_A POKYTIS</i>
Visi	0.887	- 5.64 %	0.828	- 19.4 %	+18.87%
Pasyvūs	0.826	- 21.03 %	0.813	- 32.02 %	-1.1 %
Aktyvūs	0.884	- 1.77 %	0.890	- 2.2 %	+15.86 %
Tarpiniai	0.900	- 6.34 %	0.920	- 9.7 %	+4.17 %



19 pav. Penkto eksperimento PAP kriterijaus rezultatai



20 pav. Penkto eksperimento PAVP kriterijaus rezultatai



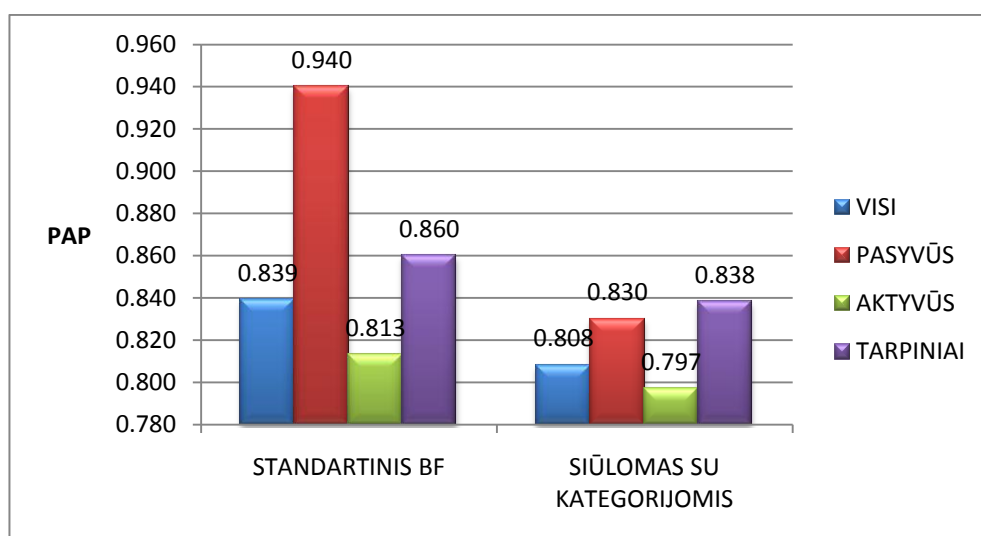
21 pav. Penkto eksperimento ĶA kriterijaus rezultatai

4.3.6. Visų eksperimentų apibendrinimas

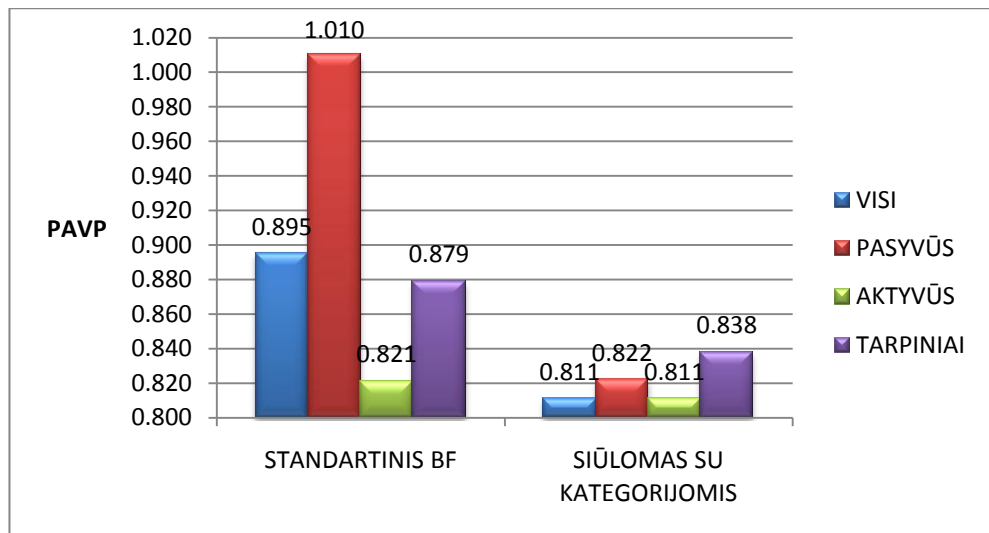
Šiame skyriuje pateikiami galutiniai apibendrinti rezultatai iš visų bandymų. Iškelta prielaida, kad rekomendacijų skaičiavimas suskirsčius produktus į kategorijas sumažina skaičiavimo paklaidas, pasitvirtino, nes siūlomas modelis pagerino rekomendacijų tikslumą visų vartotojų atžvilgiu ir kiekvienos vartotojų grupės atskirai atžvilgiu. Ypač svarbu pabrėžti, kad geriausi rezultatai yra gauti labiausiai problemiška vartotojų grupei – pasyviems vartotojams. Jiems tikslumas pagėrėjo net 11 ir 17 procentų. Taip pat nereikia pamiršti IĀ, kuri pablogėjo 5% vertinant visų vartotojų atžvilgiu, tačiau aktyviems vartotojams ji pagėrėjo 3%. Taip pat reikia pastebėti, kad atmetus trečio bandymo, kuris buvo atliktas taikant per didelius slenksčius testuojamai duomenų bazei, rezultatus, kurie įtakoją labai stipriai IĀ sumažėjimą, bendras visų bandymų IĀ pokytis nebūtų neigiamas.

10 lentelė. Visų eksperimentų rezultatų vidurkis.

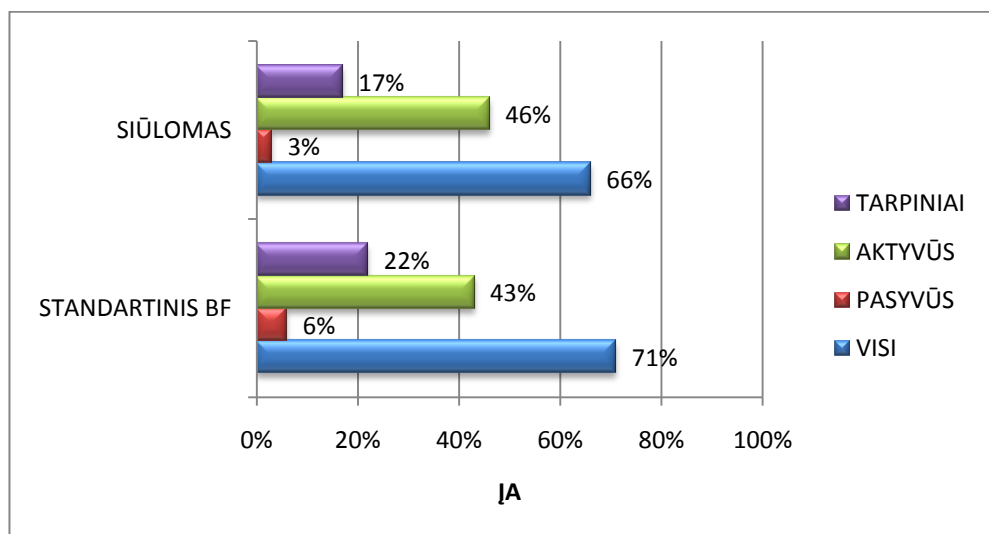
METODAS/ VARTOTOJŲ GRUPĖS	KRITERIJAI/ REZULTATAI				
<i>STANDARTINIS BF</i>	<i>PAP</i>	<i>IĖRČIŲ DALIS</i>	<i>PAVP</i>	<i>GRUPĖS SVORIS</i>	<i>IĀ</i>
Visi	0.839	100%	0.895	100%	71%
Pasyvūs	0.940	9%	1.009	37%	6%
Aktyvūs	0.813	59%	0.821	25%	43%
Tarpiniai	0.860	32%	0.879	38%	22%
<i>SIŪLOMAS SU KATEGORIJOMIS</i>	<i>PAP</i>	<i>PAP POKYTIS</i>	<i>PAVP</i>	<i>PAVP POKYTIS</i>	<i>IĀ POKYTIS</i>
Visi	0.808	-4%	0.811	-11%	-5%
Pasyvūs	0.830	-11%	0.822	-17%	-3%
Aktyvūs	0.797	-2%	0.811	-2%	+3%
Tarpiniai	0.838	-3%	0.838	-4%	-5%



22 pav. Visų eksperimentų PAP kriterijaus vidurkis



23 pav. Visų eksperimentų PAVP kriterijaus vidurkis



24 pav. Visų eksperimentų IA kriterijaus vidurkis

Šiame tiriamajame darbe iškeltos prielaidos ir tikslai buvo įgyvendinti. Tai patvirtino bendro filtravimo modelio paremti panašiais vartotojais ir siūlomo modelio gautų rezultatų lyginamoji analizė. Galutinės viso darbo išvados yra pateikiamos 5 skyriuje.

5. IŠVADOS IR PASIŪLYMAI

5.1. IŠVADOS

1. Pagrindinis tikslas, kuriuo buvo siekiama sumažinti rekomendacijų įverčių nuokrypį, buvo įgyvendintas pritaikius skaičiavimams siūlomąjį modelį, kuris sujungia dviejų rekomendacinių sistemų rekomendavimo metodų atskirus ypatumus, skaičiavimo tikslumo nuokrypio atžvilgiu, visuose bandymuose buvo geresnis už standartinį bendro filtravimo metodą.
2. Vienas iš tikslų buvo pagerinti rekomendacijų tikslumą pasyviems vartotojams. Šis tikslas buvo pasiektas taikant siūlomą metodą su kategorijomis. Nors tarp visų vartotojų grupių pasyviems vartotojams nuokrypis liko didžiausias, tačiau jis sumažėjo labiausiai. Atskiriems nuokrypio vertinimo kriterijams rezultatas pagerėjo vidutiniškai 11% ir 17%.
3. Bendrai vertinant rezultatus, kurie buvo gauti apskaičiavus vidutinius visų penkių bandymų rezultatus, bendras ĮA rezultatas pablogėjo 4%. Tačiau vertinant atskirus bandymus, esant nepalankioms skaičiavimo sąlygoms, siūlomas modelis ĮA dviejuose bandymuose iš penkių pagerino atitinkamai 4% ir 18%.
4. Vertinant ĮA atskiroms vartotojų grupėms, tai pasyviems vartotojams visuose bandymuose daugiau ar mažiau ĮA sumažėjo, tačiau tai yra natūralu, nes siūlomas modelis atlieka atskirų kategorijų skaičiavimus, taip sutrumpindamas visų vartotojų įverčių raidas. Kadangi pasyvūs vartotojai turi trumpiausią įverčių raidą, tai jiems daugiausia sumažėja ĮA. Aktyviems ir vidutiniams vartotojams dėl jų žymiai ilgesnės įverčių raidos ĮA sumažėja mažiau, bei esant tam tikroms sąlygoms netgi padidėja, palyginti su bendro filtravimo metodu.
5. ĮA padidėjimas su siūlomu modeliu atsiranda dėl to, kad nemažai filmų priklauso kelioms kategorijoms. Dėl to, vartotojas įvertinęs tokį filmą, turi daugiau šansų būti atrinktas kaip panašus vartotojas daugiau kartų, priklausomai nuo to, kelioms kategorijoms priklauso filmas.
6. Nepasitikėjimo rekomendacine sistema problemos sumažinimui baigiamajame darbe buvo aprašyta aiški rekomendacija, kuri suteikia vartotojui daugiau informacijos apie rekomenduotoją ir rekomenduojamą produktą, bei patektas tokios rekomendacijos pavyzdys.
7. Vartotojų panašumo slenksčio (toliau – VPS) naudojimas yra labai individualus kiekvienam duomenų rinkiniui. Eksperimentuose naudojamai duomenų bazei geriausias VPS buvo 0,2.

8. VPS nustatymas leidžia išvengti visiškai nepanašių vartotojų, kurie lemia didesnę rekomendacijų skaičiavimo nuokrypį, tačiau per didelis VPS gali labai stipriai sumažinti įverčių apimtį (toliau – $\mathcal{I}\mathcal{A}$), nes bus surasta mažai panašių vartotojų.
9. Vienodų įvertintų filmų slenkstis (toliau – $\mathcal{V}\mathcal{I}\mathcal{F}\mathcal{S}$) taip pat kaip ir VPS sumažina rekomendacijų skaičiavimo nuokrypį, tačiau per didelis $\mathcal{V}\mathcal{I}\mathcal{F}\mathcal{S}$ labai sumažina $\mathcal{I}\mathcal{A}$, nes iš skaičiavimo proceso iškrenta labai daug ypač pasyvių, vartotojų, kurių įverčių raida yra labai trumpa.
10. Eksperimentams naudojamai duomenų bazei optimaliausias $\mathcal{V}\mathcal{I}\mathcal{F}\mathcal{S}$ yra 10 filmų.
11. Taikant šiai duomenų bazei standartinį bendro filtravimo metodą, VPS turėtų būti 0,2 ir $\mathcal{V}\mathcal{I}\mathcal{F}\mathcal{S}$ turėtų būti 10. Su šiais parametrais atliktame antrame bandyme buvo pasiektas geriausias santykis tarp skaičiavimo nuokrypio ir $\mathcal{I}\mathcal{A}$. Nors trečiame bandyme, kai $\mathcal{V}\mathcal{I}\mathcal{F}\mathcal{S}$ buvo 20, skaičiavimo nuokrypis dar pamažėjo keliais procentais, tačiau beveik 10 % sumažėjo $\mathcal{I}\mathcal{A}$, o pasyviems vartotojams buvo apskaičiuota tik pusė procento rekomendacijų.

5.2. PASIŪLYMAI

1. Šios srities tiriamasis darbas Lietuvoje yra vienas iš pirmųjų, todėl jis gali būti naudojamas kaip pagrindas ar literatūros šaltinis tolesniems tyrimams.
2. Šiame darbe aprašytas siūlomas modelis gali būti taikomas realiai rekomendacinei sistemai.
3. Šiame darbe į kategorijas buvo skirstomi filmai. Tolesniems tyrimams, būtų galima naudoti duomenų bazę, kurioje būtų įvairūs produktai, pavyzdžiui knygos, muzika, automobiliai ir pan. Tačiau laisvai prieinamos tokios realios duomenų bazės šiuo metu nėra, dėl šios priežasties buvo pasirinkta MovieLens sukurta, laisvai prieinama duomenų bazė.
4. Minėtos duomenų bazės sukūrimui būtų galima darbe aprašytam modeliui sukurti taikomąją programą, kuri būtų pritaikoma socialiniam tinklui, kuris apima minėtus produktus. Per tam tikrą laikotarpį būtų surinkti reikalingi tyrimams duomenys.
5. Tolesniems tyrimams galima būtų naudoti žymiai didesnius duomenų rinkinius, kurie apima 1000000 arba daugiau įverčių. Tačiau su tokiu duomenų rinkiniu yra labai sudėtinga dirbti, nes testai atliekami labai ilgai, o gautų rezultatų analizė reikalautų didesnių kompiuterio ir laiko išteklių.

LITERATŪROS SĄRAŠAS

1. **Adomavičius G. ir Kwon Y.Ok** New Recommendation Techniques for Multicriteria Rating Systems [Konferencija]. - Minnesota, JAV : IEEE Intelligent Systems, 2007 m., Tom. 22, nr. 3, p. 48-55.
2. **Adomavičius G. ir Tuzhilin A.** Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [Žurnalas] // IEEE Trans. Knowledge and Data Engineering, 2005 m., p. 734-749.
3. **Balabanovic M. ir Shoham Y.** Fab: Content-Based, Collaborative Recommendation [Žurnalas] // Comm. ACM, 1997 m., Tom. 40, p. 66-72.
4. **Belkin N.J. ir Croft W.B.** Information Filtering and Information Retrieval: Two Sides of the Same Coin? [Konferencija] // Communications of the ACM, 1992 m., p. 29-38.
5. **Bonhard P. [et al.]** Accounting for Taste: Using Profile Similarity to Improve Recommender Systems" [Konferencija] // Conference on Human Factors in Computing Systems. Montreal, Quebec, Kanada : ACM, 2006, p. 1057-1066.
6. **Bonhard P. ir Sasse M. A.** Knowing me, knowing you - using profiles and social networking to improve recommender systems [Žurnalas]. - Hingham, JAV : Kluwer Academic Publishers, 2006 m., Tom. 24., p. 84-98.
7. **Bouyd D. M. ir Ellison N. B.** Social network sites: definitions, history and scholarship [Žurnalas]. - Mičiganas, JAV : Journal of Computer - Mediated Communication, 2007 m., Tom. 13, p. 1-12.
8. **Breese J.S., Heckerman D. and Kadie C.** Empirical Analysis of Predictive Algorithms for Collaborative Filtering [Conference] // Proc. 14th Conf. Uncertainty in Artificial Intelligence, 1998 m., p. 43-52.
9. **Burke R.D., Hammond K.J. and Young B.C.** The FindMe Approach to Assisted Browsing [Conference] // IEEE Expert, 1997 m., p. 32-40.
10. **Burke Robin** Hybrid Recommender Systems: Survey and Experiments [Žurnalas]. - Kalifornija, JAV : Springer Netherlands, 2004 m., Tom. 12, p. 331-370.
11. **Canny J.F.** Collaborative Filtering with Privacy via Factor Analysis [Konferencija] // Proc. of SIGIR 2002, Tampere, Finland, 2002 m., p. 238-245.
12. **Das A. [et al.]** Google News Personalization: Scalable Online Collaborative Filtering [Konferencija] // ACM WWW, Kanada, 2007 m., p. 271-280.
13. **Ding S. [et al.]** Boosting collaborative filtering based on statistical prediction errors [Konferencija] // Proceedings of ACM Conference on Recommender Systems, Luiziana, Šveicarija : ACM, 2008 m., p. 3-10.

14. **Herlocker J.L. [et al.]** An algorithmic framework for performing collaborative filtering [Konferencija] // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999 m., p. 230 - 237 .
15. **Huang Shiu-li** Comparision of Utility-Based Recommendation Methods [Konferencija] // PACIS 2008 Proceedings. - Taiwan, 2008 m., p. 1-12.
16. **Linden G., Smith B. ir York J.** Amazon.com Recommendations: Item-to-Item Collaborative Filtering [Žurnalas]. IEEE Internet Computing. 2003 m., Tom. 7, p. 76-80.
17. **Massa P. ir Avesani P.** Trust-Aware Collaborative Filtering for Recommender Systems [Knygos skyrius] // On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE. Springer Berlin / Heidelberg, 2004 m., Tom. 3290, p. 492-508.
18. **McLaughlin M.R. ir Herlocker J.L.** A collaborative filtering algorithm and evaluation metric that accurately model the user experience [Konferencija] // Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004 m., p. 329 - 336.
19. **Palau J. [et al.]** Collaboration analysis in recommender systems using social networks [Knygos skyrius] // Cooperative Information Agents VIII. - Vokietija : Springer, 2004. - Tom. 3191, p. 137-151.
20. **Papagelis M. ir Plexousakis D.** Qualitative Analysis of User-Based and Item-Based Prediction Algorithms for Recommendation Agents [Knygos skyrius] // Cooperative Information Agents VIII. Springer Berlin / Heidelberg, 2004 m., Tom. 3191, p. 152-166.
21. Prieiga per internetą, GroupLens [Tinkle] // <http://www.grouplens.org/>. - [Žiūrėta 2009 m. 11]. - <http://www.grouplens.org/node/73>.
22. Prieiga per internetą, Jason Weber [Tinkle] // <http://www.jweb.us>. - [Žiūrėta 2009 m. 11]. - http://jweb.us/source_code/?p=239.
23. Prieiga per internetą, Wikipedia [Tinkle] // <http://www.wikipedia.org>. - Wikimedia Foundation, Inc.. - [Žiūrėta 2009 m. 10]. - http://en.wikipedia.org/wiki/Social_web.
24. **Rashid A. M. [et al.]** ClustKNN: A highly scalable hybrid model & memory based algorithm [Konferencija]. - Philadelphia, Pennsylvania, USA : WEBKDD '06, 2006 m., p. 1-10.
25. **Resnick P. [et al.]** GroupLens: An Open Architecture for Collaborative Filtering of Netnews [Konferencija] // Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work. - Chapel Hill, 1994 m., p. 175-186.
26. **Sarwar B.M. [et al.]** Item-based collaborative filtering recommendation algorithms [Konferencija] // Proc. of the Tenth International World Conference. 2001 m., p. 285-295.

27. **Shardanand U. ir Maes P.** Social Information Filtering: Algorithms for Automating “Word of Mouth” [Konferencija] // In Proceedings of ACM 95 Conference on Human Factors in Computing Systems. - Denver, 1995 m., p. 210-217.
28. **Shi Y., Larson M. ir Hanjalic A.** Exploiting User Similarity based on Rated-Item Pools for Improved User-Based Collaborative Filtering [Konferencija] // ACM Conference On Recommender Systems. - Njujorkas, JAV : ACM, 2009 m., p. 125-132.
29. **Vozalis E. ir Margarits K. G.** Analysis of recommender systems algorithms [Konferencija] // HERCMA. - Aténai, Graikija : University of Macedonia, 2003 m., p. 1-14.
30. **Zanker M. [et al.]** Comparing Recommendation Strategies in a Commercial Context [Žurnalas]. IEEE Intelligent Systems, 2007 m., Tom. 22, p. 69-73.
31. **Zhang J. ir Pu P.** A recursive prediction algorithm for collaborative filtering recommender systems [Konferencija] // ACM conference on Recommender Systems. - Mineapolis, JAV : ACM, 2007 m., p. 57-64.