UNIVERZITET U NOVOM SADU
EKONOMSKI FAKULTET U SUBOTICI

UNAPREĐENJE POSTUPAKA ZA OTKRIVANJE ASOCIJATIVNIH PRAVILA O KORIŠĆENJU WEB SAJTOVA

DOKTORSKA DISERTACIJA

Mentor: Prof. dr Zita Bošnjak

Kandidat: mr Maja Dimitrijević

Subotica, 2016. godine

UNIVERZITET U NOVOM SADU
EKONOMSKI FAKULTET U SUBOTICI

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj: RBR	
Identifikacioni broj: IBR	
Tip dokumentacije: TD	Monografska dokumentacija
Tip zapisa: TZ	Tekstualni štampani materijal
Vrsta rada (dipl., mag., dokt.): VR	Doktorska disertacija
Ime i prezime autora: AU	Maja Dimitrijević
Mentor (titula, ime, prezime, zvanje): MN	dr Zita Bošnjak, redovni profesor
Naslov rada: NR	Unapređenje postupaka za otkrivanje asocijativnih pravila o korišćenju web sajtova
Jezik publikacije: JP	Srpski
Jezik izvoda: JI	srp. / eng.
Zemlja publikovanja: ZP	Republika Srbija
Uže geografsko područje: UGP	Autonomna Pokrajina Vojvodina
Godina: GO	2016.
Izdavač: IZ	autorski reprint
Mesto i adresa: MA	Segedinski put 9-11, 24000 Subotica
Fizički opis rada: FO	8 poglavlja / 148 stranica / 17 slika / 23 tabele / 14 grafikona / 142 reference
Naučna oblast: NO	Informacioni sistemi
Naučna disciplina: ND	Algoritmi i mere interesantnosti asocijativnih pravila

Predmetna odrednica, ključne reči: PO	Asocijativna pravila, mere interesantnosti, otkrivanje znanja o korišćenju web sajtova
UDK	
Čuva se: ČU	Biblioteka Ekonomskog fakulteta u Subotici, Segedinski put 9-11, 24000 Subotica
Važna napomena, VN:	
Izvod: IZ	Algoritmi za otkrivanje asocijativnih pravila u web log podacima imaju tendenciju generisanja prevelikog broj pravila u kojima je analitičarima podataka teško snaći se pri odabiru stvarno korisnih pravila. U okviru ove disertacije predložena je metoda za eliminaciju neinteresantnih, odnosno asocijativnih pravila statistički očekivanih u odnosu na opštija asocijativna pravila. Definicija uslova za eliminaciju bazira se na statističkoj Z-score meri, definisanoj lokalno, na skupu transakcija koje sadrže opštiji skup atributa. Predložena je modifikacija standardnih matematičkih mera interesantnosti, kojom se povećava kvalitet rangiranja preostalih asocijativnih pravila. Implementiran je softverski sistem koji obuhvata sve faze procesa otkrivanja asocijativnih pravila o korišćenju web sajtova. Eksperimentalno je ispitan učinak smanjenja veličine skupa otkrivenih asocijativnih pravila, kao i rangiranje preostalih asocijativnih pravila prema različitim merama interesantnosti, na dva stvarna skupa podataka o korišćenju web sajtova.
Datum prihvatanja teme od strane Senata, DP:	27.01.2012.
Datum odbrane: DO	
Članovi komisije: (ime i prezime / titula / zvanje / naziv organizacije / status) KO	predsednik: _____ član: _____ član: _____ član: _____ član: _____

UNIVERSITY OF NOVI SAD
FACULTY OF ECONOMICS SUBOTICA

KEY WORD DOCUMENTATION

Accession number: ANO	
Identification number: INO	
Document type: DT	Monograph documentation
Type of record: TR	Textual printed material
Contents code: CC	PhD thesis
Author: AU	Maja Dimitrijević
Mentor: MN	Zita Bošnjak, PhD Full Profesor
Title: TI	Improving web usage association rule discovery methods
Language of text: LT	Serbian
Language of abstract: LA	English
Country of publication: CP	Serbia
Locality of publication: LP	Vojvodina
Publication year: PY	2016.
Publisher: PU	Author's reprint
Publication place: PP	9-11 Segedinski put Street, 24000 Subotica
Physical description: PD	8 chapters / 148 pages / 17 figures / 23 tables / 14 charts / 142 references
Scientific field SF	Information systems
Scientific discipline SD	Association rule discovery algorithms and interestingness measures

Subject, Key words SKW	Association rules, interestingness measures, knowledge discovery in web usage data
UC	
Holding data: HD	Faculty Library, 9-11 Segedinski put Street, 24000 Subotica
Note: N	
Abstract: AB	Association rule mining algorithms applied to web usage data tend to generate huge numbers of rules, which makes it difficult for the data analysts to select truly useful rules. A method for elimination of uninteresting association rules, which are statistically expected with respect to more general association rules is proposed. The condition for elimination is based on the statistical Z-score measure, defined locally, in the set of transactions that contain the more general attribute set. A modification of the standard statistical interestingness measures is proposed, aiming to enhance the quality of the remaining association rule ranking. A software system encompassing all phases of web usage association rule discovery process is implemented. The experiments are conducted on two real life web usage data sets, showing the high degree of pruning of uninteresting association rules based on the proposed elimination methods, as well as the improvement of the association rule ranking based on the modified interestingness measures.
Accepted on Senate on: AS	27.01.2012.
Defended: DE	
Thesis Defend Board: DB	president: _____ member: _____ member: _____ member: _____ member: _____

Zahvaljujem se mentoru prof. dr Ziti Bošnjak na savetima i nesebičnoj podršci tokom izrade ove doktorske disertacije; mom suprugu i deci na strpljenju i ljubavi tokom ovog našeg zajedničkog poduhvata; i svima koji su direktno ili indirektno, doprineli ovom istraživanju.

SADRŽAJ

Uvod

Zahvaljujući ogromnom protoku podataka preko hiper-tekst transfer protokola tokom poslednjih godina na web serverima se konstantno generišu velike količine podataka o korišćenju web sajtova. Ovi podaci se najčešće nalaze u formi tekstualnih web server log datoteka, i sadrže informacije o tome koje stranice određenog web sajta su korisnici posetili u kom trenutku (Facca & Lanzi, 2005; Berendt, Hollink, Luczak-Rösch, Möller & Vallet, 2011). Metode koje pripadaju popularnoj istraživačkoj oblasti *„otkrivanje znanja u bazama podataka“* primenjuju se i u cilju otkrivanja „prethodno nepoznatih, potencijalno korisnih paterna“ u ovim ogromnim repozitorijumima podataka o korišćenju web sajtova (Facca & Lanzi, 2005; Wu & Kumar, 2009).

Analitički deo procesa otkrivanja znanja u bazama podataka razvijen je u posebnu istraživačku oblast pod nazivom *„rudarenje podataka“ („data mining“)*. Jedna od najpopularnijih data mining metoda je otkrivanje asocijativnih pravila u podacima (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996; Hand, Mannila & Smyth, 2001; Wu & Kumar, 2009). Primena metoda iz data mining oblasti radi analize podataka koji postoje na *World Wide Web* mreži tokom poslednjih godina razvila se u posebnu oblast istraživanja nazvanu *„web mining“* (Berendt, 2004; Etzioni, 1996; Singh & Singh, 2010). Primena asocijativnih pravila u cilju otkrivanja znanja o korišćenju web sajtova je jedna od popularnih metoda koja pripada *web mining* istraživačkoj oblasti (Liu, 2007; Srivastava, Cooley, Deshpande & Tan, 2000).

Znanje sadržano u formi otkrivenih asocijativnih pravila o korišćenju web sajtova može se koristiti u cilju unapređenja dizajna web sajtova, povećanja njihove posećenosti i iskorišćenosti. Vlasnicima web sajtova vezanih za elektronsko poslovanje primena ovog znanja može povećati profit, dok korisnicima olakšava i ubrzava korišćenje web sajtova. Pored toga, otkrivena asocijativna pravila koriste se u raznim sistemima za preporuku web stranica, kao i u sistemima za povećanje performansi web servera keširanjem web stranica (Berendt, 2004; Chakrabarti, 2000; Liu, 2007; Srivastava, Cooley, Deshpande & Tan, 2000).

Algoritmi za otkrivanje asocijativnih pravila prvobitno su razvijeni za otkrivanje znanja u podacima iz takozvanih transakcionih baza podataka. Ovakve baze podataka

karakteristične su za podatke o proizvodima kupljenim u super-marketima, gde svaka transakcija predstavlja skup artikala u datoj potrošačkoj korpi. Proces otkrivanja asocijativnih pravila podrazumeva pronalaženje korelacija u podacima, u kojima prisustvo jednog skupa atributa u određenoj transakciji implicira prisustvo drugog skupa atributa u istoj transakciji, sa određenim stepenom sigurnosti (Agrawal, Imielinski & Swami, 1993; Agrawal, Mannila, Srikant & Toivonen, 1995).

U domenu primene asocijativnih pravila na podatke o korišćenju web sajtova, pod transakcijom se najčešće podrazumeva web sesija, dok se atributi odnose na prisustvo pojedinih web objekta u web sesiji (Srivastava et al., 2000; Liu, 2007). Pri tome se pod web sesijom podrazumeva skup web objekata koje je određeni korisnik posetio u toku jednog pretraživanja web sajta. Asocijativna pravila tada podrazumevaju implikacije koje imaju formu *„Postojanje jednog skupa web objekata implicira postojanje drugog skupa web objekata u istoj web sesiji".*

Jedan od osnovnih problema koji negativno utiče na upotrebljivost asocijativnih pravila u raznim domenima primene je tendencija postojećih algoritama za otkrivanje asocijativnih pravila da generišu preveliki broj pravila u kojima se analitičari podataka teško snalaze pri odabiru stvarno korisnih pravila (Tan, Kumar & Srivastava, 2004; Wu & Kumar, 2009; Zaki, 2004). Kako bi se umanjio ovaj problem, u literaturi je predložen veliki broj matematičkih funkcija koje se mogu koristiti kao mere interesantnosti asocijativnih pravila. Mnogobrojna istraživanja predlažu i analiziraju primenu raznovrsnih mera interesantnosti asocijativnih pravila, ukazujući pri tome da nijedna od predloženih mera ne daje apsolutno kvalitetne rezultate (Carvalho, Freitas & Ebecken, 2005; Geng & Hamilton, 2006; Hilderman & Hamilton, 2013).

U slučaju asocijativnih pravila o korišćenju web sajtova, problem generisanja prevelikog broja asocijativnih pravila je dodatno pogoršan usled snažne korelacije između različitih web stranica, koja je najčešće posledica hiperlink strukture web sajtova. Generiše se preveliki broj asocijativnih pravila sa visokim vrednostima standardnih mera interesantnosti, koja su zapravo očekivana za analitičare podataka i web eksperte. Takva pravila negativno utiču na kvalitet znanja sadržan u skupu otkrivenih asocijativnih pravila o korišćenju web sajtova, smanjujući njegovu upotrebljivost (Cercone & An, 2002; Dimitrijević & Bošnjak, 2010; Facca & Lanzi, 2005; Sahar, 2010).

Shodno gore navedenom, opšti cilj ovog istraživanja je *razvoj teorijskog okvira i unapređenje metoda za pronalaženje i vrednovanje asocijativnih pravila u web server log podacima (C1).* U okviru opšteg cilja definisani su potciljevi:

C1.1: Uporedna analiza primene različitih mera interesantnosti pravila kroz aspekt njihove upotrebljivosti za analizu web server log podataka.

C1.2: Formulisanje smernica pri odabiru mera interesantnosti asocijativnih pravila.

Kao ispunjenje ovako definisanih ciljeva, očekivani teoretski doprinos ovog istraživanja je razvoj metoda kojima se povećava kvalitet otkrivenog znanja, kroz povećanje stvarne interesantnosti i korisnosti otkrivenih asocijativnih pravila u web server log podacima. Fokus istraživanja je na eliminisanju neinteresantnih asocijativnih pravila o korišćenju web sajtova, kao i na analizi različitih matematičkih mera interesantnosti i formulisanju smernica pri njihovim odabiru, tako da one što kvalitetnije rangiraju otkrivena asocijativna pravila. Na ovaj način povećava se upotrebljivost skupa otkrivenih asocijativnih pravila o korišćenju web sajtova od strane analitičara podataka.

Praktični cilj ovog istraživanja je *razvoj softverskog sistema za analizu web log podataka sa mogućnošću odabira postojećih i generisanja novih mera interesantnosti asocijativnih pravila (C2).* U okviru praktičnog cilja definisani su potciljevi:

C2.1: Analiza funkcionalnosti odredjenog softverskog sistema za data mining i predlog njegovog proširenja

C2.2: Proširenje funkcionalnosti postojećeg softverskog sistema u cilju poboljšanja kvaliteta otkrivenog znanja

Očekuje se da predloženi softverski sistem za analizu web log podataka sadrži proširenu funkcionalnost, kojom se povećava upotrebljivost skupa otkrivenih asocijativnih pravila o korišćenju web sajtova. Ovaj softverski sistem treba da implementira mogućnost eliminacije neinteresantnih asocijativnih pravila, kao i ugradnju novih mera interesantnosti, kojima se povećava kvalitet rangiranja otkrivenih asocijativnih pravila.

Ova doktorska disertacija je organizovana u sedam poglavlja. U prvom poglavlju date su definicije asocijativnih pravila, kao jedne od data mining metoda i opisana je njihova uloga u procesu otkrivanja znanja. Date su opšte definicije osnovnih pojmova i algoritama vezanih za proces otkrivanja asocijativnih pravila. Takođe je dat kratak pregled raznovrsnih domena u kojima se primenjuje znanje dobijeno otkrivanjem asocijativnih pravila.

Drugo poglavlje daje opšti pregled metoda korišćenih u web mining istraživačkoj oblasti. Poseban fokus je na metodama vezanim za otkrivanje znanja u web log podacima. Opisana je struktura podataka sadržanih u web log datotekama, kao i metode za pretprocesiranje ovih podataka, kojima se oni pripremaju za primenu data mining algoritama. Data je definicija asocijativnih pravila primenjenih u web mining-u, sa osvrtom na specifične probleme primene asocijativnih pravila u ovom domenu.

Uporedna analiza algoritama za otkrivanje asocijativnih pravila data je u trećem poglavlju. Dat je pregled najznačajnijih algoritama za otkrivanje asocijativnih pravila. Potom se analizira uticaj izbora parametara ovih algoritama na kvalitet otkrivenog znanja, pri čemu je dat pregled metoda za evaluaciju kvaliteta otkrivenih asocijativnih pravila.

Četvrto poglavlje se detaljnije bavi merama interesantnosti asocijativnih pravila i njihovim uticajem na kvalitet otkrivenog znanja o korišćenju web sajtova. U okviru ovog poglavlja predložena je primena metoda za eliminaciju neinteresantnih asocijativnih pravila i kvalitetnije rangiranje preostalih asocijativnih pravila prema modifikovanim merama interesantnosti. Predložene metode su poređene sa nekim od prethodnih istraživanja, kako teoretski, tako i na primerima iz stvarnih skupova podataka.

Softverski sistem, kojim se otkrivaju asocijativna pravila u web log podacima prikazan je u petom poglavlju. Sistem integriše sve faze procesa otkrivanja asocijativnih pravila, od pripreme web log podataka, otkrivanja asocijativnih pravila, do metoda za eliminisanje neinteresantnih asocijativnih pravila i kvalitetnije rangiranje preostalih asocijativnih pravila primenom standardnih i modifikovanih mera interesantnosti.

Šesto poglavlje sadrži rezultate eksperimentalnog istraživanja, koje je izvršeno korišćenjem implementiranog softverskog sistema za otkrivanje asocijativnih pravila,

primenjeno na dva stvarna skupa web log podataka. Poseban akcenat je dat na ispitivanju efikasnosti predloženih metoda za eliminaciju i kvalitetnije rangiranje asocijativnih pravila o korišćenju web sajtova, primenom različitih mera interesantnosti. Prikazani su rezultati smanjenja veličine skupa otkrivenih asocijativnih pravila o korišćenju web sajtova primenom predloženih metoda za eliminaciju neinteresantnih asocijativnih pravila. Poređeni su rezultati rangiranja asocijativnih pravila primenom standardnih i modifikovanih mera interesantnosti.

U zaključnom poglavlju dat je pregled teoretskih i praktičnih doprinosa ove disertacije. Komentarisana je ispunjenost postavljenih naučnih i pragmatičnih ciljeva i dat kratak pregled osnovnih rezultata istraživanja. Pored toga, predloženi su pravci budućeg istraživanja.

1 Asocijativna pravila

1.1 Definicije asocijativnih pravila

Algoritam za otkrivanje asocijativnih pravila primenljiv na otkrivanje paterna u transakcionim bazama podataka predložen je početkom devedesetih godina dvadesetog veka (Agrawal, Imielinski & Swami, 1993). U nastavku dajemo formalne definicije i opšte prihvaćenu terminologiju osnovnih pojmova u domenu otkrivanja asocijativnih pravila (Agrawal, Mannila, Srikant & Toivonen, 1995; Aggarwal & Yu, 1998; Pang-Ning, Steinbach & Kumar, 2006).

Obzirom da su metode za pronalaženje asocijativnih pravila prvobitno predložene za otkrivanje korelacija u podacima vezanim za prisustvo artikala u potrošačkim korpama kupaca u marketima, standarnda terminologija vezana za proces otkrivanja asocijativnih pravila je preuzeta upravo iz ovog domena.

Neka je $I = \{i_1, \dots, i_m\}$ neki skup elemenata, i neka je *stavka (item)* neki element skupa I ($i_k \in I$).

Definicija 1.1

Transakciona baza podataka $T = \{T_1, \dots, T_n\}$ je skup transakcija T_i, pri čemu $T_i \subseteq I$.

Definicija 1.2

Transakcija T_i *sadrži* skup stavki $X \subseteq I$ ako $X \subseteq T_i$.

Definicija 1.3

Support skupa stavki X u skupu transakcija $T = \{T_1, \dots, T_n\}$ dat je formulom $support(X) = \frac{|T^X|}{|T|}$, pri čemu $|T^X|$ označava broj transakcija koje sadrže skup stavki X, dok je $|T|$ kardinalitet skupa T.

Definicija 1.4

Asocijativno pravilo je implikacija oblika $X \rightarrow Y$, pri čemu $X, Y \subseteq I$ i $X \cap Y = \emptyset$.

Snaga asocijativnog pravila izražava se koristeći *support* i *confidence* mere, čiju definiciju dajemo u nastavku.

Definicija 1.5

Support mera asocijativnog pravila $X \rightarrow Y$ jednaka je support vrednosti skupa $X \cup Y$:

$$support(X \rightarrow Y) = support(X \cup Y) = \frac{|T^{X \cup Y}|}{|T|}$$

Support mera asocijativnog pravila $X \rightarrow Y$ u skupu transakcija $T = \{T_1, \dots, T_n\}$ predstavlja verovatnoću pojavljivanja skupa $X \cup Y$ u nekoj transakciji.

Definicija 1.6

Confidence mera asocijativnog pravila $X \rightarrow Y$ u skupu transakcija $T = \{T_1, \dots, T_n\}$ definisana je kao:

$$conf(X \rightarrow Y) = \frac{|T^{X \cup Y}|}{|T^X|}$$

Confidence mera asocijativnog pravila $X \rightarrow Y$ u skupu transakcija $T = \{T_1, \dots, T_n\}$ predstavlja uslovnu verovatnoću pojavljivanja skupa Y u nekoj transakciji ako ona sadrži skup X.

Definicija 1.7

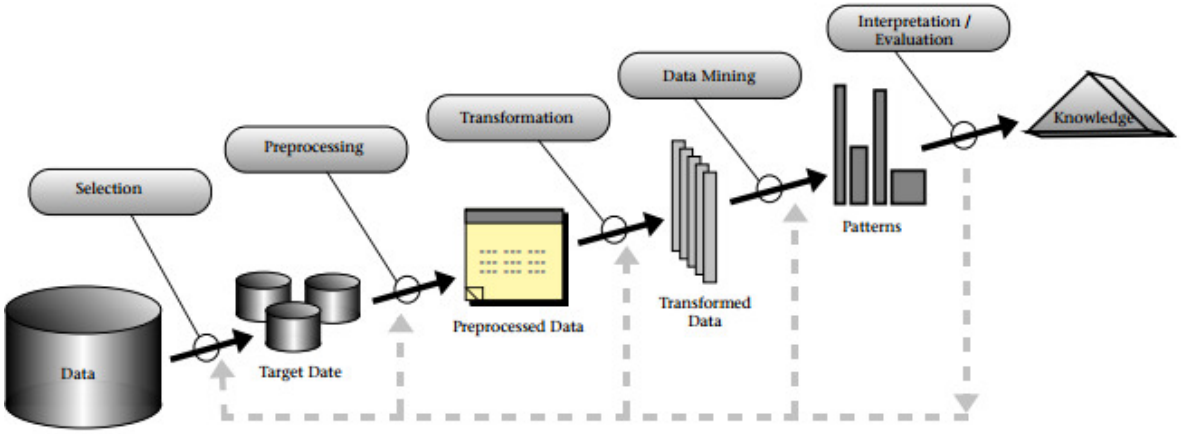
Skup elemenata $X \subseteq I$ je *frekventan* u skupu transakcija $T = \{T_1, \dots, T_n\}$ ako njegov support prelazi zadati minimalni support prag.

1.2 Uloga asocijativnih pravila u procesu otkrivanja znanja

Otkrivanje asocijativnih pravila u velikim bazama podataka je jedna od najčešće korišćenih metoda u okviru popularne istraživačke oblasti „Otkrivanje znanja u bazama podataka“. Prema jednoj od definicija, otkrivanje znanja u bazama podataka je “netrivijalni proces identifikovanja validnih, prethodno nepoznatih, potencijalno korisnih, korisniku razumljivih paterna u podacima” (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996; Wu & Kumar, 2009).

Proces otkrivanja znanja u bazama podataka (Knowledge Discovery in Databases – KDD) sastoji se iz više koraka, kao što je prikazano na slici 1.1. Tu se podrazumeva selekcija podataka iz baze podataka, njihovo pretprocesiranje i transformacija. Potom se na tako pripremljene podatke primenjuju metode čiji je cilj otkrivanje paterna u podacima. Ovaj centralni, analitički deo KDD procesa je izdvojen kao posebna naučno-istraživačka

disciplina pod nazivom *data mining* (rudarenje podataka) (Hand, Mannila & Smyth, 2001). Poslednji korak KDD procesa predstavlja evaluacija otkrivenih paterna, pri čemu se neki od njih mogu interpretirati kao novo znanje (Fayyad, Piatetsky-Shapiro & Uthurusamy, 1996).



Slika 1.1 Proces otkrivanja znanja (slika javno dostupna na Internetu)

Data mining je interdisciplinarna oblast koja se nalazi na preseku između statistike, mašinskog učenja, upravljanja bazama podataka i veštačke inteligencije. Prema jednoj od definicija, data mining je netrivijalni proces identifikovanja takvih paterna u velikim skupovima podataka, koji imaju sledeće karakteristike: validni, novi, korisni i razumljivi (Han, Kamber & Pei, 2011).

Za razliku od statistike, data mining se bavi analizom velikih baza podataka koji su prikupljani u različite svrhe, ali ne u cilju otkrivanja znanja u njima. U statistici podaci se najčešće prikupljaju na određene načine, kako bi se na efikasan način odgovorilo na unapred definisana pitanja. Za razliku od toga, data mining se bavi analizom podataka koji su prikupljeni u neke druge svrhe. Cilj data mining algoritama je da se na efikasan način otkriju neočekivani modeli, odnosno paterni, u već prikupljenim podacima (Hand, Mannila & Smyth, 2001; Han, Kamber & Pei, 2011).

Definicije pojmova vezanih za algoritme otkrivanja asocijativnih pravila biće date u sledećem poglavlju.

1.3 Algoritmi za otkrivanje asocijativnih pravila

Algoritmi za otkrivanje asocijativnih pravila prvi put su primenjeni u takozvanim transakcionim bazama podataka. Klasičan primer skupa podataka na kome se može primeniti proces otkrivanja asocijativnih pravila je skup svih transakcija nastalih prilikom kupovine artikala u nekom marketu tokom nekog vremenskog perioda. Pri tome, transakcija predstavlja skup artikala u korpi potrošača prilikom pojedine kupovine. Asocijativno pravilo je tada implikacija oblika $\{X_1, \dots, X_k\} \rightarrow \{Y_1, \dots, Y_l\}$, pri čemu su X_i i Y_i pojedini artikli u prodavnici, i kojoj je pridružen izračunati stepen poverenja. Klasičan primer asocijativnog pravila je: „Kupci koji kupuju mleko i hleb takođe kupuju pivo, u 35% slučajeva“ (Brin et al., 1997).

Proces otkrivanja asocijativnih pravila u datom skupu transakcija podrazumeva pronalaženje svih asocijativnih pravila čije support (definicija 1.5) i confidence (definicija 1.6) vrednosti prelaze zadate minimalne pragove (Agrawal & Srikant, 1994; Agrawal, et.al., 1995).

Support mera ima značajnu osobinu anti-monotonosti, koja se koristi u algoritmima za otkrivanje asocijativnih pravila, kako bi se povećala efikasnost algoritama.

Confidence mera se može koristiti kao aproksimacija pouzdanosti pravila. Pored toga, ona je razumljiva analitičarima podataka jer označava uslovnu verovatnoću.

Problem otkrivanja asocijativnih pravila izražen koristeći support i confidence mere može biti formalno formulisan na sledeći način:

Definicija 1.8

Ako je dat skup stavki I i skup transakcija T, problem otkrivanja asocijativnih pravila je definisan kao pronalaženje skupa svih asocijativnih pravila čija vrednost support mere prelazi zadati *minsupp* prag, i čija vrednost confidence mere prelazi zadati *minconf* prag.

Naivni „brute-force“ pristup za rešavanje problema otkrivanja svih asocijativnih pravila za dati skup stavki I i skup transakcija T, podrazumevao bi izračunavanje support i confidence vrednosti za svako moguće asocijativno pravilo, kojih ima eksponencijalno mnogo. Međutim, zahvaljujući osobini anti-monotonosti support mere, postoje algoritmi koji smanjuju prostor pretraživanja i čine ovaj proces znatno efikasnijim.

U većini slučajeva proces otkrivanja asocijativnih pravila odvija se u dve faze:

1) Otkrivanje svih frekventnih skupova stavki za zadati minimalni support prag
2) Generisanje svih asocijativnih pravila na osnovu prethodno otkrivenih frekventnih skupova stavki, koja prelaze zadati minimalni confidence prag.

Algoritmi koji realizuju drugu fazu procesa otkrivanja asocijativnih pravila su relativno jednostavni i efikasni. Međutim, prva faza ovog procesa je računski zahtevna. Prvo, broj potencijalno frekventnih skupova stavki raste eksponencijalno sa veličinom skupa I. Drugo, potrebno je višestruko prolaženje kroz bazu podataka kako bi se izvršilo prebrojavanje transakcija koje sadrže potencijalno frekventne skupove stavki.

Postoje razne verzije i optimizacije algoritama za otkrivanje asocijativnih pravila čiji će detaljniji pregled biti dat u poglavlju 3.

1.4 Domeni primene asocijativnih pravila

Pored otkrivanja paterna u potrošačkim korpama, asocijativna pravila pronalaze primenu u vrlo raznovrsnim domenima. Ona se primenjuju u trgovini, medicinskoj dijagnostici, bioinformatici, za otkrivanje paterna u podacima dostupnim na web-u, i u mnogim drugim oblastima. U ovom poglavlju dajemo samo neke od primera istraživanja u kojima se asocijativna pravila primenjuju u pomenutim domenima.

Asocijativna pravila se odavno koriste u softverskim aplikacijama u oblasti trgovine, u cilju analize preferenci kupaca, kao i preporučivanja proizvoda na osnovu drugih proizvoda koje je dati kupac već kupio (Sarwar et al., 2000; Schafer, Konstan & Riedl, 2001).

U sistemima za podršku odlučivanju (decision support systems) mogu se koristiti takozvana klasifikaciona asocijativna pravila (CARs). Ova vrsta asocijativnih pravila sadrži određeni ciljni, tzv. target-atribut sa desne strane pravila, i koristi se u okviru algoritama za klasifikaciju. Neki od algoritama za otkrivanje klasifikacionih asocijativnih pravila predloženi su u istraživanjima (Nguyen, Vo, Hong & Thanh, 2013; Nguyen & Nguyen, 2015).

Asocijativna pravila se primenjuju u raznim domenima u oblasti medicine. Jedan od mnogobrojnih primera primene asocijativnih pravila u medicinskoj dijagnostici je

istraživanje (Soni, Ansari, Sharma & Soni, 2011), gde se ona koriste u okviru sistema za predikciju srčanih bolesti. Još jedan primer je istraživanje (Wang & Zheng, 2012), gde se otkrivaju asocijativna pravila koja povezuju nivo endokrinih hormona kod pacijenata sa određenim bolestima.

Primena asocijativnih pravila u oblasti genetike i molekularne biologije je široko zastupljena (Fernald, Capriotti, Daneshjou, Karczewski & Altman, 2011). Na primer, Manda, McCarthy & Bridges (2013) koriste multi-level asocijativna pravila za otkrivanje paterna u ontologiji gena. Shaikh & Beyene (2015) predlažu korišćenje asocijativnih pravila za otkrivanje grupa genotipa vezanih za invazivne komplikacije oboljenja virusom sa zapadnog nila.

Oblasti u kojima se primenjuju asocijativna pravila vrlo su raznovrsne. Na primer, asocijativna pravila se primenjuju u okviru sistema za otkrivanje kriminalnih prevara (Phua, Lee, Smith & Gayler, 2010). Kamsu-Foguem, Rigal & Mauget (2013) primenjuju asocijativna pravila za povećanje kvaliteta procesa prizvodnje. Abdullah, Herawan & Deris (2014) otkrivaju asocijativna pravila u podacima o upisu studenata u visokoškolsku ustanovu.

Postoje mnogobrojna istraživanja koja se bave primenom asocijativnih pravila za otkrivanje paterna u podacima dostupnim na web-u (web mining), što je detaljnije razmatrano u sledećem poglavlju.

2 Web mining

Ogromne količine podataka dostupnih na jedinstvenoj svetskoj mreži (*web* u daljem tekstu) predstavljaju plodno tle za primenu metoda iz oblasti otkrivanja znanja u bazama podataka. Istraživačka oblast nazvana *„rudarenje web podataka“* (*web mining*) podrazumeva korišćenje data mining metoda za automatsko otkrivanje i ekstrakciju informacija iz web dokumenata i servisa (Etzioni, 1996; Berendt, 2004; Singh & Singh, 2010).

2.1 Pregled web mining metoda

U okviru web mining-a kao oblasti istraživanja izdvajaju se tri podoblasti: rudarenje sadržaja na web-u (*web content mining*), rudarenje strukture web-a (*web structure mining*), i rudarenje podataka o korišćenju web-a (*web usage mining*) (Madria, Bhowmick, Ng & Lim, 1999; Kosala & Blockeel, 2000; Anand, Mulvenna & Chevalier, 2004).

2.1.1 Rudarenje sadržaja na web-u

Rudarenje sadržaja na web-u (web content mining) odnosi se na otkrivanje potencijalno korisnih informacija u indeksiranim sadržajima na web-u, koji mogu biti različitih vrsta: tekstualni, slike, audio, video (Chakrabarti, 2000; Subašić & Berendt, 2009). Pri tome, posebno se izdvaja oblast koja se bavi rudarenjem raznovrsnog sadržaja na web-u, pod nazivom „multimedia data mining“ (Oh et.al., 2003; Zaiane et al, 1998).

Ogroman deo podataka na web-u su nestruktuirani tekstualni podaci, koji se ne nalaze u formi tabela ili baze podataka. Rudarenjem tekstualnih sadržaja bavi se oblast pod nazivom „tekst data mining“. Rudarenje tekstualnih sadržaja obuhvata kategorizaciju teksta, klasifikaciju, klasterovanje dokumenata, otkrivanje potencijalno interesantnih paterna i pravila u tekstu (Hotho, Nürnberger & Paaß, 2005; Tan, 1999; Miner, 2012).

Jedan od pravaca istraživanja bavi se metodama reprezentacije tekstualnih dokumenata. Pri tome se može koristiti vektorska reprezentacija skupa reči koji se pojavljuju u datom tekstu korišćenjem na različite načine definisanih istaknutih svojstava (features) (Clifton, Cooley & Rennie, 2004; Kaski et.al., 1998). Kao istaknuta svojstva mogu se koristiti

ključne reči koje se pojavljuju u obučavajućem skupu podataka. Pri tome se može koristiti metoda Latentnog semantičkog indeksiranja (Deerwester et al., 1990), kojom se smanjuje dimenzija originalnog vektora kojim je predstavljen dati dokument. Na primer, jedna od tehnika koje se koriste u tu svrhu je „stemming“, gde se reči koje imaju isti koren svrstavaju u istu grupu otklanjanjem nastavaka. Za iscrpan pregled metoda iz tekst mining oblasti mogu se pogledati mnogobrojni pregledni radovi kao što su (Hotho, Nürnberger & Paaß, 2005; Miner, 2012).

2.1.2 Rudarenje strukture web-a

Rudarenje strukture web-a (web structure mining) primenjuje se na podatke o povezanosti dokumenata na web-u preko hiperlink strukture (Kautz, Selman & Shah, 1997; Getoor & Diehl, 2005). Pri tome se čitava struktura web-a može posmatrati kao graf povezanih dokumenata (Büchner et.al., 2000). Neke od metoda koje se primenjuju u okviru rudarenja strukture web-a inspirane su tehnikama iz oblasti društvenih mreža i analize citiranosti (Chakrabarti, 2000; Kosala & Blockeel, 2000). Na primer, na osnovu ulaznih (incoming) i izlaznih (outgoing) linkova mogu se otkrivati specifične vrste web stranica, kao što su „hubs“ i „authorities“ (Borodin et.al., 2001; Ding et.al., 2004).

Popularni su algoritmi kojima se modeluje topologija web-a, kao što je algoritam HITS. Postoje i njegove optimizacije u kojima se link strukturi web-a dodaje informacija o značenju sadržaja na koji se odnose linkovi, kao i metode za filtriranje izuzetaka (outliers) (Ding et al., 2002). Čuveni algoritam PageRank (Page et al., 1999) i njegove mnogobrojne optimizacije (Langville & Meyer, 2011) implementirane su u okviru Google pretraživača u cilju otkrivanja kvalitetnih i popularnih web stranica, što se koristi prilikom njihovog rangiranja.

2.1.3 Rudarenje podataka o korišćenju web-a

Rudarenje podataka o korišćenju web-a (web usage mining) odnosi se na otkrivanje znanja u podacima o korišćenju web sajtova (Cooley, Mobasher & Srivastava, 1997; Dong, 2009; Kosala & Blockeel, 2000; Singh & Singh, 2010). Repozitorijum ovih podataka najčešće predstavljaju tekstualne log datoteke koje se generišu na web serverima širom sveta, a u kojima se čuvaju podaci o pristupima web dokumentima i ostalim web objektima od strane posetioca web sajtova. Otkrivanje asocijativnih pravila o korišćenju

web sajtova, što je tema ove doktorske disertacije, pripada upravo ovoj podoblasti web mining-a.

Praktične primene web usage mining metoda mogu se svrstati u kategorije: (a) personalizacija web sadržaja, (b) unapređenje efikasnosti navigacije kroz keširanje web stranica, (c) unapređenje dizajna web sajta, (d) povećanje satisfakcije kupaca u slučaju e-commerce web sajtova (Facca & Lanzi, 2005; Berendt et al., 2011).

Personalizacija web sadržaja može se postići poređenjem ponašanja korisnika (njegove navigacione putanje) sa tipičnim ponašanjem prethodnih korisnika. Na ovaj način korisniku se mogu preporučiti linkovi koje su prethodni korisnici, koji se ponašaju na sličan način takođe posetili. Ova metoda najčešće se realizuje u okviru sistema za preporuku sadržaja (Gavalas & Kenteris, 2011; Niwa & Honiden, 2006; Schafer, Konstan & Riedl, 2001). Pri tome su u algoritme implementirane u okviru sistema za preporuku sadržaja često ugrađene i konceptne ontologije generisane na osnovu ekspertskog znanja (Liang & Wang, 2004; Szomszor et.al., 2007).

Keširanje web stranica u cilju unapređenja efikasnosti navigacije kroz web sajt predloženo je u mnogobrojnim istraživanjima (Bonchi et al., 2001; Yang & Zhang, 2003). Postoje i alati za analizu uspešnosti keširanja web stranica u cilju smanjenja opterećenja web servera, kao što je (Wang, Balasubramanian, Krishnamurthy & Wetherall, 2013) kojim se generišu izveštaji o keširanju web stranica na osnovu opterećenja mreže, parsiranja web stranica, i aktivnosti na samom web pretraživaču.

Unapređenje strukture web sajta primenom web usage mining metoda je takođe aktuelna oblast istraživanja. Na primer, Carmona et al. (2012) predlažu unapređenje e-commerce web sajta kombinacijom klasterovanja podataka o ponašanju korisnika web sajta i otkrivanju asocijativnih o korišćenju web sajta. Fu, Shih, Creado & Ju (2002) koriste metod za unapređenje strukture web sajta, baziran na klasifikaciji web stranica u dve kategorije – indeksne strane i sadržajne strane, na osnovu heuristika kao što su vrsta web stranice (html), broj linkova na web stranici, učestalost sesija u kojima je data web stranica poslednja posećena i prosečno vreme koje korisnik provodi na datoj stranici.

Unapređenje strukture web sajta analizom paterna o korišćenju web sajta primenljivo je i u oblasti e-learning sistema (Romero & Ventura, 2010). Pri tome, moguće je

inkorporirati i ontologiju generisanu od strane eksperta, kako bi se navigacioni paterni obogatili podacima o konceptima i njihovim relacijama (Becker & Vanzin, 2010).

Implementirani su mnogobrojni sistemi u kojima se web mining metode primenjuju u cilju povećanja satisfakcije kupaca i profita u okviru aplikacija elektronskog poslovanja (Facca & Lanzi, 2005; Carmona et al., 2012). Pri tome, moguće je primeniti taksonomiju na podacima o korisnicima u okviru web usage mining inteligentnog sistema, kao što je predloženo u (Devi, Devi, Rani & Rao, 2012).

2.2 Struktura web log podataka

Podaci o korišćenju web sajtova čuvaju se na web serveru u obliku tekstualnih *„web server log datoteka*“. U ovom poglavlju opisan je format web server log datoteka, a potom i pregled metoda njihove pripreme za web usage mining proces.

2.2.1 Originalni format web log podataka

Struktura web server log datoteka može biti u *„Common log“* formatu, ili u *„Extended common log*“ formatu. Svaka linija ove datoteke odnosi se na po jedan zahtev za web objektom koji se nalazi na web serveru (Chitraa, Davamani & Selvdoss, 2010; Cooley, Mobasher & Srivastava, 1999; Singh & Singh, 2010).

U common log formatu, svaka linija web log datoteke sadrži sledeće stavke:

- *IP adresa sa koje stiže zahtev*
- *korisničko ime ukoliko postoji log sistem na web sajtu*
- *datum i vreme stizanja zahteva na web server*
- *status kod*
- *broj poslatih bajtova*

U extended common log formatu, pored gore navedenih podataka, linija datoteke takođe sadrži:

- *url web objekta koji je bio zahtevan pre aktuelnog objekta (referrer url)*
- *string koji opisuje ime i verziju korisnikovog web pretraživača*

Pored korišćenja podataka iz web server log datoteka, postoje i drugi načini sakupljanja podataka o korišćenju web sajtova, kao što su web kolačići (cookies), formulari za

registraciju korisnika na web sajtu, kao i klijentski apleti za prikupljanje podataka (Chitraa, Davamani & Selvdoss, 2010).

Web kolačići sadrže jedinstveni identifikator posetioca web sajta, koji se šalje na web server sa svakim novim zahtevom za web objektom. Na ovaj način mogu se nepogrešivo povezati zahtevi za web objektima sa korisnicima web sajta koji su ih uputili na web server. Međutim, problem sa korišćenjem web kolačića je što su oni često zabranjeni na korisnikovom web pretraživaču zbog privatnosti podataka (Singh & Singh, 2010).

Formulari za registraciju korisnika web sajta retko se koriste kao izvor podataka o posetama web sajtu, i to samo u retkim slučajevima gde korisnici sami odlučuju da se registruju i unesu svoje podatke (Chitraa, Davamani & Selvdoss, 2010).

Klijentski apleti za prikupljanje podataka su pouzdan način prikupljanja podataka o korišćenju web sajtova. Međutim, oni se takođe mogu koristiti samo u slučajevima kada korisnici eksplicitno prihvate njihovu instalaciju na svom web pretraživaču (Tao, Hong & Su, 2008).

2.2.2 Pretprocesiranje web log podataka

Kako bi se na web log podatke mogle primeniti web usage mining metode neophodno je izvršiti pripremu, odnosno pretprocesiranje ovih podataka (Cooley, Mobasher & Srivastava, 1999).

Proces pretprocesiranja podataka iz web server log datoteka sastoji se iz najmanje dva koraka: (a) eliminacija irelevantnih web zahteva, (b) rekonstrukcija korisničkih web sesija.

Eliminacija irelevantnih web zahteva

Postoje dve vrste web zahteva koji se skladište u web log datotekama, a koji su irelevantni za proces otkrivanja znanja u podacima o korišćenju web sajtova (a) zahtevi za irelevantnim web objektima, (b) robotski web zahtevi.

U web log datotekama čuvaju se svi web zahtevi koje klijentski web pretraživač šalje na web server, uključujući zahteve za web objektima ugrađenim u web stranice, kao što su slike, multimedijalni sadržaji, datoteke koje sadrže razne skriptove, stilove, i slično.

Eliminacija ovakvih irelevantnih objekata je obično relativno jednostavan proces. Prepoznavanje irelevantnih objekata najčešće se zasniva na prepoznavanju standardnih ekstenzija koje su deo naziva web objekata datog tipa (jpg, png, mov, css, i slično) (Chitraa, Davamani & Selvdoss, 2010). Ovakvi irelevantni objekti se eliminišu prilikom učitavanja sadržaja web log datoteka u okviru sistema za otkrivanje znanja u podacima o korišćenju web sajtova, a u skupu podataka se zadržavaju samo zahtevi za relevantnim web dokumentima nekog web sajta.

Pod robotskim web zahtevima podrazumevaju se zahtevi za web objektima koji nisu inicirani od strane stvarnog posetioca web sajta, nego od strane mašina za pretraživanje i indeksiranje Web-a. Robotski web zahtevi „zagušuju" skup podataka i mogu negativno uticati na proces otkrivanja znanja o ponašanju korisnika web sajtova. Jedan deo ovakvih web zahteva je moguće eliminisati prepoznavanjem imena poznatih web robota u *url* adresi sa koje web zahtev dolazi. Dodatno prečišćavanje robotskih web zahteva moguće je izvršiti primenom raznih heuristika, koje se zasnivaju na vremenskoj udaljenosti između različitih web zahteva koji stižu sa iste IP adrese, ukupnom broju web zahteva koji u relativno kratkom intervalu stižu sa IP adrese, i slično (Cooley, Mobasher & Srivastava 1999; Chitraa, Davamani & Selvdoss, 2010).

Rekonstrukcija korisničkih web sesija

Pod pojmom *web sesije* podrazumeva se niz zahteva za relevantnim web dokumentima nekog web sajta, upućenih na web servere od strane nekog posetioca web sajta tokom jedne njegove sesije pretraživanja.

Web zahtevi koji pristižu na web server od strane mnoštva posetioca web sajtova hostovanih na tom web serveru beleže se u web log datoteke onim redosledom kojim pristižu na web server. Da bi se izvršila rekonstrukcija web sesija potrebno je jedinstveno identifikovati posetioca web sajta, a potom razbiti skup svih web zahteva načinjenih od strane datog posetioca na web sesije (Li & Feng, 2009; Spiliopoulou, Mobasher, Berendt & Nakagawa, 2003).

Najjednostavniji način za identifikaciju posetioca web sajta je oslanjati se na IP adresu sa koje pristiže dati web zahtev. U slučaju proxy servera, gde više posetioca web sajta koriste deljenu IP adresu, potrebno je koristiti sofisticiranije metode za identifikaciju

posetioca, koje mogu uzimati u obzir topologiju web sajta, na osnovu koje se nizovi web zahteva spajaju u moguće web sesije.

Niz svih web zahteva načinjenih od strane određenog posetioca web sajta razbija se u skup web sesija, tako da se jedna web sesija odnosi na jedan događaj pretraživanja web sajta od strane datog posetioca. U ovu svrhu mogu se koristiti heuristike zasnovane na vremenskom intervalu između različitih web zahteva i/ili informacije o topologiji web sajta (Dettmar, 2004; Li & Feng, 2009).

2.3 Primena asocijativnih pravila u web mining-u

Kada se asocijativna pravila primenjuju u cilju otkrivanja znanja o korišćenju web sajtova, pod *stavkom (item)* podrazumeva se *web objekat* nekog web sajta, pod *transakcijom* se podrazumeva *web sesija*, dok se pod *transakcionom bazom podataka* podrazumeva *skup svih web sesija* za dati web sajt u nekom određenom vremenskom periodu (Srivastava et al., 2000; Liu, 2007).

U opštem slučaju asocijativna pravila o korišćenju web sajtova tada imaju oblik: *„Postojanje jednog skupa web objekata u web sesiji implicira postojanje drugog skupa web objekata u datoj sesiji, sa određenim stepenom poverenja“*. Primer asocijativnog pravila o korišćenju web sajta bi mogao biti „Ako web sesija sadrži web stranice *home.html* i *contents.html*, tada sadrži i web stranicu *products.html*, sa stepenom poverenja 40%“.

Asocijativna pravila o korišćenju web sajtova se u mnogim istraživanjima otkrivaju implementacijom neke verzije Apriori algoritma (Nanopoulos, Katsaros & Manolopoulos, 2002; Joshi, Joshi & Yesha, 2003; Dimitrijević & Bošnjak, 2011).

Cercone & An (2002) porede rezultate primene različitih mera interesantnosti na rangiranje asocijativnih pravila o korišćenju web sajtova. Međutim, veliki broj asocijativnih pravila rangiranih među prvih 10 u ovom istraživanju imaju ekstremno visoke vrednosti confidence mere (Cercone & An, 2002), što je vrlo verovatno rezultat direktne povezanosti web stranica hiperlink strukturom. Kako bi se umanjio uticaj hiperlink strukture na kvalitet otkrivenih pravila o korišćenju web sajtova potrebno je

umanjiti interesantnost pravila koja sadrže web stranice direktno povezane hiperlinkovima (Dimitrijević, Subić & Bošnjak, 2014).

Semantički obogaćena asocijativna pravila o korišćenju web sajtova mogu se generisati korišćenjem informacije o konceptima i njihovim relacijama, pored podataka o url web stranice koja je sadržana u web sesijama (Becker & Vanzin, 2010; Senkul & Salin, 2012).

Posebna linija istraživanja bavi se otkrivanjem asocijativnih pravila o korišćenju web sajtova na način koji neće ugroziti privatnost korisnika web sajtova (Yan, Jiajin & Dongmei, 2010; Dimitrijević & Krunić, 2014).

Aktuelna oblast je i otkrivanje fazi asocijativnih pravila, koja se takođe može primeniti u web mining domenu (Wong, Shiu & Pal, 2001).

Asocijativna pravila o korišćenju web sajtova imaju primenu u sistemima za preporuku web stranica (Kazienko, 2007; Lazcorreta, Botella & Fernández-Caballero, 2008; Zhang & Jiao, 2007), sistemima za preporuku za unapređenje dizajna web sajtova (Carmona et al. 2012), sistemima za keširanje web stranica (Nanopoulos, Katsaros & Manolopoulos, 2002), i raznim e-commerce aplikacijama (Devi, Devi, Rani & Rao, 2012).

Jedan od osnovnih problema prilikom korišćenja asocijativnih pravila o korišćenju web sajtova je visoka korelacija između web stranica, koja je najčešće rezultat topologije web sajta, a ne stvarnog interesovanja korisnika sajta (Cercone & An, 2002; Cooley, 2003; Dimitrijević & Bošnjak, 2010; Kazienko & Pilarczyk, 2008; Lee, Lo & Fu, 2011; Sahar, 2010; Zaki, 2004). Posledica toga je da statističke mere interesantnosti neadekvatno rangiraju otkrivena asocijativna pravila o korišćenju web sajtova, dajući prednost asocijativnim pravilima koja sadrže web stranice povezane hiperlink strukturom web sajta. Stoga su u okviru ovog istraživanja (poglavlja 4.2 i 4.3) predložene metode koje umanjuju ovaj problem, utičući na smanjenje interesantnosti i eliminaciju jednog dela otkrivenih asocijativnih pravila o korišćenju web sajtova.

3 Uporedna analiza AP algoritama

U ovom poglavlju dajemo analizu najznačajnijih algoritama za otkrivanje asocijativnih pravila u transakcionim bazama podataka, kao i metoda za njihovu evaluaciju.

3.1 Pregled najznačajnijih algoritama za otkrivanje asocijativnih pravila

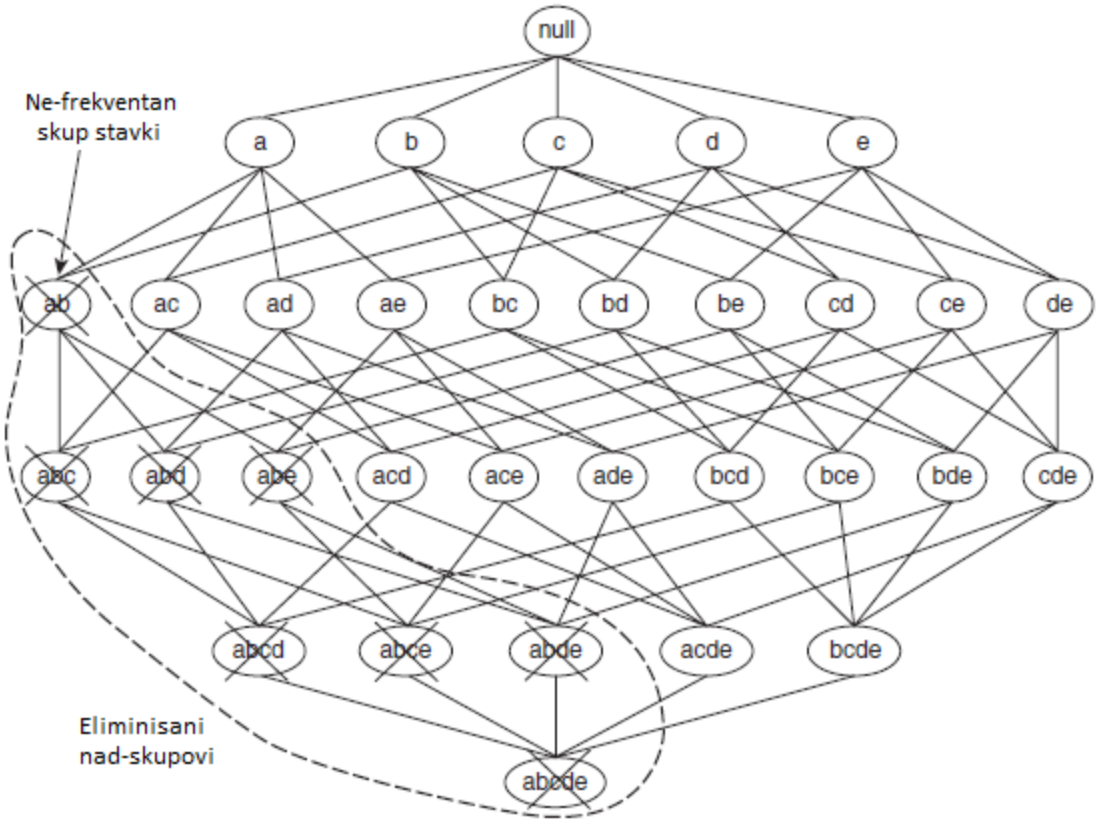
3.1.1 Algoritam „Apriori“

Apriori je prvi predloženi i još uvek jedan od najpopularnijih algoritama za generisanje frekventnih skupova, koji se zasniva na osobini *anti-monotonosti* support mere (Agrawal & Srikant, 1994).

Definicija 3.1

Neka je dat skup stavki I i neka mera f definisana na skupu svih podskupova od I. Mera f je *anti-monotona* ako važi: $\forall X, Y \subseteq I \ \ (X \subseteq Y) \rightarrow f(X) \geq f(Y)$

Lako se može pokazati da support mera zadovoljava princip anti-monotonosti. Ova osobina koristi se u Apriori algoritmu da se smanji prostor pretraživanja prilikom generisanja svih kandidata za frekventne skupove stavki, kao što je prikazano na slici 3.1 i objašnjeno u nastavku.



Slika 3.1: Eliminacija kandidata za frekventne skupove stavki koristeći osobinu anti-monotonosti

Generisanje frekventnih skupova vrši se po principu „prvo u širinu“ (breath-first). Najpre se generišu svi frekventni skupovi stavki dužine 1. Na slici 3.1 njih čine sve stavke na prvom nivou mreže, obeležene slovima a, b, c, d, e. U svakom sledećem koraku na osnovu skupova dužine k čiji support prelazi zadati minimalni prag (frekventni skupovi), generišu se svi kandidati za frekventne skupove dužine $k+1$. Potom se izračunava support kandidata i zadržavaju kandidati čiji support prelazi minimalni prag.

Generisanje frekventnih skupova stavki je računski zahtevan deo procesa otkrivanja asocijativnih pravila od koga zavisi efikasnost čitavog procesa (Ceglar & Roddick, 2006). Stoga je optimizacija algoritama za generisanje frekventnih skupova stavki privukla veliku pažnju istraživača i postoje razne verzije ovog algoritma.

3.1.2 Optimizacije „Apriori“ algoritma

Apriori algoritam smanjuje broj kandidata za frekventne skupove stavki, čime se smanjuje računska kompleksnost algoritma, zahtevi za memorijom, kao i broj ulazno/izlaznih operacija. Ipak, problemi koje algoritam ne otklanja u potpunosti su potreba da se više puta skenira čitava baza podataka, kao i činjenica da je generisanje kandidata za frekventne paterne računski zahtevno. Optimizacije algoritma za generisanje frekventnih skupova predložene u literaturi imaju za cilj smanjenje broja prolaska kroz bazu podataka, kao i primenu raznih tehnika za smanjenje veličine skupa kandidata stavki ugrađene u proces njihovog generisanja.

Na primer, algoritam Apriori-TID, predložen od autora originalnog Apriori algoritma (Agrawal & Srikant, 1994) zahteva samo jedan prolazak kroz bazu podataka. AprioriHybrid algoritam predložen u istom radu, kombinuje originalni Apriori i Apriori-TID, pri čemu se postiže veća efikasnost od oba kombinovana algoritma (Agrawal & Srikant, 1994). Hipp et al. dalje unapređuju Apriori-TID algoritam koristeći heš stabla umesto brojača (Hipp et al. , Güntzer & Nakhaeizadeh, 2000). Apriori-Brave algoritam predložen u (Bodon, 2003) uključuje kriterijum za odlučivanje o kombinovanju između Apriori i Apriori-TID algoritama u zavisnosti od veličine slobodne memorije.

Algoritam u okviru kojeg se particioniše baza podataka predložen je u (Savasere, Omiecinski & Navathe, 1995). U prvoj fazi ovog algoritma baza podataka se deli na male particije koje se ne preklapaju. One se procesiraju nezavisno kako bi se pronašli njihovi kandidati za frekventne skupove podataka. Veličina particija se određuje tako da one mogu stati u osnovnu memoriju, te se ne zahteva prolazak kroz bazu podataka na disku. U drugoj fazi frekventni skupovi stavki pronalaze se na osnovu ovih kandidata, što zahteva jedan prolazak kroz bazu podataka. U okviru Dynamic Itemset Counting algoritma predloženog u radu (Brin et al., 1997) takođe se vrši particionisanje baze podataka, čime se smanjuje broj prolazaka kroz bazu podataka. Prebrojavanje transakcija u kojima se kandidat za frekventni skup stavki pojavljuje vrši se dinamički, pri čemu se koriste particionisani blokovi baze podataka.

3.1.3 „FP-growth“ algoritam

Popularni algoritam za generisanje frekventnih paterna „FP-growth“ prvi put je predložen u radu (Han, J., Pei, J., & Yin, Y. 2000), a potom proširen u verziji (Han, J., Pei, J., Yin, Y., & Mao, R. 2004). Algoritam se bazira na korišćenju kompleksnih struktura koje sadrže kondenzovanu reprezentaciju frekventnih paterna. Korišćenjem ovih struktura pri otkrivanju frekventnih paterna i generisanju asocijativnih pravila smanjuje se broj prolaženja kroz bazu podataka, čime se znatno povećava efikasnost izvršavanja algoritma.

Stablo frekventnih paterna (FP-tree) je struktura koja sadrži:

- koreni čvor označen „null“
- skup podstabala koji predstavljaju prefikse paterna
- tabelu zaglavlja frekventnih stavki

Pri tome su u podstablima, kao i u tabeli sve frekventne stavke sortirane prema njihovoj frekventnosti u transakcionoj bazi podataka.

Svaki čvor u podstablima sadrži tri polja:

- naziv stavke
- broj, koji se odnosi na broj transakcija koje sadrže deo paterna kojeg čine čvorovi stabla između korena i datog čvora
- vezu ka daljim čvorovima, koji se odnose na paterne koji dati čvor mogu nastaviti

Svaki element tabele zaglavlja sadrži tri polja:

- naziv stavke
- frekventnost stavke, odnosno broj pojavljivanja stavke u transakcijama
- vezu na listu čvorova u stablu sa istim nazivom stavke

Za generisanje opisanog stabla frekventnih paterna prolazi se kroz bazu podataka da bi se izračunala frekventnost svake stavke, posle čega se one sortiraju prema njihovoj frekventnosti. Potom se prolazi kroz bazu podataka da bi se svaka transakcija dodala u stablo frekventnih paterna, pri čemu se inkrementalno menjaju brojevi pojavljivanja pojedinih prefiks-paterna za datu transakciju.

Algoritam FP-growth potom koristi stablo frekventnih paterna da generiše sve frekventne paterne, pri čemu nije potreban nijedan dodatni prolazak kroz bazu podataka.

Mana algoritama koji koriste FP-stablo je što se veličina stabla povećava eksponencijalno sa brojem frekventnih stavki (Ceglar & Roddick, 2006). Dodatni problem je što u slučajevima interaktivnog data mining procesa, gde korisnik može promeniti minimalni support prag, čitavo FP-stablo mora biti ponovo generisano (Zhao & Bhowmick, 2003).

Postoje različite unapređene verzije osnovnog FP-growth algoritma kojima se povećava efikasnost algoritma korišćenjem kompaktnijih FP-stabala. Na primer, COFI-tree je dvostruko povezano stablo kojim se omogućuje prolazak kroz čvorove podstabala frekventnih paterna u oba pravca, predloženo u (El-Hajj & Zaïane, 2003). CATS-tree sadrži sve stavke, uključujući i one koje nisu frekventne, pa je njegova veličina veća u odnosu na FP-tree, ali olakšava interaktivni data mining proces (Cheung & Zaiane, 2003). CT-tree je kompaktnije stablo frekventnih stavki, a njegovi autori pokazuju da je algoritam kojim se generišu frekventni paterni u tom slučaju brži (Sucahyo & Gopalan, 2003). Isti autori predlažu još kompaktniju strukturu nazvanu CFP-tree, kojom se smanjuje veličina FP-stabla za oko 50% (Gopalan & Sucahyo, 2004).

3.2 Uticaj izbora parametara na kvalitet otkrivenog znanja

Osnovni parametri algoritma za otkrivanje asocijativnih pravila su minimalni support i confidence prag. Izbor ovih parametara utiče na veličinu skupa otkrivenih pravila, kao i na karakteristike otkrivenih asocijativnih pravila.

3.2.1 Odabir parametara algoritma za otkrivanje asocijativnih pravila

Odabirom minimalnog support i confidence praga tokom procesa otkrivanja asocijativnih pravila eliminiše se jedan deo neinteresantnih pravila. Međutim, korišćenje support i confidence mera interesantnosti ima ozbiljne nedostatke (Kotsiantis & Kanellopoulos, 2006; Wu & Kumar, 2009).

Formulacija confidence mere interesantnosti je takva da se ne uzima u obzir verovatnoća pojavljivanja desne strane pravila u podacima. Asocijativna pravila čija je desna strana sama po sebi visoko frekventna imaju povišene confidence vrednosti, što često ne reflektuje stvarnu statističku korelaciju leve i desne strane pravila. U mnogim

slučajevima asocijativno pravilo može imati visoku confidence vrednost, a da pri tome postoji čak i negativna statistička korelacija između leve i desne strane pravila.

Osobina anti-monotonosti support mere asocijativnih pravila koristi se u cilju smanjenja veličine prostora pretraživanja tokom izvršavanja algoritama za generisanje frekventnih skupova (Agrawal et al., 1995; Aggarwal & Yu, 1998; Pang-Ning, Steinbach & Kumar, 2006). Postavljanjem minimalnog support praga unapred se eliminiše jedan deo pravila koja su apriori neinteresantna analitičarima podataka, čime se smanjuje prostor pretraživanja. Međutim, problem sa korišćenjem support mere je što ona ne reflektuje direktno statističku korelaciju leve i desne strane asocijativnih pravila.

Tan, Kumar & Srivastava (2004) pokazuju da se postavljanjem minimalnog support praga na nisku vrednost većinom eliminišu asocijativna pravila koja imaju nizak stepen korelacije leve i desne strane. Pri tome, u istom istraživanju pokazano je i da se povećanjem vrednosti minimalnog support praga postiže podjednaka eliminacija pravila čija su leva i desna strana negativno ili nisko korelirane, kao i pravila čija su leva i desne strana visoko korelirane (Tan, Kumar & Srivastava, 2004). Dakle, povećavanjem vrednosti minimalnog support praga problem generisanja prevelikog broja asocijativnih pravila se može ublažiti, ali tada dolazi do gubitka velikog broja pravila čija su leva i desna strana visoko korelirane, i koja stoga mogu biti potencijalno interesantna analitičarima podataka.

Kako bi se umanjile negativne posledice korišćenja support i confidence mere pri generisanju asocijativnih pravila, u literaturi je predložen veliki broj matematičkih funkcija koje se takođe mogu koristiti kao mere interesantnosti asocijativnih pravila. Mnogobrojna istraživanja predlažu i analiziraju matematičke funkcije kao mere interesantnosti asocijativnih pravila, ukazujući pri tome da nijedna od predloženih funkcija ne daje apsolutno kvalitetne rezultate (Carvalho, Freitas & Ebecken, 2005; Geng & Hamilton, 2006; Hilderman & Hamilton, 2013).

3.2.2 Metode za evaluaciju kvaliteta otkrivenih asocijativnih pravila

Jedan od najvećih problema pri korišćenju otkrivenih asocijativnih pravila od strane analitičara podataka je što algoritmi za generisanje asocijativnih pravila rezultuju prevelikim brojem pravila (Hilderman & Hamilton, 2013; Tan, Kumar & Srivastava 2004).

Stoga je znatna količina istraživanja vezana za otkrivanje i povećanje upotrebljivosti asocijativnih pravila usmerena na definisanje kriterijuma za odabir kvalitetnih asocijativnih pravila (Geng & Hamilton, 2006; Kontonasios, Spyropoulou & De Bie, 2012).

U cilju evaluacije kvaliteta otkrivenih asocijativnih pravila predloženo je mnoštvo metoda koje se mogu podeliti u dve osnovne grupe – subjektivne i objektivne.

Subjektivne metode za evaluaciju asocijativnih pravila zasnivaju se na inkorporaciji subjektivnih preferenci u proces otkrivanja znanja i uključenju ljudskog eksperta u sam proces. Ove metode najčešće se baziraju na *vizualizaciji*, *definisanju paterna*, ili korišćenju *subjektivnih mera interesantnosti* (Sahar, 2010).

Vizualizacija asocijativnih pravila

Metode bazirane na vizualizaciji podrazumevaju korišćenje sistema koji omogućuju interakciju sa korisnicima tokom procesa otkrivanja asocijativnih pravila. Pri tome je uloga korisnika u ovakvim sistemima da interpretiraju i verifikuju otkrivena asocijativna pravila (Berendt, 2002).

Neki komercijalni sistemi za otkrivanje asocijativnih pravila sadrže prikaz svih otkrivenih pravila koja zadovoljavaju uslove minimalnog support i confidence praga u obliku dvodimenzionalnog grafa, gde ose predstavljaju levu, odnosno desnu stranu pravila.

Pored toga, predložene su i druge metode, kao što je korišćenje mozaik grafova pri vizualizaciji skupa otkrivenih pravila, čime se pored pravila mogu vizualizirati i tabele kontigencije koje odgovaraju datim pravilima (Hofmann, Siebes & Wilhelm, 2000). Pored toga, Ben Said, Guillet, Richard, Picarougne & Blanchard (2013) predlažu način vizualizacije otkrivenih asocijativnih pravila gde se uzima u obzir korelacija između leve i desne strane asocijativnog pravila.

Definisanje paterna asocijativnih pravila

Primenom metoda baziranim na paternima definisanim od strane analitičara podataka, koje otkrivena asocijativna pravila treba da zadovolje otkriva se samo jedan deo asocijativnih pravila, koja zadovoljavaju unapred definisane paterne (Baralis, Cagliero, Cerquitelli & Garza, 2012). Proces otkrivanja asocijativnih pravila postaje računski efikasniji ukoliko su zadati paterni takvi da se mogu ugraditi u algoritme za otkrivanje

asocijativnih pravila i smanjiti prostor pretraživanja (Ayad, El-Makky & Taha, 2001; Ng, Lakshmanan, Han & Pang, 1998).

Subjektivne metode za evaluaciju asocijativnih pravila

Subjektivne mere interesantnosti asocijativnih pravila bazirane na znanju iz određenog domena primene je vrlo teško jasno definisati i inkorporirati u proces otkrivanja znanja (Tan, Kumar & Srivastava 2004).

Carvalho, Freitas & Ebecken (2005) ispituju korelaciju između objektivnih mera interesantnosti i stvarnog interesovanja eksperata za otkrivena asocijativna pravila. Izvršena je eksperimentalna evaluacija različitih objektivnih mera interesantnosti i zaključeno da ne postoji mera interesantnosti koja zaista odgovara stvarnom interesovanju eksperata za otkrivena asocijativna pravila (Carvalho, Freitas & Ebecken, 2005).

Jedan od pristupa definisanju subjektivnih mera interesantnosti bazira se na korišćenju konceptnih hijerarhija (Pohle, Spiliopoulou, 2002). Tada se konceptna hijerarhija za dati domen primene definiše unapred od strane eksperta, a potom se mogu otkrivati asocijativna pravila koja su neočekivana jer sadrže stavke koje nisu blisko povezane u datoj konceptnoj hijerarhiji.

Još jedan pristup definisanju subjektivnih mera interesantnosti bazira se na dodeljivanju vrednosti pojedinim stavkama koje se mogu naći u asocijativnom pravilu. Otkrivanje potencijalno interesantnih asocijativnih pravila tada se bazira na otkrivanju onih pravila koja sadrže vrednije stavke. Ovaj pristup je primenljiv na domen otkrivanja asocijativnih pravila u potrošačkim korpama, gde je vrednost stavki određena profitom ili cenom pojedinih artikala.

Pregled subjektivnih metoda za evaluaciju asocijativnih pravila dat je u mnogobrojnim istraživanjima (Hilderman & Hamilton, 2013; Zhang, Zhang, Nie & Shi, 2009).

Objektivne metode za evaluaciju asocijativnih pravila

Objektivne metode za evaluaciju asocijativnih pravila uzimaju u obzir samo osobine skupa podataka na kome se otkrivaju asocijativna pravila i ne zahtevaju ekspertsko znanje. Od analitičara podataka zahteva se samo da specificira minimalni prag

interesantnosti, radi filtriranja neinteresantnih asocijativnih pravila. U ovu svrhu koriste se razne mere interesantnosti, definisane kao matematičke funkcije, čiji pregled je dat u poglavlju 4.1.

4 Mere interesantnosti AP algoritama

U ovom poglavlju dajemo pregled mera interesantnosti asocijativnih pravila, kao i predlog modifikovanih metoda za eliminaciju i rangiranje asocijativnih pravila o korišćenju web sajtova. Cilj predloženih metoda je povećanje kvaliteta otkrivenih asocijativnih pravila i njihove upotrebljivosti od strane analitičara podataka.

4.1 Pregled matematičkih funkcija kao mera interesantnosti asocijativnih pravila

Matematičke mere interesantnosti koje se koriste za rangiranje otkrivenih asocijativnih pravila definisane su na tabeli kontigencije asocijativnog pravila.

Tabela kontigencije asocijativnog pravila oblika $X \rightarrow Y$ data je u Tabeli 4.1. Oznaka X odnosi se na prisustvo skupa elemenata u transakcijama, dok se oznaka $\neg X$ odnosi na odsustvo skupa elemenata u transakcijama. Oznaka $n(X)$ predstavlja broj transakcija koje sadrže skup X, oznaka $n(\neg X)$ predstavlja broj transakcija koje ne sadrže skup X. Oznaka $n(XY)$ odnosi se na $n(X \cup Y)$, što predstavlja broj transakcija koje sadrže elemente oba skupa X i Y. Oznaka N predstavlja ukupan broj svih transakcija u skupu.

	Y	$\neg Y$	
X	$n(XY)$	$n(X\neg Y)$	$n(X)$
$\neg X$	$n(\neg XY)$	$n(\neg X\neg Y)$	$n(\neg X)$
	$n(Y)$	$n(\neg Y)$	N

Tabela 4.1. Tabela kontigencije za asocijativno pravilo oblika $X \rightarrow Y$

Piatetsky-Shapiro (1991) navodi tri osnovne osobine koje bi bilo koja objektivna mera interesantnosti *M* trebala da zadovolji (Piatetsky-Shapiro, 1991).

O1: *M* = 0 ako su *X* i *Y* statistički nezavisni, odnosno ako je $P(XY) = P(X)P(Y)$

O2: *M* se monotono povećava sa $P(XY)$ kada su $P(X)$ i $P(Y)$ konstantni.

O3: *M* monotono opada sa $P(X)$ (ili sa $P(Y)$) kada se ostali parametri ne menjaju.

U tabeli 4.2 dato je 10 matematičkih funkcija definisanih na tabeli kontigencije asocijativnog pravila, kao i pregled osnovnih osobina koje one ispunjavaju ili neispunjavaju kao mere interesantnosti asocijativnih pravila.

Osobina (O1) koju ne zadovoljavaju 6 od 10 matematički funkcija navedenih u tabeli 4.2 je unekoliko preoštro formulisana. Na primer, *lift* mera ne zadovoljava ovaj uslov, jer je njena vrednost jednaka 1 kada su *X* i *Y* statistički nezavisni, dok je u slučaju pozitivne korelacije veća od 1, a u slučaju negativne korelacije manja od 1. Tako neki istraživači predlažu formulaciju osobine (O1) gde se traži da mera interesantnosti ima neku konstantnu vrednost (a ne obavezno vrednost 0) kada su *X* i *Y* statistički nezavisni (Tan et al. 2004).

Naziv mere	Formula	O1	O2	O3
Support	$P(XY)$	-	+	-
Confidence	$P(Y\|X)$	-	+	-
Lift	$\frac{P(XY)}{P(X)P(Y)}$	-	+	+
Added value	$P(Y\|X) - P(Y)$	+	+	+
Leverage	$P(Y\|X) - P(X)P(Y)$	-	+	+
Conviction	$\frac{P(X\neg Y)}{P(X)P(\neg Y)}$	-	+	-
Piatetsky-Shapiro	$P(XY) - P(X)P(Y)$	+	+	+
Odds ratio	$\frac{P(XY)P(\neg X\neg Y)}{P(X\neg Y)P(\neg XY)}$	-	+	+
Yule's Q	$\frac{P(XY)P(\neg X\neg Y) - P(X\neg Y)P(\neg XY)}{P(XY)P(\neg X\neg Y) + P(X\neg Y)P(\neg XY)}$	+	+	+
Yule's Y	$\frac{\sqrt{P(XY)P(\neg X\neg Y)} - \sqrt{P(X\neg Y)P(\neg XY)}}{\sqrt{P(XY)P(\neg X\neg Y)} + \sqrt{P(X\neg Y)P(\neg XY)}}$	+	+	+

Tabela 4.2. Matematičke funkcije kao mere interesantnosti asocijativnih pravila

U tabeli 4.2 oznaka $P(X) = \frac{n(X)}{N}$ odnosi se na verovatnoću pojavljivanja skupa stavki X u nekoj transakciji. Oznaka $P(Y|X) = \frac{n(XY)}{n(X)}$ odnosi se na uslovnu verovatnoću pojavljivanja skupa stavki Y u nekoj transakciji, kada je poznato da se X u njoj takođe pojavljuje. Oznaka $P(XY)$ odnosi se na verovatnoću pojavljivanja elemenata skupa X i skupa Y u istoj transakciji, odnosno važi: $P(XY) = P(X \cup Y) = \frac{n(X \cup Y)}{N}$.

Support mera se koristi kao mera značaja asocijativnog pravila. Njena osobina anti-monotonosti je pri tome veoma korisna radi smanjenja prostora pretraživanja prilikom izvršavanja algoritma za pronalaženje potencijalnih asocijativnih pravila (poglavlje 3.1).

Confidence i *conviction* mere se često koriste kao mere tačnosti asocijativnog pravila. Međutim, one mogu dovesti do zbunjujućih rezultata, naročito kada je visok *support* desne strane pravila.

Conviction mera ima sličnosti sa lift merom. Međutim, ona nije simetrična u odnosu na levu i desnu stranu pravila, jer uzima se uzima u obzir prisustvo leve, i odsustvo desne strane pravila u transakcijama. Interesantna činjenica je međutim da je *conviction* mera monotona u odnosu na *confidence* i *lift*.

Lift mera se često koristi u data mining aplikacijama kao mera devijacije u odnosu na statističku nezavisnost (Brin et al. 1997; Brijs et al. 1999). Međutim, mana lift mere je njena osetljivost u odnosu na support leve i desne strane asocijativnog pravila. Tendencija je da asocijativna pravila koja imaju nizak support leve ili desne strane asocijativnog pravila imaju izrazito visoke lift vrednosti, i obrnuto. *Added value i Piatetsky-Shapiro* mere interesantnosti su srodne sa *lift* merom, ali nemaju ovu osobinu, te se nekada koriste umesto *lift* mere.

Odds ratio kao mera interesantnosti odnosi se na odnos izgleda da se dobije različit rezultat slučajne promenljive. Ako ne bi postojala korelacija između *X* i *Y*, tada bi izgled da se *X* pronađe u transakciji trebao biti jednak bez obzira da li *Y* postoji u toj transakciji ili ne. Tako se odnos izgleda da se leva strana pravila pronađe u transakciji ako ona sadrži ili ne sadrži desnu stranu pravila, i obrnuto, može koristiti da se odredi snaga korelacije između leve i desne strane pravila. Vrednosti *odds ratio* mere kreću se od 0 u slučaju maksimalno negativne korelacije, do ∞ u slučaju maksimalno pozitivne

korelacije. *Yule's Q i Yule's Y* mere interesantnosti su normalizacije u odnosu na *odds ratio*, koje njegovu vrednost svode u interval $[-1, +1]$.

Detaljan pregled većeg broja matematičkih funkcija koje se mogu koristiti kao mere interesantnosti asocijativnih pravila, a dolaze iz različitih oblasti primenjene matematike, dat je u preglednim radovima (Tan et al. 2004; Geng & Hamilton, 2006; Hilderman & Hamilton, 2013).

4.2 Uticaj mera interesantnosti na kvalitet otkrivenog znanja

Kao što je izloženo u prethodnim poglavljima, i pored velikog broja metoda koje pomažu analitičarima podataka pri odabiru potencijalno korisnih asocijativnih pravila, ova oblast istraživanja ostaje aktuelna, a nijedna metoda ne daje dovoljno dobre rezultate.

U slučaju asocijativnih pravila o ponašanju posetioca web sajtova, gde postoji povezanost web objekata kroz hiperlink strukturu web sajta, dodatno se pogoršava postojeći problem generisanja prevelikog broja asocijatvnih pravila. Ogroman je broj pravila sa visokim vrednostima statističkih mera interesantnosti, a koja zapravo ne reflektuju interesovanja posetioca za sadržaj web objekata, već su rezultat povezanosti web objekata hiperlinkovima, koje su posetioci jednostavno primorani da prate kako bi došli do željenih informacija (Cercone & An, 2002; Cooley, 2003; Dimitrijević & Bošnjak, 2014; Kazienko & Pilarczyk, 2008; Liu, 2007; Padmanabhan & Tuzhilin, 2002; Sahar, 2010; Zaki, 2004).

U cilju ublažavanja ovog problema, u ovom poglavlju predlažemo primenu modifikovanih metoda za eliminaciju neinteresantnih asocijativnih pravila. Pored toga, u poglavlju 4.3. predlažemo primenu modifikovanih mera interesantnosti asocijativnih pravila o korišćenju web sajtova. Primena predloženih metoda ima cilj da smanji veličinu skupa otkrivenih asocijativnih pravila, i da pomogne da se preostala asocijativna pravila kvalitetnije rangiraju.

U nastavku koristimo terminologiju koja se odnosi specifično na asocijativna pravila o korišćenju web sajtova. Napominjemo da se neke od predloženih metoda mogu

uopštiti i primeniti i na druge vrste transakcionih baza podataka, što ostavljamo van domena ove disertacije.

4.2.1 Definicija statističke Z-score mere asocijativnog pravila

Korišćenjem statističkih testova nezavisnosti kao što je χ^2 test mogu se otkriti i eliminisati asocijativna pravila koja povezuju statistički nezavisne skupove atributa (Liu, Hsu & Ma, 1999). Jedna od statističkih mera korišćena u okviru algoritma za otkrivanje statistički opravdanih asocijativnih pravila je Z-score mera (Hämäläinen, 2010; Dimitrijević & Bošnjak, 2014).

U okviru ovog istraživanja predlažemo korišćenje Z-score mere pri otkrivanju i eliminaciji statistički očekivanih, neinteresantnih asocijativnih pravila, i to na dva različita načina. U prvom slučaju koristimo je za otkrivanje i eliminaciju asocijativnih pravila koja povezuju skupove web objekata koji su statistički nezavisni u skupu svih sesija na način predložen u (Hämäläinen, 2010). U drugom slučaju, predlažemo njeno korišćenje za otkrivanje i eliminaciju asocijativnih pravila koja su statistički očekivana u odnosu na opštija pravila, što je izloženo u narednim poglavljima.

U ovom poglavlju dajemo osnovne definicije i notaciju vezane za statističku Z-score meru kada se ona primenjuje na asocijativna pravila o korišćenju web sajtova. U nastavku teksta koristimo ovu notaciju, kao i terminologiju prilagođenu otkrivanju asocijativnih pravila u domenu web mining-a.

Osnovne definicije i notacija

Neka je I skup svih web objekata datog sajta, i neka je S skup sesija $\forall s \in S, s \subseteq I$.

Neka je S^X skup svih sesija koje sadrže skup web objekata $X \subseteq I$.

Neka su D^X, odnosno D^Y slučajni događaji koji predstavljaju pojavljivanje skupa X, odnosno skupa Y u nekoj sesiji $s \in S$, pri čemu su X i Y skupovi web objekata $X, Y \subseteq I$. Neka je $D^{X \cup Y}$ slučajni događaj koji predstavlja pojavljivanje skupa $X \cup Y$ u nekoj sesiji $s \in S$.

Pri ovoj notaciji napominjemo da važi:

$D^{X \cup Y} = D^X \cap D^Y$, kao i $S^{X \cup Y} = S^X \cap S^Y$.

Ako su D^X i D^Y nezavisni slučajni događaji u skupu sesija S, tada je verovatnoća pojavljivanja slučajnog događaja $D^{X\cup Y}$ data kao $p = P(X)P(Y)$, pri čemu je $P(X) = \frac{|S^X|}{n}$, $P(Y) = \frac{|S^Y|}{n}$, $n = |S|$.

Neka je M slučajna promenljiva koja uzima vrednost očekivanog broja pojavljivanja slučajnog događaja $D^{X\cup Y}$ u skupu sesija S. M ima binomnu raspodelu sa srednjom vrednosti $\mu = np$, varijansom $\sigma^2 = np(1-p)$ i standardnom devijacijom $\sigma = \sqrt{np(1-p)}$.

Neka je C stvarni broj pojavljivanja slučajnog događaja $D^{X\cup Y}$ u skupu sesija S.

Definicija 4.1

Z-score kao *mera statističke zavisnosti skupova web objekata* X i Y u skupu sesija S definiše se kao:

$$Z^S(X,Y) = \frac{C-\mu}{\sigma}$$

Z-score označava za koliko standardnih devijacija se izmereni broj zajedničkog pojavljivanja skupova X i Y u skupu sesija S razlikuje od njihovog očekivanog broja zajedničkog pojavljivanja.

Definicija 4.2

Z-score kao *mera interesantnosti asocijativnog pravila* $X \to Y$, $X, Y \subseteq I$ definiše se kao Z-score mera statističke zavisnosti skupova X i Y u skupu sesija S:

$$Z^S(X \to Y) = Z^S(X,Y)$$

Lema 4.1

Z-score je simetričan u odnosu na levu i desnu stranu asocijativnog pravila, odnosno važi:

$$Z^S(X \to Y) = Z^S(Y \to X) = Z^S(X,Y)$$

Dokaz: Sledi direktno iz definicije 4.2.

Z-score vrednost u okolini nule ukazuje na statističku nezavisnost skupova web objekata X i Y. Shodno tome, asocijativno pravilo čija je Z-score vrednost u okolini

nule se može eliminisati iz skupa svih pravila kao neinteresantno, odnosno statistički očekivano.

U radu (Hämäläinen, 2010) vršena je eliminacija statistički očekivanih pravila, čiji Z-score ne prelazi zadati minimalni Z-score prag. U okviru ovog istraživanja na sličan način vršimo eliminaciju statistički očekivanih asocijativnih pravila, kao jedan od prvih koraka prečišćavanja skupa otkrivenih asocijativnih pravila o korišćenju web sajta. Potom primenjujemo i druge metode za eliminaciju pravila koja preostaju u skupu otkrivenih pravila, ali su ipak neinteresantna analitičarima podataka, što je opisano u narednim poglavljima.

4.2.2 Relacija opšte/specifično asocijativno pravilo

U ovom poglavlju dajemo formalnu definiciju pojma opšte/specifično asocijativno pravilo, na kome su zasnovane metode predložene za prečišćavanje skupa asocijativnih pravila o korišćenju web sajtova opisane u narednim poglavljima. Predložene metode su primenljive na skupove podataka u kojima postoji visok stepen korelacije između atributa koji čine asocijativna pravila, kao što je to slučaj sa asocijativnim pravilima o korišćenju web sajtova.

Definicija 4.3

Ako su skupovi web objekata $X, X' \subseteq I$ takvi da $S^X \subseteq S^{X'}$ onda je X' opštiji skup web objekata u odnosu na specifičniji skup web objekata X.

Definicija 4.4

Neka su $X, X', Y \subseteq I$ skupovi web objekata, pri čemu važi $X' \cap Y = \emptyset$ i $X \cap Y = \emptyset$. Neka je X' opštiji skup web objekata u odnosu na specifičniji skup web objekata X. Tada je asocijativno pravilo $X' \rightarrow Y$ opštije u odnosu na specifičnije asocijativno pravilo $X \rightarrow Y$, dok je asocijativno pravilo $Y \rightarrow X'$ opštije u odnosu na specifičnije asocijativno pravilo $Y \rightarrow X$.

4.2.3 Lokalni Z-score kao mera statističke očekivanosti asocijativnog pravila u odnosu na opštije asocijativno pravilo

U ovom poglavlju dajemo definiciju lokalne Z-score mere asocijativnog pravila, koju koristimo pri formulisanju uslova za eliminaciju asocijativnih pravila koja su statistički očekivana u odnosu na opštija asocijativna pravila.

4.2.3.1 Opšte definicije i notacija

Neka X' označava opštiji skup web objekata u odnosu na specifičniji skup web objekata X. Neka $S^{X'}$ označava skup svih sesija koje sadrže skup web objekata X', pri čemu je $|S^{X'}| = \text{n'}$. Neka je D' slučajni događaj koji predstavlja pojavljivanje skupa $X \cup Y$ u nekoj sesiji $s \in S^{X'}$.

Ako su D^X i D^Y nezavisni slučajni događaji u skupu sesija $S^{X'}$ (odnosno kada je dat događaj $D^{X'}$), tada je verovatnoća pojavljivanja slučajnog događaja D' u skupu sesija $S^{X'}$ data kao:

$$p' = P(D') = P\left(D^{X\cup Y}\middle|D^{X'}\right) = P\left(D^X \cap D^Y\middle|D^{X'}\right) = P\left(D^X\middle|D^{X'}\right) P\left(D^Y\middle|D^{X'}\right)$$

Ako dalje označimo $P'(X) = P\left(D^X\middle|D^{X'}\right)$, $P'(Y) = P\left(D^Y\middle|D^{X'}\right)$, tada važi

$$P'(X) = \frac{\left|S^{X\cup X'}\right|}{n'} = \frac{\left|S^X \cap S^{X'}\right|}{n'}\ ,\ \ P'(Y) = \frac{\left|S^{Y\cup X'}\right|}{n'} = \frac{\left|S^Y \cap S^{X'}\right|}{n'}.$$

Pri tome važi $P'(X) = \frac{|S^X|}{n'}$ jer je X' opštiji skup web objekata u odnosu na specifičniji skup web objekata X (definicija 4.3).

Neka je M' slučajna promenljiva koja uzima vrednost broja pojavljivanja slučajnog događaja D' u skupu sesija $S^{X'}$. Tada M' ima binomnu raspodelu sa srednjom vrednosti $\mu' = n'p'$, varijansom $\sigma'^2 = n'p'(1-p')$ i standardnom devijacijom $\sigma' = \sqrt{n'p'(1-p')}$.

Neka je C' izmereni broj pojavljivanja slučajnog događaja D' u skupu sesija $S^{X'}$.

Neka su su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \cap Y = \emptyset$ i $X \cap Y = \emptyset$, i neka je X'opštiji skup web objekata u odnosu na specifičniji skup web objekata X.

Koristeći navedenu notaciju, u nastavku definišemo lokalni Z-score kao meru statističke zavisnosti između dva skupa web objekata (definicija 4.5). Potom definišemo lokalni Z-score kao meru statističke očekivanosti asocijativnog pravila u odnosu na opštije asocijativno pravilo (definicije 4.6a i 4.6b).

Definicija 4.5

Lokalni Z-score kao meru statističke zavisnosti skupova web objekata X i Y u skupu sesija $S^{X'}$ definišemo kao:

$$Z^{X'}(X,Y) = \frac{C' - \mu'}{\sigma'}, \qquad pri\ čemu\ je$$

$$C' = \left|S^{X \cup Y \cup X'}\right| = \left|S^X \cap S^Y \cap S^{X'}\right| = |S^X \cap S^Y|$$

$$\mu' = |S^{X'}| \cdot p'$$

$$p' = \frac{|S^{X \cup X'}| \cdot |S^{Y \cup X'}|}{|S^{X'}| \cdot |S^{X'}|} = \frac{|S^X \cap S^{X'}| \cdot |S^Y \cap S^{X'}|}{|S^{X'}| \cdot |S^{X'}|} = \frac{|S^X| \cdot |S^Y \cap S^{X'}|}{|S^{X'}| \cdot |S^{X'}|}$$

$$\sigma' = \sqrt{|S^{X'}| \cdot p'(1-p')}$$

Lokalni Z-score predstavlja meru statističke zavisnosti skupova web objekata X i Y, ali u okviru skupa sesija $S^{X'}$, pri čemu je X' opštiji skup web objekata u odnosu na specifičniji skup web objekata X.

Preciznije, lokalni Z-score označava za koliko standardnih devijacija se izmereni broj zajedničkog pojavljivanja skupova web objekata X i Y razlikuje od njihovog očekivanog broja zajedničkog pojavljivanja u skupu sesija $S^{X'}$.

Kada je vrednost $Z^{X'}(X,Y)$ u okolini nule, skupovi web objekata X i Y su statistički nezavisni u skupu sesija $S^{X'}$.

Definicija 4.6a

Lokalni Z-score kao meru statističke očekivanosti asocijativnog pravila $X \rightarrow Y$ u odnosu na opštije asocijativno pravilo $X' \rightarrow Y$ definišemo kao lokalni Z-score skupova web objekata X i Y:

$$Z^{X'}(X \rightarrow Y) = Z^{X'}(X,Y)$$

Definicija 4.6b

Lokalni Z-score kao meru statističke očekivanosti asocijativnog pravila $Y \rightarrow X$ u odnosu na opštije asocijativno pravilo $Y \rightarrow X'$ definišemo kao lokalni Z-score skupova web objekata X i Y:

$$Z^{X'}(Y \rightarrow X) = Z^{X'}(X, Y)$$

Lokalni Z-score asocijativnog pravila u odnosu na opštije asocijativno pravilo je očigledno simetričan u odnosu na levu i desnu stranu asocijativnog pravila.

4.2.3.2 Uslov za eliminaciju statistički očekivanih asocijativnih pravila u odnosu na opštija asocijativna pravila

Definicija 4.7a

Neka je X' opštiji skup web objekata u odnosu na specifičniji skup web objekata X. Neka u skupu svih asocijativnih pravila postoje asocijativna pravila $X \rightarrow Y$ i $X' \rightarrow Y$. Asocijativno pravilo $X \rightarrow Y$ smatra se *statistički očekivano u odnosu na opštije asocijativno pravilo* $X' \rightarrow Y$ ako važi $Z^{X'}(X, Y) < minZ$.

Definicija 4.7b

Neka je X' opštiji skup web objekata u odnosu na specifičniji skup web objekata X. Neka u skupu svih asocijativnih pravila postoje asocijativna pravila $Y \rightarrow X$ i $Y \rightarrow X'$. Asocijativno pravilo $Y \rightarrow X$ smatra se *statistički očekivano u odnosu na opštije asocijativno pravilo* $Y \rightarrow X'$ ako važi $Z^{X'}(X, Y) < minZ$.

Asocijativno pravilo koje je *statistički očekivano u odnosu na opštije asocijativno pravilo* koje takođe postoji u skupu svih otkrivenih pravila može se eliminisati iz skupa otkrivenih pravila kao neinteresantno. Pri tome, uslov za eliminaciju dozvoljava variranje stvarnog broja zajedničkog pojavljivanja skupova X i Y u odnosu na njihov očekivani broj zajedničkog pojavljivanja u skupu sesija $S^{X'}$, ograničeno parametrom $minZ$, kojim se definiše maksimalni broj standardnih devijacija za ovo variranje.

U okviru eksperimentalnog istraživanja (poglavlje 6) vršena je eliminacija statistički očekivanih asocijativnih pravila u odnosu na opštija asocijativna pravila, koja su otkrivena u stvarnim skupovima podataka. Pri tome parametar $minZ$ uzima vrednosti

iz uobičajenog intervala od 2 do 4, analogno prethodnom istraživanju (Hämäläinen, 2010).

Opšta definicija statistički očekivanog asocijativnog pravila u odnosu na opštije asocijativno pravilo data definicijama 4.7a i 4.7b obuhvata dva različita slučaja:

1. *Asocijativna pravila statistički očekivana u odnosu na opštija asocijativna pravila manje dužine* – slučaj razmatran u poglavlju 4.2.4
2. *Asocijativna pravila statistički očekivana u odnosu na opštija asocijativna pravila jednake dužine, koja postoje u prisustvu konceptne hijerarhije web objekata* – slučaj razmatran u poglavlju 4.2.6.

Predloženi metod eliminacije statistički očekivanih pravila u prisustvu opštijih pravila je upoređen sa dva postojeća metoda kojima se eliminišu trivijalna, odnosno očekivana asocijativna pravila u poglavlju 4.2.7. Motivacija za eliminaciju asocijativnih pravila primenom predložene metode potkrepljena je primerima asocijativnih pravila iz stvarnih skupova podataka u poglavlju 4.2.8.

4.2.4 Eliminacija statistički očekivanih asocijativnih pravila u odnosu na opštija asocijativna pravila manje dužine

U ovom poglavlju razmatramo eliminaciju dužih asocijativnih pravila, ukoliko su ona statistički očekivana u odnosu na kraća i opštija asocijativna pravila.

Definicija 4.8

Neka su su $X, Y \subseteq I$ skupovi web objekata pri čemu važi $X \cap Y = \emptyset$. Dužinu asocijativnog pravila $X \rightarrow Y$ definišemo kao kardinalitet skupa $X \cup Y$.

Lema 4.2a

Neka su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \subset X,\ X \cap Y = \emptyset$. Tada je asocijativno pravilo $X' \rightarrow Y$ kraće od asocijativnog pravila $X \rightarrow Y$.

Dokaz:

$$X' \subset X \Rightarrow |X' \cup Y| < |X \cup Y| \Rightarrow X' \rightarrow Y \; je \; kraće \; od \; X \rightarrow Y \; (\text{def. } 4.8)$$

Lema 4.2b

Neka su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \subset X,\ X \cap Y = \emptyset$. Tada je asocijativno pravilo $Y \to X'$ kraće od asocijativnog pravila $Y \to X$.

Dokaz: analogno lemi 4.2a.

Lema 4.3a

Neka su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \subset X,\ X \cap Y = \emptyset$. Tada je asocijativno pravilo $X' \to Y$ opštije u odnosu na asocijativno pravilo $X \to Y$.

Dokaz:

$X' \subset X \Rightarrow (\forall s \in S)(X \subseteq \mathrm{s} \Rightarrow X' \subseteq \mathrm{s}) \Rightarrow S^X \subseteq S^{X'} \Rightarrow$

$\Rightarrow X' \to Y$ je opštije asocijativno pravilo u odnosu na $X \to Y$ $(def.4.4)$

Asocijativno pravilo $X \to Y$ za koje postoji kraće asocijativno pravilo oblika $X' \to Y$, takvo da je $X' \subset X$ može se eliminisati iz skupa svih asocijativnih pravila ako je $X \to Y$ statistički očekivano u odnosu na $X' \to Y$ (definicija 4.7a).

Lema 4.3b

Neka su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \subset X,\ X \cap Y = \emptyset$. Tada je asocijativno pravilo $Y \to X'$ opštije u odnosu na asocijativno pravilo $Y \to X$.

Dokaz: analogno lemi 4.3a.

Asocijativno pravilo $Y \to X$ za koje postoji kraće asocijativno pravilo oblika $Y \to X'$, takvo da je $X' \subset X$, može se eliminisati iz skupa svih asocijativnih pravila ako je $Y \to X$ statistički očekivano u odnosu na $Y \to X'$ (definicija 4.7b).

U okviru softverskog sistema (poglavlje 5) implementiran je algoritam za eliminaciju asocijativnih pravila statistički očekivanih u odnosu na kraća i opštija pravila. Eliminacija se vrši prema nivoima dužine asocijativnih pravila, počevši od najdužih, završivši sa eliminacijom asocijativnih pravila dužine 3. Kratka asocijativna pravila dužine 2 nije moguće eliminisati ovom metodom. Rezultati eksperimentalnog istraživanja gde je ova eliminacija primenjena na stvarne skupove otkrivenih asocijativnih pravila data je u poglavlju 6.2.3.

4.2.5 Konceptna hijerarhija web objekata

U ovom poglavlju definišemo konceptnu hijerarhiju web objekata, koja se može generisati na osnovu asocijativnih pravila o korišćenju web sajtova čija je vrednost confidence mere približno jednaka 1. Pri tome, vrednost confidence mere asocijativnog pravila $X \to Y$, gde su $X, Y \subseteq I$ skupovi web objekata je u skladu sa opštom definicijom confidence mere (definicija 1.6) data kao:

$$conf(X \to Y) = \frac{|S^{X \cup Y}|}{|S^X|} = \frac{|S^X \cap S^Y|}{|S^X|}$$

Kao što potvrđuju brojna eksperimentalna istraživanja, u skupu asocijativnih pravila o korišćenju web sajtova postoji znatan broj pravila sa vrednostima confidence mere u okolini 1 (Huang, 2007; Huang & An, 2002; Dimitrijević & Bošnjak, 2010). Ovo su najčešće kratka asocijativna pravila oblika $x \to x'$, pri čemu su x i x' web stranice.

Ovakva pravila se obično javljaju u jednom od dva slučaja:

1) Na web stranici x' se nalazi direktan hiperlink ka web stranici x, i to je jedini hiperlink ka web stranici x na web sajtu.
2) Na web sajtu postoji hijerarhijski meni na kome je web stranica x' direktno nadređena stavka za web stranicu x, i jedini način da se poseti web stranica x je ako se prethodno poseti web stranica x'.

U navedenim slučajevima kada postoji takva hiperlink struktura web sajta gde je web stranica x hijerarhijski direktno podređena web stranici x', i jedini način da se poseti web stranica x je ako se prethodno poseti web stranica x', možemo smatrati da je web stranica x potkoncept u odnosu na web stranicu x', koja je njen direktni natkoncept.

Pored toga, asocijativna pravila sa maksimalnom confidence vrednosti se javljaju i u slučaju kada neki od irelevantnih web objekata ugrađenih u neku web stranicu greškom nije eliminisan tokom pretprocesiranja web log podataka. Tada postoji asocijativno pravilo $o \to x'$, $conf(o \to x') = 1$, gde je x' web stranica koja sadrži ugrađem web objekat o. U tom slučaju će metodom eliminacije asocijativnih pravila datom u poglavlju 4.2.6 većina drugih asocijativnih pravila koja sadrže web objekat o biti eliminisana kao statistički očekivana u odnosu na odgovarajuća asocijativna pravila koja sadrže web stranicu x', i neće opterećivati skup otkrivenih asocijativnih pravila.

Definicija 4.9

Ako su $x, x' \in I$ web objekti i postoji asocijativno pravilo $x \rightarrow x'$, $conf(x \rightarrow x') = 1$, tada je x' *natkoncept* u odnosu na x.

Koristeći sva pravila $x \rightarrow x'$ čija je vrednost confidence mere jednaka 1, može se ekstrahovati deo konceptne hijerarhije web sajta. Ovako generisana konceptna hijerarhija ne može obuhvatiti sve veze tipa potkoncept/natkoncept između web stranica sajta, za čega bi idealno bilo konsultovati ekspertsko znanje, ali predstavlja aproksimaciju dela konceptne hijerarhije web sajta.

Prednosti korišćenja ovako definisane konceptne hijerarhije su:

- Zasniva se na strogo definisanim pravilima otkrivenim analizom samih podataka.
- Ne zahteva se korišćenje ekspertskog znanja.
- Ne ograničava se na web sajtove gde su srodne web stranice organizovane ili obeležene na određeni način.
- Ne zahteva se parsiranje hiperlink strukture web sajta, što može biti kompleksan zadatak.

Pri ekstrahovanju ovakve konceptne hijerarhije treba uzeti u obzir da postoji mogućnost da vrednost confidence mere pravila $x \rightarrow x'$ bude visoka, ali ne i jednaka 1, a da pri tome ipak postoji veza potkoncept/natkoncept između web stranica x i x' kao u gore navedenim slučajevima 1) i 2). Na primer, moguće je da jedan deo sesija koje sadrže web stranicu x ne sadrži web stranicu x' samo zato što jedan deo posetioca web sajta čuva direktan „bookmark“ na web stranicu x, ili je pronalazi direktno preko web pretraživača, bez prethodne posete početnoj stranici web sajta, i bez posete web stranici x'. Iz tog razloga uvodimo definiciju 4.9a:

Definicija 4.9a

Ako su $x, x' \in I$ web objekti i postoji asocijativno pravilo $x \rightarrow x'$, $conf(x \rightarrow x') = 1 - \varepsilon$, gde je ε nenegativna mala vrednost, tada je x' *natkoncept* u odnosu na x .

U skladu sa gore navedenim, prilikom ekstrahovanja konceptne hijerarhije u okviru eksperimentalnog istraživanja (poglavlje 6) koristimo minimalni prag vrednosti confidence mere pravila $x \rightarrow x'$, koji je nešto manji od 1.

4.2.6 Eliminacija statistički očekivanih asocijativnih pravila u odnosu na opštija asocijativna pravila jednake dužine

U ovom poglavlju razmatramo eliminaciju asocijativnih pravila koja su statistički očekivana u odnosu na opštija asocijativna pravila jednake dužine, u prisustvu relacije „natkoncept“.

Napomena: U nastavku koristimo Ax kao skraćenu oznaku za $A \cup \{x\}$, pri čemu $x \in I$, $A \subseteq I$. Takođe koristimo x kao skraćenu oznaku za $\{x\}$.

Lema 4.4a

Neka su $x, x' \in I$ web objekti takvi da postoji asocijativno pravilo $x \to x'$, $conf(x \to x') = 1$. Neka su $A, B \subseteq I$ skupovi web objekata takvi da važi $B \neq \emptyset$, $A \cap B = \emptyset$, $x, x' \notin A \cup B$. Tada je asocijativno pravilo $Ax' \to B$ opštije u odnosu na asocijativno pravilo $Ax \to B$.

Dokaz:

$$conf(x \to x') = 1 \Rightarrow \frac{|S^{x \cup x'}|}{|S^x|} = 1 \Rightarrow |S^{x \cup x'}| = |S^x| \Rightarrow |S^x| \cap |S^{x'}| = |S^x| \Rightarrow$$

$$S^x \subseteq S^{x'} \Rightarrow S^{Ax} \subseteq S^{Ax'} \Rightarrow Ax' \to B \; je \; opštije \; a.p. od \; Ax \to B \; (\text{def } 4.4)$$

Lema 4.4b

Neka su $x, x' \in I$ web objekti takvi da postoji asocijativno pravilo $x \to x'$, $conf(x \to x') = 1$. Neka su $A, B \subseteq I$ skupovi web objekata takvi da važi $A \neq \emptyset$, $A \cap B = \emptyset$, $x, x' \notin A \cup B$. Tada je asocijativno pravilo $A \to x'B$ opštije u odnosu na specifičnije asocijativno pravilo $A \to xB$.

Dokaz: Analogno dokazu leme 4.4a.

Dakle, asocijativno pravilo oblika $Ax \to B$ može se eliminisati iz skupa svih asocijativnih pravila ako je ono statistički očekivano u odnosu na neko asocijativno pravilo oblika $Ax' \to B$ (definicija 4.7a). Analogno tome, asocijativno pravilo oblika $A \to xB$ može se eliminisati iz skupa svih asocijativnih pravila ako je ono statistički očekivano u odnosu na neko asocijativno pravilo oblika $A \to x'B$ (definicija 4.7b).

Značaj ove metode eliminacije je i u tome što je ona primenljiva kako na duga, tako i na kratka asocijativna pravila (pravila koja imaju samo po jednu web stranicu sa obe

strane), koja su zbog svoje jednostavnosti najčešće korišćena od strane analitičara podataka (Kazienko, 2009; Schafer, Konstan & Riedl, 2001).

Eliminacija otkrivenih asocijativnih pravila prema lemama 4.4a i 4.4b implementirana je u okviru softverskog sistema (poglavlje 5). U eksperimentalnom istraživanju ova metoda eliminacije je primenjena na asocijativna pravila koja preostaju u skupu otkrivenih pravila posle primene svih ostalih predloženih metoda eliminacije (poglavlje 6.2.4).

4.2.6.1 Specijalan slučaj – klaster pravila

Specijalan slučaj relacije „natkoncept" je situacija u kojoj postoji ciklična struktura, gde su dve ili više web stranice međusobno u relaciji potkoncept/natkoncept, i to u oba pravca.

Lema 4.5a

Neka su $x, x' \in I$ web objekti takvi da postoje asocijativna pravila $x \to x'$, $x' \to x$, $conf(x \to x') = 1$, $conf(x' \to x) = 1$. Neka su $A, B \subseteq I$ skupovi web objekata takvi da važi $A \neq \emptyset$, $A \cap B = \emptyset$, $x, x' \notin A \cup B$. Tada su vrednosti statističkih mera interesantnosti asocijativnih pravila $Ax' \to B$ i $Ax \to B$ jednake.

Dokaz:

$conf(x \to x') = 1 \ \wedge conf(x' \to x) = 1 \Rightarrow S^x = S^{x'} \Rightarrow S^{Ax} = S^{Ax'}$

Sledi da su vrednosti svih mera interesantnosti koje su bazirane na veličini skupova sesija koje sadrže levu i/ili desnu stranu asocijativnog pravila jednake za ova dva pravila.

Lema 4.5b

Neka su $x, x' \in I$ web objekti takvi da postoje asocijativna pravila $x \to x'$, $x' \to x$, $conf(x \to x') = 1$, $conf(x' \to x) = 1$. Neka su $A, B \subseteq I$ skupovi web objekata takvi da važi $A \neq \emptyset$, $A \cap B = \emptyset$, $x, x' \notin A \cup B$. Tada su vrednosti statističkih mera interesantnosti asocijativnih pravila $A \to x'B$ i $A \to xB$ jednake.

Dokaz: Analogno dokazu leme 4.5a.

Asocijativna pravila oblika $Ax' \to B$ i $Ax \to B$, odnosno oblika $A \to x'B$ i $A \to xB$ za koje važi lema 4.5a, odnosno lema 4.5b, nazivamo *klaster pravilima*.

U konkretnim skupovima podataka za koje je vršeno eksperimentalno istraživanje (poglavlje 6) postoje samo slučajevi gde su dve web stranice međusobno u relaciji potkoncept/natkoncept. U opštem slučaju, gde su više „klaster“ web stranica međusobno u relaciji „potkoncept/natkoncept“, postojao bi čitav „klaster“ asocijativnih pravila sa jednakim vrednostima statističkih mera interesantnosti.

Očigledno, klaster pravila su suvišna u skupu svih pravila. Dovoljno je zadržati jedno od pravila koje čine klaster, koje je tada predstavnik klastera, a eliminisati ostala pravila iz klastera kao trivijalna.

Obzirom da su klaster pravila trivijalna i da je njihova eliminacija računski jednostavna, u okviru eksperimentalnog istraživanja ova pravila se eliminišu u prvom koraku prećišćavanja skupa otkrivenih asocijativnih pravila (poglavlje 6.2.1). Potom se primenjuju ostale kompleksnije metode eliminacije statistički očekivanih pravila, na skup asocijativnih pravila prethodno prečišćen eliminisanjem klaster pravila.

4.2.7 Poređenje sa prethodnim istraživanjima

Otkrivanje generalizovanih asocijativnih pravila

Klasična metoda za otkrivanje generalizovanih asocijativnih pravila prvi put je predložena u radu (Srikant & Agrawal, 1997), a potom korišćena u brojnim istraživanjima (Hong, Lin & Wang, 2003; Kotsiantis & Kanellopoulos, 2006; Tseng & Lin, 2007; Yang, 2005). Metoda se bazira na očekivanoj vrednosti confidence mere specifičnog pravila u odnosu na kraća, opštija pravila. Prvi korak pri generalizaciji asocijativnih pravila je generisanje novih asocijativnih pravila u kojima su atributi zamenjeni njihovim natkonceptima iz zadate konceptne hijerarhije, unapred definisane od strane eksperta. Potom se vrši eliminacija onih specifičnijih pravila čija je confidence vrednost očekivana u odnosu na confidence vrednost opštijih pravila.

U nastavku ovog poglavlja povezujemo pomenutu metodu generalizacije asocijativnih pravila sa relacijom „opštije/specifičnije“ asocijativno pravilo. Potom dajemo paralelu između očekivane confidence mere generalizovanog asocijativnog pravila i lokalne Z-score mere definisane u poglavlju 4.2.3.

U okviru pomenute klasične metode generalizacije, asocijativno pravilo R se eliminiše iz skupa svih pravila ukoliko važi $conf(R) < M \cdot expconf(R)$, pri čemu je M zadati koeficijent, dok $expconf(R)$ označava očekivani confidence pravila R u odnosu na neko opštije pravilo R'.

U zavisnosti od toga da li se skup atributa koji je opštiji/specifičniji nalazi sa leve ili desne strane pravila R, odnosno pravila R', Srikant & Agrawal (1997) razlikuju dva slučaja:

1. Opštiji koncept se nalazi sa leve strane pravila, odnosno pravilo R je oblika $X \to Y$, a pravilo R' je oblika $X' \to Y$, pri čemu je X' opštiji skup atributa u odnosu na skup atributa X. Tada je očekivani confidence definisan kao:
$$expconf(X \to Y) = conf(X' \to Y)$$
2. Opštiji koncept se nalazi sa desne strane pravila, odnosno pravilo R je oblika $Y \to X$, a pravilo R' *je* oblika $Y \to X'$, pri čemu je X' opštiji skup atributa u odnosu na skup atributa X. Tada je očekivani confidence definisan kao:
$$conf(Y \to X\ ) = conf(Y \to X') \cdot conf(X' \to X)$$

Lemama 4.6 i 4.7 pokazujemo da asocijativna pravila čiji je confidence jednak očekivanom u odnosu na neko opštije pravilo, imaju lokalni Z-score upravo jednak 0.

Lema 4.6

Neka su su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \cap Y = \emptyset$, $X \cap Y = \emptyset$ i $X' \to Y$ je opštije asocijativno pravilo u odnosu na $X \to Y$ (def. 4.4). Tada važi:
$Z^{X'}(X \to Y) = 0\ \ akko\ \ conf(X \to Y) = conf(X' \to Y)$.

Dokaz:

$S^X \subseteq S^{X'}(def.4.3;\ def.4.4) \Rightarrow S^{X \cup X'} = S^X$

$$Z^{X'}(X \to Y) = 0\ akko\ \left|S^{X \cup Y \cup X'}\right| = \frac{\left|S^{X \cup X'}\right| \cdot \left|S^{Y \cup X'}\right|}{|S^{X'}|}\quad (def.4.5)$$

$$akko\ \ |S^{X \cup Y}| = \frac{|S^X| \cdot \left|S^{Y \cup X'}\right|}{|S^{X'}|}$$

$$akko \ \frac{|S^{X\cup Y}|}{|S^X|} = \frac{|S^{X'\cup Y}|}{|S^{X'}|}$$

$$akko \ \ conf(X \to Y) = conf(X' \to Y).$$

Lema 4.7

Neka su su $X, X', Y \subseteq I$ skupovi web objekata takvi da važi $X' \cap Y = \emptyset, \ X \cap Y = \emptyset$ i $Y \to X'$ je opštije asocijativno pravilo u odnosu na $Y \to X$ (def. 4.4). Tada važi:

$Z^{X'}(Y \to X) = 0 \ \ akko \ \ conf(Y \to X) = conf(Y \to X') \cdot conf(X' \to X)$

Dokaz:

$S^X \subseteq S^{X'}(def.4.3;\ def.4.4) \Rightarrow S^{X\cup X'} = S^X$

$conf(Y \to X) = \ conf(Y \to X') \cdot conf(X' \to X)$

$$akko \ \frac{|S^{Y\cup X}|}{|S^Y|} = \frac{|S^{Y\cup X'}|}{|S^Y|} \cdot \frac{|S^{X\cup X'}|}{|S^{X'}|}$$

$$akko \ |S^{X\cup Y\cup X'}| = \frac{|S^{X\cup X'}| \cdot |S^{Y\cup X'}|}{|S^{X'}|}$$

$akko \ Z^{X'}(Y \to X) = 0 \ \ (def.4.5).$

Zaključujemo da će predloženim postupkom eliminacije statistički očekivanih pravila baziranim na lokalnoj Z-score meri svakako biti eliminisana pravila koja imaju tačno očekivanu vrednost confidence mere u odnosu na neko opštije pravilo. Međutim, smatramo da je u odnosu na metod korišćen u (Srikant & Agrawal, 1997), gde se odnos očekivane i stvarne vrednosti confidence mere poredi sa zadatom konstantnom M, statistički opravdanije koristiti lokalnu Z-score meru za utvrđivanje statističke očekivanosti asocijatnih pravila u odnosu na opštija pravila.

Eliminacija trivijalnih pravila

U radovima (Webb & Zhang, 2002) i (Huang & Webb, 2005) vrši se eliminacija takozvanih trivijalnih pravila. Za razliku od metode za otkrivanje generalizovanih asocijativnih pravila (Srikant & Agrawal, 1997), eliminacija trivijalnih pravila ne podrazumeva korišćenje bilo kakve konceptne ontologije.

Prevedeno na notaciju korišćenu u ovoj disertaciji, u radovima (Webb & Zhang, 2002) i (Huang & Webb, 2005) asocijativno pravilo oblika $X \rightarrow Y$ označeno je kao *trivijalno* ako postoji *kraće* asocijativno pravilo $X' \rightarrow Y$ takvo da važi $X' \subset X$ i $conf(X \rightarrow Y) = conf(X' \rightarrow Y)$. Lemom 4.6 dokazano je da su ovako definisana trivijalna pravila specijalan slučaj obuhvaćen opštim uslovom eliminacije asocijativnih pravila statistički očekivanih u odnosu na opštija asocijativna pravila (poglavlje 4.2.4).

Korišćenje konceptne ontologije

Brojna istraživanja predlažu korišćenje ontologije koncepata definisane od strane eksperta, pri čemu se pravi mapiranje između web stranica i opštijih koncepata. Na primer, ovakva ontologija se koristi za povećanje kvaliteta predikcije posete sledećoj web stranici u istraživanjima (Becker & Vanzin, 2010; Mabroukeh & Ezeife, 2009). Sličniji našem istraživanju je rad (Senkul & Salin, 2012), gde se predlaže korišćenje konceptne hijerarhije za povećanje kvaliteta otkrivenih asocijativnih pravila o ponašanju korisnika web sajtova. Ovaj metod se takođe oslanja na ekspertsku konceptnu hijerarhiju u cilju smanjenja broja otkrivenih pravila. Međutim, mana ovog pristupa je što se gubi nivo granularnosti informacije sadržane u otkrivenim asocijativnim pravilima. Umesto url web stranica, otkrivena asocijativna pravila sadrže samo podatke o konceptima kojima web stranice pripadaju (Senkul & Salin, 2012). U okviru istraživanja (Lee, Lo & Fu, 2011) hijerarhijska struktura foldera aproksimira konceptnu hijerarhiju, na kojoj se bazira predloženi metod za povećanje tačnosti predikcije sledeće posećene stranice. Mana ovog pristupa je njegova ograničenost na web sajtove organizovane u okviru hijerarhijske strukture foldera, pri čemu su nazivi foldera vidljivi u samoj url adresi web stranica.

Jedna od prednosti metode predložene u okviru ove disertacije je što se ona oslanja na statistički validnu meru očekivanosti asocijativnih pravila u odnosu na opštija asocijativna pravila. Pored toga, metoda nije ograničena na neku određenu vrstu web sajtova, niti se zahteva ekspertsko znanje pri definisanju konceptne hijerarhije. Iako se ne menja granularnost otkrivenih asocijativnih pravila, koja i dalje sadrže informaciju o url adresi zahtevanih web stranica, veličina skupa otkrivenih pravila se znatno smanjuje, čineći ga upotrebljivijim za analitičare podataka.

4.2.8 Primeri iz stvarnog skupa podataka

U ovom poglavlju definiciju uslova za eliminaciju statistički očekivanih pravila u prisustvu opštijih pravila potkrepljujemo primerima iz stvarnog skupa podataka o korišćenju web sajta Fakulteta organizacionih nauka u Beogradu.

Br	Levo	Desno	Conf	Lift	Z	L-Z
1	/istrazivanjeirazvoj/index.html /vesti/index.html	/ofakultetu/index.html	0.33	5.7	10.3	< 2.0
1a	/istrazivanjeirazvoj/index.html	/ofakultetu/index.html	0.32	5.6	15.7	---
2	/vesti/Konkursdrstudije2013.pdf	/postdiplomske/index.html	0.37	4.2	16.2	< 2.0
2a	/postdiplomske/doktorske/index.html	/postdiplomske/index.html	0.34	3.9	22.6	---
3	/ofakultetu/index.html	/osnovnestudije/isit/index.html	0.13	2.0	9.6	< 2.0
3a	/ofakultetu/index.html	/osnovnestudije/index.html	0.37	2.1	17.4	---
3b	/osnovnestudije/index.html	/osnovnestudije/isit/index.html	0.33	5.2	70.8	---

Tabela 4.3. Primeri statistički očekivanih pravila u prisustvu opštijih pravila

U Tabeli 4.3 dato je sedam pravila otkrivenih u ovom skupu podataka. Kolona „Levo“ sadrži skup atributa na levoj strani pravila, a kolona „Desno“ sadrži skup atributa na desnoj strani pravila. Pravilo sa rednim brojem 1 sadrži dva atributa sa leve strane, dok su ostala pravila kratka, odnosno sadrže po jedan atribut sa obe strane.

Kolone „Conf“, „Lift“ i „Z“ sadrže vrednosti confidence, lift i Z-score mere interesantnosti za svako pravilo, posmatrano u skupu svih web sesija. Ove vrednosti su povišene za sva pravila, što znači da su sva pravila označena kao potencijalno interesantna analitičaru podataka ako se uzmu u obzir samo ove statističke mere interesantnosti.

Kolona „L-Z“ sadrži vrednost lokalne Z-score mere u odnosu na opštije asocijativno pravilo. Za pravila 1a, 2a, 3a i 3b ne postoji opštije pravilo u skupu svih pravila, pa je vrednost lokalne Z-score mere nedefinisana. Za pravila sa rednim brojevima 1, 2 i 3 vrednost lokalne Z-score mere u odnosu na opštije pravilo je niska (manja od 2.0). Ova pravila su označena za eliminaciju kao statistički očekivana u prisustvu opštijih pravila, sa dozvoljenim pragom devijacije $minZ = 2.0$.

Pravilo sa rednim brojem 1 je statistički očekivano u odnosu na opštije i kraće pravilo 1a (slučaj razmatran u poglavlju 4.2.4). Primetimo da je pri tome confidence vrednost pravila 1 približno jednaka confidence vrednosti pravila 1a (kolona „Conf“), što je u skladu sa očekivanom confidence vrednosti (Srikant & Agrawal, 1997).

Pravilo sa rednim brojem 2 je statistički očekivano u odnosu na opštije pravilo 2a, pri čemu je *{/postdiplomske/doktorske/index.html}* opštiji koncept u odnosu na *{/vesti/Konkursdrstudije2013.pdf}* (slučaj razmatran u poglavlju 4.2.6). Primetimo da je pri tome confidence vrednost pravila 2 približno jednaka confidence vrednosti pravila 2a (kolona „Conf“), što je takođe u skladu sa očekivanom confidence vrednosti.

Pravilo sa rednim brojem 3 je statistički očekivano u odnosu na opštije pravilo 3a, pri čemu je *{/osnovnestudije/index.html}* opštiji koncept u odnosu na *{/osnovnestudije-isit/index.html}* (slučaj razmatran u poglavlju 4.2.6). Pri tome, confidence vrednost pravila 3 je približno jednaka proizvodu confidence vrednosti pravila 3a i pravila 3b (kolona „Conf“), što je takođe u skladu sa očekivanom confidence vrednosti.

U skladu sa definisanim uslovima za eliminaciju specifičnijih pravila u prisustvu opštijih pravila (definicije 4.7a i 4.7b), pravila 1, 2 i 3 biće eliminisana iz skupa svih pravila. Ova pravila su neinteresantna i zbunjuju analitičara podataka jer su trivijalna, odnosno očekivana u odnosu na opštija pravila 1a, 2a, odnosno 3a. Pri tome, pravila 1 i 2 bila bi visoko rangirana primenom standardnih mera interesantnosti (kolone Conf, Lift i Z), čime bi dodatno zbunjivala analitičare podataka.

4.3 Mere interesantnosti u web mining-u

Nakon eliminacije neinteresantnih asocijativnih pravila primenom metoda predloženih u prethodnom poglavlju, mogu se koristiti standardne statističke mere interesantnosti za rangiranje potencijalno korisnih asocijativnih pravila (poglavlje 4.1). Međutim, ove mere interesantnosti definisane su podrazumevajući nezavisnost atributa koji čine asocijativno pravilo. Asocijativna pravila o korišćenju web sajtova koja sadrže web stranice povezane meni i hiperlink strukturom web sajta mogu imati „neopravdano“ povišene vrednosti standardnih statističkih mera interesantnosti.

Čak i posle eliminacije statistički očekivanih pravila primenom metoda predloženih u poglavlju 4.2, u skupu otkrivenih asocijativnih pravila može preostati znatan broj pravila čije vrednosti standardnih statističkih mera interesantnosti ne odgovaraju njihovoj stvarnoj interesantnosti sa stanovišta analitičara podataka.

U ovom poglavlju predlažemo modifikaciju mera interesantnosti jednog dela asocijativnih pravila o korišćenju web sajtova, koja su delimično očekivana i ne previše interesantna analitičarima podataka, a ipak imaju povišene vrednosti statističkih mera interesantnosti. Smatramo da se na ovaj način vrednosti mera interesantnosti asocijativnih pravila o korišćenju web sajtova približavaju njihovoj stvarnoj interesantnosti sa stanovišta analitičara podataka.

4.3.1 Lokalna interesantnost asocijativnih pravila u odnosu na zajednički natkoncept

Predložena modifikacija interesantnosti odnosi se na asocijativna pravila čija leva i desna strana sadrže web stranice koje su potkoncept zajedničkog natkoncepta u okviru konceptne hijerarhije definisane u poglavlju 4.2.3. Kada se statističke mere interesantnosti, čija se vrednost računa u odnosu na veličinu skupa svih sesija primene na ovakva pravila, ona bivaju „nepravedno“ visoko rangirana.

Predlažemo da se na takva asocijativna pravila primene mere interesantnosti čija se vrednost računa lokalno, u skupu sesija koje sadrže zajednički natkoncept. Obzirom da je veličina ovog skupa znatno manja od veličine skupa svih sesija, vrednosti statističkih mera interesantnosti, kao što su lift, added value i Z-score na ovaj način se znatno smanjuju.

Na primer, vrednost lokalne mere interesantnosti kratkog asocijativnog pravila oblika $w_1 \rightarrow w_2$, pri čemu su w_1 i w_2 web stranice, čiji je zajednički natkoncept web stranica w', može se tumačiti na sledeći način. Posetioci web sajta koji su posetili web stranicu w', i pri tome su takođe posetili web stranicu w_1, zainteresovani su i za web stranicu w_2, sa određenim stepenom poverenja.

4.3.2 Definicija lokalnih mera interesantnosti

U nastavku definišemo uslov pod kojim predlažemo primenu lokalne mere interesantnosti asocijativnih pravila (definicija 4.10). Potom definišemo sledeće lokalne mere interesantnosti: lokalni lift (definicija 4.11), lokalni added value (definicija 4.12) i lokalni Z-score (definicija 4.13).

Napomena: U nastavku i dalje koristimo x kao skraćenu oznaku za $\{x\}$, i Ax kao skraćenu oznaku za $A \cup \{x\}$, $x \in I$, $A \subseteq I$.

Definicija 4.10

Neka su $x_1, x_2, x' \in I$ web objekti takvi da je x' natkoncept za x_1 i za x_2 (def. 4.9). Neka su $A, B \subseteq I$ skupovi web objekata takvi da važi $x_1, x_2, x' \notin A \cup B$, $A \cap B = \emptyset$. Tada asocijativno pravilo oblika $Ax_1 \rightarrow Bx_2$ zadovoljava uslov za definisanje lokalnih mera interesantnosti.

Primetimo da pod uslovima datim definicijom 4.10 važi: $S^{Ax_1}, S^{Bx_2} \subseteq S^{x'}$, odnosno da su Ax_1 i Bx_2 specifičniji skupovi atributa u odnosu na opštiji skup atributa $\{x'\}$ (def.4.3).

Lokalni lift

Standardna lift mera asocijativnog pravila (poglavlje 4.1) primenjena na asocijativna pravila o korišćenju web sajtova opšteg oblika $X \rightarrow Y$, $X, Y \subseteq I$, $X \cap Y = \emptyset$, koja su otkrivena u skupu sesija S može se izraziti na sledeći način:

$$lift(X \rightarrow Y) = \frac{P(XY)}{P(X)P(Y)} = \frac{\frac{|S^{X \cup Y}|}{|S|}}{\frac{|S^X|}{|S|} \cdot \frac{|S^Y|}{|S|}} = \frac{|S^X \cap S^Y| \cdot |S|}{|S^X| \cdot |S^Y|}$$

Definicija 4.11

Za asocijativno pravilo oblika $Ax_1 \rightarrow Bx_2$ koje zadovoljava uslove date definicijom 4.10, definišemo *lokalni lift* kao meru interesantnosti:

$$liftLocal(Ax_1 \rightarrow Bx_2) = \frac{\frac{|S^{Ax_1 \cup Bx_2}|}{|S^{x'}|}}{\frac{|S^{Ax_1}|}{|S^{x'}|} \cdot \frac{|S^{Bx_2}|}{|S^{x'}|}} = \frac{|S^{Ax_1} \cap S^{Bx_2}| \cdot |S^{x'}|}{|S^{Ax_1}| \cdot |S^{Bx_2}|}$$

4.3.2.1 Lokalni added value

Standardna added value mera interesantnosti asocijativnog pravila (poglavlje 4.1) primenjena na asocijativna pravila o korišćenju web sajtova opšteg oblika $X \rightarrow Y$, $X, Y \subseteq I, X \cap Y = \emptyset$, koja su otkrivena u skupu sesija S može se izraziti na sledeći način:

$$AV(X \rightarrow Y) = P(Y|X) - P(Y) = \frac{|S^{X \cup Y}|}{|S^X|} - \frac{|S^Y|}{|S|} = \frac{|S^X \cap S^Y|}{|S^X|} - \frac{|S^Y|}{|S|}$$

Definicija 4.12

Za asocijativno pravilo oblika $Ax_1 \rightarrow Bx_2$ koje zadovoljava uslove date definicijom 4.10, definišemo *lokalni added value* kao meru interesantnosti:

$$AVLocal(Ax_1 \rightarrow Bx_2) = \frac{|S^{Ax_1 \cup Bx_2}|}{|S^{Ax_1}|} - \frac{|S^{Bx_2}|}{|S^{x'}|} = \frac{|S^{Ax_1} \cap S^{Bx_2}|}{|S^{Ax_1}|} - \frac{|S^{Bx_2}|}{|S^{x'}|}$$

4.3.2.2 Lokalni Z-score kao mera interesantnosti

Standardna Z-score mera interesantnosti asocijativnog pravila o korišćenju web sajtova opšteg oblika $X \rightarrow Y$ data je definicijom 4.2 (poglavlje 4.2.1).

Definicija 4.13

Za asocijativno pravilo oblika $Ax_1 \rightarrow Bx_2$ koje zadovoljava uslove date definicijom 4.10, definišemo *lokalni Z-score* kao meru interesantnosti u skupu sesija $S^{x'}$:

$$ZLocal^{S^{x'}}(Ax_1 \rightarrow Bx_2) = \frac{C' - \mu'}{\sigma}, \qquad pri\ čemu\ je$$

$$C' = |S^{Ax_1 \cup Bx_2}|, \qquad \mu' = |S^{x'}| \cdot p',$$

$$p' = \frac{|S^{Ax_1}| \cdot |S^{Bx_2}|}{|S^{x'}| \cdot |S^{x'}|}, \qquad \sigma' = \sqrt{|S^{x'}| \cdot p'(1 - p')}$$

Obzirom da važi $|S^{x'}| \leq |S|$, može se pokazati da za svaku prethodno definisanu lokalnu meru interesantnosti LI baziranu na standardnoj statističkoj meri interesantnosti I važi: $LI(Ax_1 \rightarrow Bx_2) \leq I(Ax_1 \rightarrow Bx_2)$.

Tri prethodno definisane lokalne mere interesantnosti (lokalni lift, lokalni added value i lokalni Z-score) su korišćene u eksperimentalnom istraživanju (poglavlje 6.3) i implementirane u softverski sistem (poglavlje 5.6). Po analogiji, relativno je jednostavno definisati i druge lokalne mere interesantnosti bazirane na standardnim statističkim merama (poglavlje 4.1). Pri tome, treba imati u vidu da neke mere interesantnosti, kao što je confidence, ne zavise od veličine skupa svih sesija. Za takve mere interesantnosti nema smisla definisati lokalnu meru, jer bi ona za svako asocijativno pravilo imala jednaku vrednost kao i standardna mera.

4.3.3 Primeri iz stvarnog skupa podataka

Motivaciju za definisanje lokalnih mera interesantnosti potkrepljujemo primerima asocijativnih pravila iz stvarnog skupa podataka o korišćenju web sajta Fakulteta organizacionih nauka u Beogradu.

U Tabeli 4.4 dat je primer sedam pravila otkrivenih u ovom skupu podataka, koja nisu eliminisana primenom metoda za eliminaciju statistički očekivanih pravila (poglavlje 4.2). Kolone „Lift“ i „Z“ označavaju vrednosti standardnih lift i Z-score mera

interesantnosti, dok kolone „L-Lift“ i „L-Z“ označavaju vrednosti lokalnih lift i Z-score mera interesantnosti.

Br	Levo	Desno	Lift	L-Lift	Z	L-Z
1	/osnovnestudije/om/index.html	/osnovnestudije/men/index.html	17.2	3.4	60.5	19.7
2	/osnovnestudije/om/index.html	/osnovnestudije/uk/index.html	30.2	6.1	73.6	27.8
3	/osnovnestudije/uk/index.html	/osnovnestudije/isit/index.html	6.7	1.3	28.4	3.8
4	/zivot/index.html	/osnovnestudije/index.html	2.6	-	13.8	-
5	/zivot/index.html	/sluzbe/index.html	6.1	-	19.7	-
6	/zivot/index.html	/ofakultetu/index.html	4.5	-	16.9	-
7	/osnovnestudije/nastava/index.html	/osnovnestudije/isit/index.html	3.6	-	15.0	-

Tabela 4.4. Primeri vrednosti lokalnih mera interesantnosti iz stvarnog skupa podataka

Prva tri pravila u tabeli 4.4 povezuju web stranice koje su potkoncepti zajedničkog natkoncepta. Web stranice */osnovnestudije/om/index.html, /osnovnestudije/men/index.html, /osnovnestudije/uk/index.html i /osnovnestudije/isit/index.html* odnose se na pojedine studijske programe osnovnih studija. Njihov zajednički natkoncept je web stranica */osnovnestudije/index.html,* koja se odnosi na osnovne studije uopšte. Pravila sa rednim brojem 4, 5, 6 i 7 povezuju web stranice koje nemaju zajednički natkoncept, pa za ova pravila nisu defnisane vrednosti lokalnih mera interesantnosti.

Vrednosti standardnih lift i Z-score mera interesantnosti za prva tri pravila su znatno povišene u odnosu na preostala četiri pravila. Međutim, za analitičare podataka i web mastere, koji su upoznati sa strukturom web sajta i značenjem ovih web stranica, prva tri pravila su dobrim delom očekivana i ne previše interesantna. Za njih su vrlo verovatno barem toliko interesantna i neka od pravila sa rednim brojem 4, 5, 6 i 7 u tabeli, koja povezuju web stranice koje nemaju zajednički natkoncept, iako imaju niže vrednosti standardnih mera interesantnosti („Lift“ i „Z“). Zaključujemo da rangiranje navedenih asocijativnih pravila prema standardnim merama interesantnosti ne odgovara potencijalnoj interesantnosti ovih asocijativnih pravila sa stanovišta analitičara podataka.

Sa druge strane, vrednosti lokalnih mera interesantnosti („L-Lift“ i „L-Z“) asocijativnih pravila 1, 2 i 3 su znatno niže od vrednosti njihovih standardnih mera interesantnosti („Lift“ i „Z“). Dakle, relativna interesantnost asocijativnih pravila 1, 2 i 3 je znatno umanjena korišćenjem lokalnih mera interesantnosti.

Pri tome, pravila sa rednim brojevima 1, 2 i 3 mogu se porediti koristeći lokalne mere interesantnosti. Na primer, pravilo sa rednim brojem 3 ima relativno nisku vrednost lokalnih mera („L-Lift“ i „L-Z“), što se može tumačiti na sledeći način. Posetioci web stranice */osnovnestudije/index.html,* koja se odnosi na osnovne studije uopšte, koji posećuju web stranicu */osnovnestudije/uk/index.html* (studijski program „Upravljanje kvalitetom”) nisu previše zainteresovani za web stranicu */osnovnestudije/isit/index.html* (studijski program „Informacione tehnologije”). Sa druge strane, pravilo sa rednim brojem 1 ima povišene vrednosti lokalnih mera, što se može tumačiti na sledeći način. Posetioci web stranice */osnovnestudije/index.html,* koja se odnosi na osnovne studije uopšte, koji posećuju web stranicu */osnovnestudije/om/index.html* (studijski program „Operacioni menadžment”) zainteresovani su za web stranicu */osnovnestudije/uk/index.html* (studijski program „Upravljanje kvalitetom”), znatno više nego što je to očekivano, na šta ukazuju visoke L-Lift i L-Z vrednosti pravila broj 2.

Rangiranje otkrivenih asocijativnih pravila primenom lokalnih i standardnih statističkih mera interesantnosti na stvarnim skupovima podataka detaljnije je analizirano u poglavlju 6.3.

5 Softverski sistem za otkrivanje asocijativnih pravila o korišćenju web sajtova

U okviru ovog istraživanja implementiran je softverski sistem koji obuhvata kompletan proces otkrivanja asocijativnih pravila o korišćenju web sajtova. Implementirane su metode za pretprocesiranje web log podataka, algoritmi za generisanje asocijativnih pravila, metode za eliminaciju neinteresantnih asocijativnih pravila, kao i metode za rangiranje otkrivenih asocijativnih pravila korišćenjem više različitih mera interesantnosti. Sistem je implementiran kao samostalna Windows aplikacija, koristeći programski jezik C#.NET i razvojno okruženje MS Visual Studio 2010.

5.1 Elementi objektno-orijentisanog dizajna

Softverski sistem za otkrivanje asocijativnih pravila o korišćenju web sajtova je dizajniran objektno-orijentisano. Korišćeni su poznati principi objektno-orijentisanog programiranja – enkapsulacija, apstrakcija, nasleđivanje i polimorfizam (Rumbaugh et al., 1991; Gamma et al., 1993; Liberty, 2005).

Enkapsulacija je mehanizam pomoću koga se delovi programskog koda i podaci objedinjuju u celinu. Princip enkapsulacije je podržan konceptom klase, koja obuhvata podatke i metode koji služe za pristup i manipulaciju tim podacima. Instance klase nazivaju se objekti, koji predstavljaju entitete. Objedinjavanjem podataka i programskog koda sprečava se manipulacija podacima od strane spoljnih delova programa, čime se smanjuje mogućnost greške pri korišćenju objekta. Jedini način da se pristupi podacima unutar objekta je slanjem „poruka“ objektu, odnosno pozivom javnih metoda klase kojoj objekat pripada. Na taj način se samo indirektno može uticati na stanje objekta. Detalji implementacije metoda klase skriveni su od „spoljnog sveta“. Oni se mogu promeniti, bez uticaja na ostale delove programskog koda, koji koriste tu klasu.

Apstrakcija podataka je kreiranje pojednostavljenog pogleda na podatke, pri čemu se nepotrebni detalji sakrivaju. U programskim jezicima generalno, apstrakcija se postiže implementiranjem apstraktnih tipova podataka. U objektno-orijentisanim programskim jezicima to se postiže implementiranjem klasa. Principi enkapsulacije i

apstrakcije u objektno-orijentisanom programiranju su usko povezani i oba se zasnivaju na implementiranju klasa, u okviru kojih se izlažu javni metodi kao interfejs za komunikaciju, dok se detalji implementacije sakrivaju.

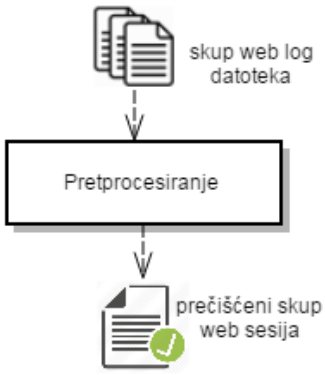
Nasleđivanje podrazumeva hijerarhijsku strukturu klasa. Pri tome, klase „deca“ (izvedene klase) mogu da naslede podatke i funkcionalnost od klasa „roditelja“ (osnovnih klasa). Ovakva relacija između klasa realizuje relaciju „je/jeste“, odnosno generalizaciju. Zahvaljujući principu nasleđivanja izbegava se dupliranje implementacije funkcionalnosti, koju izvedene klase nasleđuju od osnovnih klasa.

Princip polimorfizma označava „jedno ime, mnoštvo oblika“. U objektno-orijentisanom programiranju moguće je implementirati više programskih metoda koje imaju isto ime, ali se ponašaju različito. Postoje dva tipa polimorfizma – preopterećivanje i preklapanje. Preopterećivanje metoda postiže se implementiranjem više metoda istog imena, koji imaju raziličite tipove parametara, te se pozivaju na različit način. U tom slučaju odluka o tome koji metod će biti izvršen donosi se u trenutku kompajliranja programa. Drugi tip polimorfizma – preklapanje metoda, može se realizovati samo u sklopu sa nasleđivanjem klasa. Tada se implementira više metoda istog imena i istih tipova parametara, ali u klasama koje nasleđuju jedna drugu. Odluka o tome koji metod će biti izvršen donosi se tokom izvršavanja programa, u zavisnosti od klase kojoj pripada objekat za koji je metod pozvan.

5.2 Arhitektura sistema

Softverski sistem za otkrivanje asocijativnih pravila o korišćenju web sajtova sastoji se iz dva podsistema. Uloga prvog podsistema je pretprocesiranje web log podataka, dok se u okviru drugog otkrivaju asocijativna pravila o korišćenju web sajtova na osnovu prethodno pripremljenih web log podataka.

Podsistem za pretprocesiranje web log podataka ima ulogu da pripremi web log datoteke za proces otkrivanja asocijativnih pravila. Globalna slika ovog procesa data je na slici 5.2, a detalji su opisani u poglavlju 5.3.

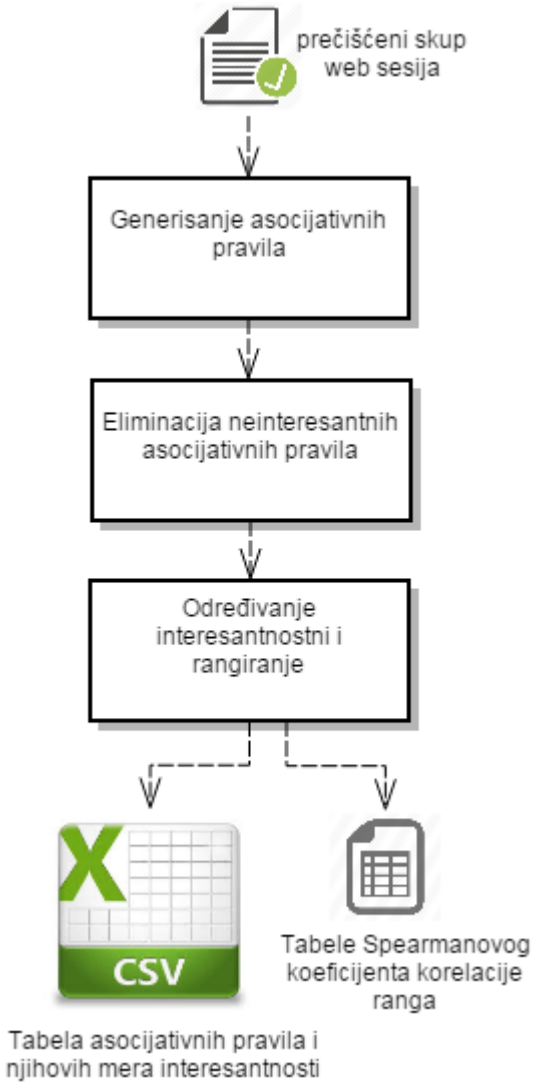


Slika 5.2. Globalni pogled na proces pretprocesiranja web log podataka

Podsistem za otkrivanje asocijativnih pravila o korišćenju web sajtova obuhvata proces generisanja asocijativnih pravila, eliminaciju neinteresantnih asocijativnih pravila, određivanje interesantnosti preostalih asocijativnih pravila i njihovo rangiranje (slika 5.3). Na ulazu se učitava datoteka koja sadrži prečišćeni skup web sesija, a na izlazu se generišu dve datoteke:

- tabela asocijativnih pravila sa njihovim vrednostima raznih mera interesantnosti
- tabela sličnosti rangiranja raznim merama interesantnosti (Spearman-ov koeficijent korelacije ranga)

Globalna slika ovog procesa data je na slici 5.3, a detalji su opisani u poglavljima 5.4, 5.5 i 5.6.

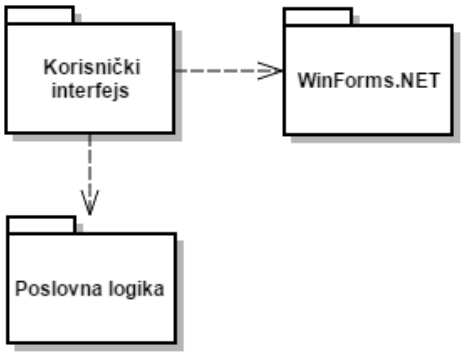


Slika 5.3. Globalni pogled na proces otkrivanja asocijativnih pravila

Oba podsistema dizajnirana su kao dvoslojne objektno-orijentisane aplikacije. Prvi sloj čini korisnički interfejs, dok je u okviru drugog sloja implementirana poslovna logika, koja uključuje metode za pretprocesiranje web log podataka, algoritme za otkrivanje asocijativnih pravila, eliminaciju neinteresantnih pravila, kao i metode za rangiranje otkrivenih pravila korišćenjem više različitih mera interesantnosti. Otkrivena asocijativna pravila i njihove mere interesantnosti čuvaju se u formi Excel datoteka, te nije korišćen treći sloj, u okviru kojeg bi se eventualno vršio pristup bazi podataka.

Objektno-orijentisani dizajn i višeslojna arhitektura ovog softverskog sistema omogućuje njegovu proširivost i unapređenje u narednim verzijama.

Korisnički interfejs trenutno je implementiran koristeći Windows.Forms.NET biblioteku klasa. U narednim verzijama ovog softverskog sistema bilo bi moguće zameniti sloj korisničkog interfejsa i eventualno ga implementirati kao Web aplikaciju, bez ikakvog uticaja na sloj poslovne logike. Globalni pogled na arhitekturu sistema prikazan je na slici 5.1.



Slika 5.1. Globalna arhitektura sistema

5.2.1 Elementi Weka data mining sistema – poređenje

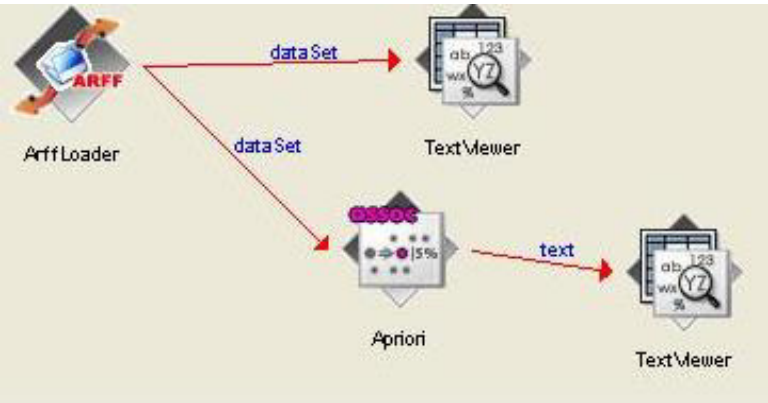
Jedan od trenutno najpopularnijih open-source data mining alata je Weka – softverski sistem održavan na Univerzitetu Waikato na Novom Zelandu (Hall et al., 2009). Weka je data mining sistem opšte namene, i pored otkrivanja asocijativnih pravila sadrži implementaciju raznih algoritama iz oblasti mašinskog učenja, kao što su klasterovanje, klasifikacija, regresiona i korelaciona analiza. Implementiran je kao objektno-orijentisana softverska aplikacija u Java programskom jeziku. Moguće je koristiti ga kao samostalnu aplikaciju, ili koristiti neke od njegovih komponenti, pozivajući ih iz sopstvenog programskog koda.

Weka sistem je korišćen za otkrivanje asocijativnih pravila o korišćenju web sajtova u našem prethodnom istraživanju (Dimitrijević & Bošnjak, 2010). Obzirom da Weka ne sadrži modul za pretprocesiranje web log podataka, u tu svrhu je korišćen poseban alat (Dettmar, 2004).

Weka sadrži implementaciju jedne od verzija Apriori algoritma za otkrivanje asocijativnih pravila (Agrawal & Srikant, 1994). Pored minimalnog support parametra, algoritam dozvoljava odabir maksimalnog broja otkrivenih pravila. Takođe su implementirane četiri mere interesantnosti: confidence, lift, leverage, i conviction.

Modul za otkrivanje asocijativnih pravila u Weka sistemu prihvata ulaznu datoteku u takozvanom Arff formatu. Arff datoteka sadrži zaglavlje koje opisuje atribute, za kojim sledi lista transakcija (u našem slučaju web sesija), od kojih svaka transakcija sadrži listu atributa i njihovih vrednosti. Postoje dva tipa formata Arff datoteke – prvi je pogodan za guste podatke, a drugi za proređene. U oba formata, web stranica se predstavlja kao binarni atribut, koji uzima true/false vrednosti u zavisnosti od toga da li ona pripada web sesiji. Pojava web stranica u web sesijama je relativno retka, te je za otkrivanje asocijativnih pravila u web log datotekama korišćen Arff format predviđen za proređene podatke. Nakon pretprocesiranja web log datoteke, rezultujuće podatke je neophodno transformisati u Arff format, što je urađeno posebnim programskim skriptama (Dimitrijević & Bošnjak, 2010).

Na slici 5.4 prikazan je proces otkrivanja asocijativnih pravila korišćenjem Weka sistema (Dimitrijević & Bošnjak, 2010). Korišćen je Weka KnowledgeFlow interface, koji omogućuje sastavljanje različitih Weka modula, koji postaju deo data mining procesa. Na ulazu se prihvata datoteka u Arff formatu, koja se može prelistavati korišćenjem TextViewer modula. Ova datoteka se učitava preko Arff loader modula, a potom se modulom Apriori otkrivaju asocijativna pravila, koja ulaze u novi TextViewer modul, gde se prikazuju rezultati.



Slika 5.4: Proces otkrivanja asocijativnih pravila u Weka sistemu (Dimitrijević & Bošnjak, 2010)

Jedna od mana eventualnog korišćenja Weka sistema u ovom istraživanju bila bi činjenica da priprema i pretprocesiranje web log podataka nisu integrisane u Weka sistem. Još važnije, u okviru Weka sistema implementirane su samo četiri mere interesantnosti asocijativnih pravila, i nije implementirana metoda za prečišćavanje asocijativnih pravila. Obzirom da je Weka sofver otvorenog koda, eventualno bi bilo moguće proširiti Weka sistem dodavanjem novih mera interesantnosti i metoda eliminacije neinteresantnih pravila. Međutim, smatramo da je razvoj nezavisnog, specijalizovanog softverskog sistema koji integriše sve korake procesa otkrivanja asocijativnih pravila o korišćenju web sajtova, uključujući nove metode predložene u poglavlju 4, efikasniji način da izvršimo sve eksperimente predviđene ovim istraživanjem.

Pored Weka sistema, na tržištu postoje i komercijalni data mining sistemi, kao što su RapidMiner, Estard Data Miner, IBM Analytics, itd. Pored nepristupačne cene, ove sisteme ne bi bilo moguće koristiti jer ne podržavaju funkcionalnosti potrebne da se izvrše eksperimenti provere efikasnosti metoda predloženih u ovom istraživanju.

5.3 Pretprocesiranje web log podataka

Uloga podsistema za pretprocesiranje web log podataka je da podatke sadržane u tekstualnim web log datotekama prečisti i pripremi za proces otkrivanja asocijativnih pravila.

Proces pretprocesiranja web log podataka (slika 5.5) obuhvata sledeće korake:

1. *Formiranje liste web log zahteva*

Na ulazu se prihvata skup web log datoteka koje sadrže podatke o posetama datom web sajtu uskladištene na web serveru. Podržana su dva standardna formata – standard i extended web log format (poglavlje 2.2.1). Parsiranjem podataka sadržanih u web log datotekama (klasa LogFile) formira se lista web log zahteva (lista objekata klase Request). Pri tome se za svaki web log zahtev skladište relevantni podaci u okviru Request objekta, koji se koriste u narednim koracima pretprocesiranja:

- Url – jedinsteni identifikator web objekta na koji se odnosi web zahtev
- IP – adresa sa koje je web zahtev upućen
- Time – vreme upućivanja web zahteva
- SessionID – jedinstveni identifikator web sesije kojoj web zahtev pripada, koji se određuje u narednim koracima pretprocesiranja

2. *Eliminacija irelevantnih web log zahteva*

Eliminacija irelevantnih web zahteva vrši se u klasi LogFile, koristeći tekstualnu datoteku u kojoj se čuvaju podaci o delovima naziva irelevantnih web log objekata (najčešće njihove ekstenzije). Irelevantni web objekti koji se eliminišu predstavljaju slike, animacije, ugrađene stilove, i slične pomoćne objekte ugrađene u web stranice.

3. *Eliminacija robotskih web zahteva*

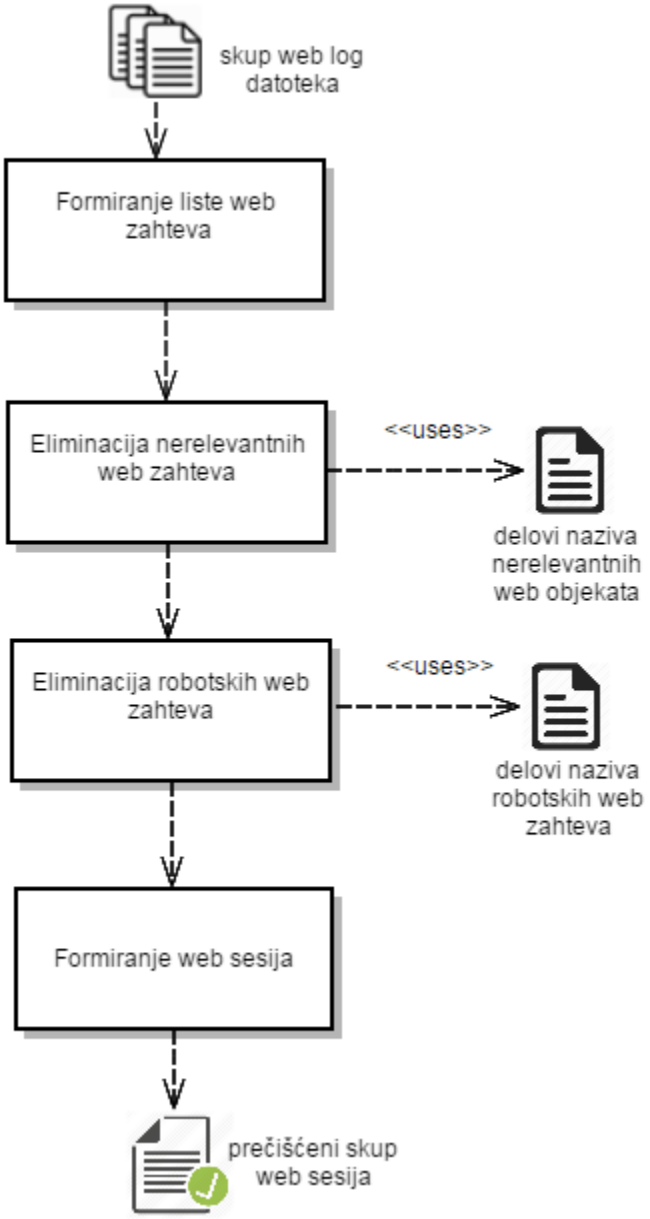
Eliminacija robotskih web zahteva vrši se u klasi LogFile, koristeći tekstualnu datoteku u kojoj se čuvaju podaci o delovima naziva user agenata koji se odnose na automatske web crawler-e (takozvane web robote).

4. *Formiranje web sesija*

Algoritam za dodeljivanje jedinstvenog identifikatora web sesije svakom web zahtevu implementiran je u klasi LogFile. Pri tome se koristi parametar SessionTimeOut, kojim se definiše maksimalno vreme između dva web zahteva unutar jedne web sesije.

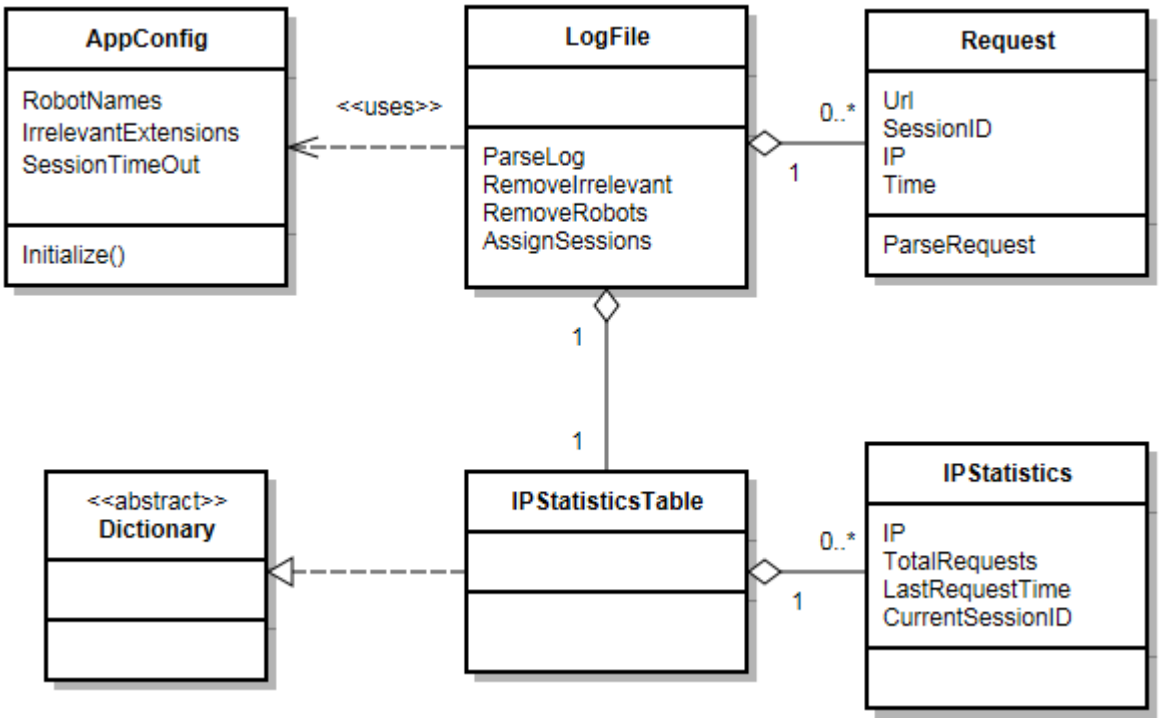
Pored toga, vrši se eliminacija kratkih sesija koje sadrže samo 1 web stranicu, dugih sesija koje sadrže više od 10 web stranica, kao i sesija koje ne sadrže osnovnu stranicu web sajta. Naime, ovakve web sesije bi mogle nastati od strane eventualno nedetektovanih web robota, ili biti nepotpune usled cache-iranja od strane web browser-a, pa se eliminišu u sklopu pretprocesiranja.

Rezultat pretprocesiranja je tekstualna datoteka koja sadrži prečišćeni skup web zahteva, iz koga su eliminisani zahtevi za irelevantnim web objektima, dok su relevantni zahtevi svrstani u web sesije.



Slika 5.5. Proces pretprocesiranja web log podataka

Dijagram klasa implementiranih u okviru modula za pretprocesiranje web log podataka dat je na slici 5.6. Na dijagramu su prikazani nazivi ključnih metoda implementiranih u okviru prikazanih klasa, dok su detalji izostavljeni.



Slika 5.6. Dijagram klasa modula za pretprocesiranje web log datoteka

Klasa AppConfig sadrži opšta podešavanja: vreme isteka web sesije, nazive datoteka u kojima se nalaze delovi naziva irelevantnih web objekata, kao i robotskih zahteva. Prilikom pokretanja sistema ovi podaci se učitavaju, a koristi ih klasa LogFile, koja sadrži metode najvišeg nivoa, kojima se realizuju potprocesi u okviru procesa pretprocesiranja:

- ParseLog – parsiranje web log datoteka i formiranje liste web zahteva
- RemoveIrrelevant – eliminacija irelevantnih web zahteva
- RemoveRobots – eliminacija robotskih web zahteva
- AssignSessions – raspodeljivanje web zahteva prema web sesijama

Svaki web zahtev predstavljen je kao objekat klase Request, dok klasa LogFile sadrži listu objekata klase Request (listu svih web zahteva).

U okviru algoritma za dodeljivanje jedinstvenog identifikatora web sesije svakom web zahtevu, koristi se klasa IPStatisticsTable, koja sadrži listu objekata IPStatistics klase. Svaki objekat IPStatistics klase odnosi se na po jednu IP adresu sa koje su upućivani web zahtevi na web server. Na osnovu vremena kada je poslednji zahtev upućen sa date IP adrese (LastRequestTime) odlučuje se o tome da li zahtev pripada već postojećoj web sesiji (CurrentSessionID), ili je potrebno formirati novu web sesiju. U slučaju formiranja nove web sesije, CurrentSessionID za datu IP adresu dobija novu vrednost.

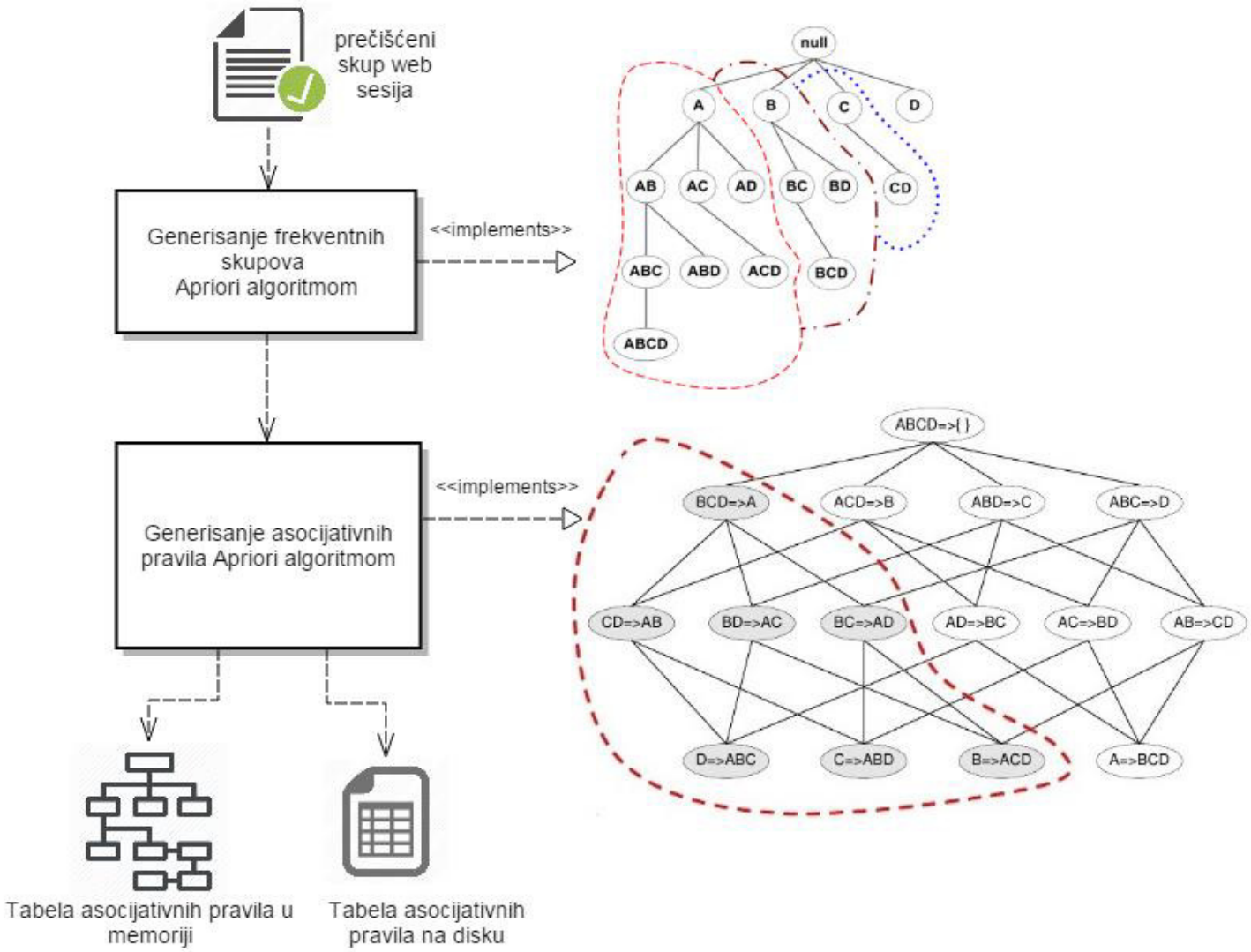
U cilju raspoređivanja web zahteva u web sesije koristi se i klasa IPStatisticsTable koja nasleđuje apstraktnu klasu Dictionary. Na taj način se implementira indeksirani pristup tabeli IPStatistics objekata, koji sadrže informaciju o vremenu kada je upućen poslednji web zahtev sa date IP adrese, pri čemu se IP adresa koristi kao indeks IPStatistics tabele.

5.4 Ugrađeni algoritmi za pronalaženje asocijativnih pravila

U okviru sistema za otkrivanje asocijativnih pravila o korišćenju web sajtova ugrađena je verzija Apriori algoritma opisanog u poglavlju 3.1.

Proces pronalaženja asocijativnih pravila odvija se u dve faze (slika 5.7). Ulazni podaci se učitavaju iz datoteke nastale procesom pretprocesiranja, koja sadrži prečišćeni skup web sesija i web zahteva koji im pripadaju. Na izlazu se generiše tabela otkrivenih asocijativnih pravila koja se koristi u narednim koracima procesa otkrivanja asocijativnih pravila. Pri tome se radi kontrole tabela asocijativnih pravila snima i u obliku datoteke na disku.

U prvoj fazi procesa pronalaze se svi frekventni skupovi stavki (web stranica). U drugoj fazi se na osnovu frekventnih skupova stavki pronalaze sva asocijativna pravila. Pri tome je u okviru obe faze procesa implementiran odgovarajući deo Apriori algoritma (slika 5.7).



Slika 5.7. Proces pronalaženja asocijativnih pravila

Na slici 5.8 prikazan je dijagram klasa i njihovih ključnih elemenata implementiranih u okviru podsistema za pronalaženje asocijativnih pravila.

Klasa App sadrži osnovne parametre algoritma za otkrivanje asocijativnih pravila, koji se postavljaju prilikom pokretanja sistema. Pored minimalnog support i confidence praga (ConfidenceTresh i SupportThresh), podešava se i maksimalna dužina otkrivenih asocijativnih pravila, koja je ekvivalentna maksimalnom broju nivoa stabla otkrivenih frekventnih skupova stavki (MaxLevels). Metoda ParseInputFile parsiranjem ulazne datoteke formira listu stavki (objekat klase ItemTable) i listu web sesija (objekat klase SessionTable). Izvršavanjem metode GenerateFreqSets generiše se tabela frekventnih skupova (objekat klase FreqSetTable) izvršavanjem implementiranog Apriori algoritma. Izvršavanjem metode GenerateRules generiše se tabela otkrivenih asocijativnih pravila (objekat klase RuleTable).

U okviru Apriori algoritma (poglavlje 3.1) generisanje frekventnih skupova se vrši prema nivoima njihove dužine, počevši od dužine 1. Svaki naredni nivo generiše se na osnovu frekventnih skupova prethodnog nivoa. Ovaj algoritam se ponavlja sve dok postoje frekventni skupovi na datom nivou dužine, ili dok maksimalni broj nivoa ne dostigne vrednost zadatog parametra MaxLevels.

Svaka web sesija (objekat klase Session) sadrži identifikacioni broj web sesije (SessionID), IP adresu sa koje su pristigli web zahtevi koji čine web sesiju (IP), kao i listu stavki (web stranica) koje čine web sesiju (objekat klase ItemTable).

Klase SessionTable i ItemTable implementiraju abstraktnu klasu Dictionary, što čini pretraživanje tabele web sesija (SessionTable), odnosno tabele web stranica (ItemTable) efikasnim. Pri tome je tabela web sesija indeksirana prema identifikacionom broju web sesije, dok je tabela web stranica indeksirana prema URL adresi web stranice.

Svakoj web stranici odgovara objekat klase Item, u okviru kojeg se čuvaju podaci o URL adresi web stranice (Url) i učestalosti pojavljivanja web stranice u skupu web sesija (Support).

Klasa Item implementira IComparable interfejs, koji omogućuje da liste web stranica unutar svakog frekventnog skupa budu sortirane alfabetski prema njihovim URL

adresama. To doprinosi povećanju efikasnosti algoritma za generisanje frekventnih skupova n-tog nivoa na osnovu frekventnih skupova (n-1)-og nivoa, što je deo Apriori algoritma.

Svakom frekventnom skupu odgovara jedan objekat klase FreqSet, koji sadrži listu web stranica koje čine frekventni skup. Pri tome je svaka web stranica predstavljena po jednim objektom klase Item.

Svaki objekat klase FreqSetLevel odnosi se na po jedan nivo frekventnih skupova generisanih tokom izvršavanja Apriori algoritma. Tako svaki objekat klase FreqSetLevel sadrži listu frekventnih skupova jednake dužine.

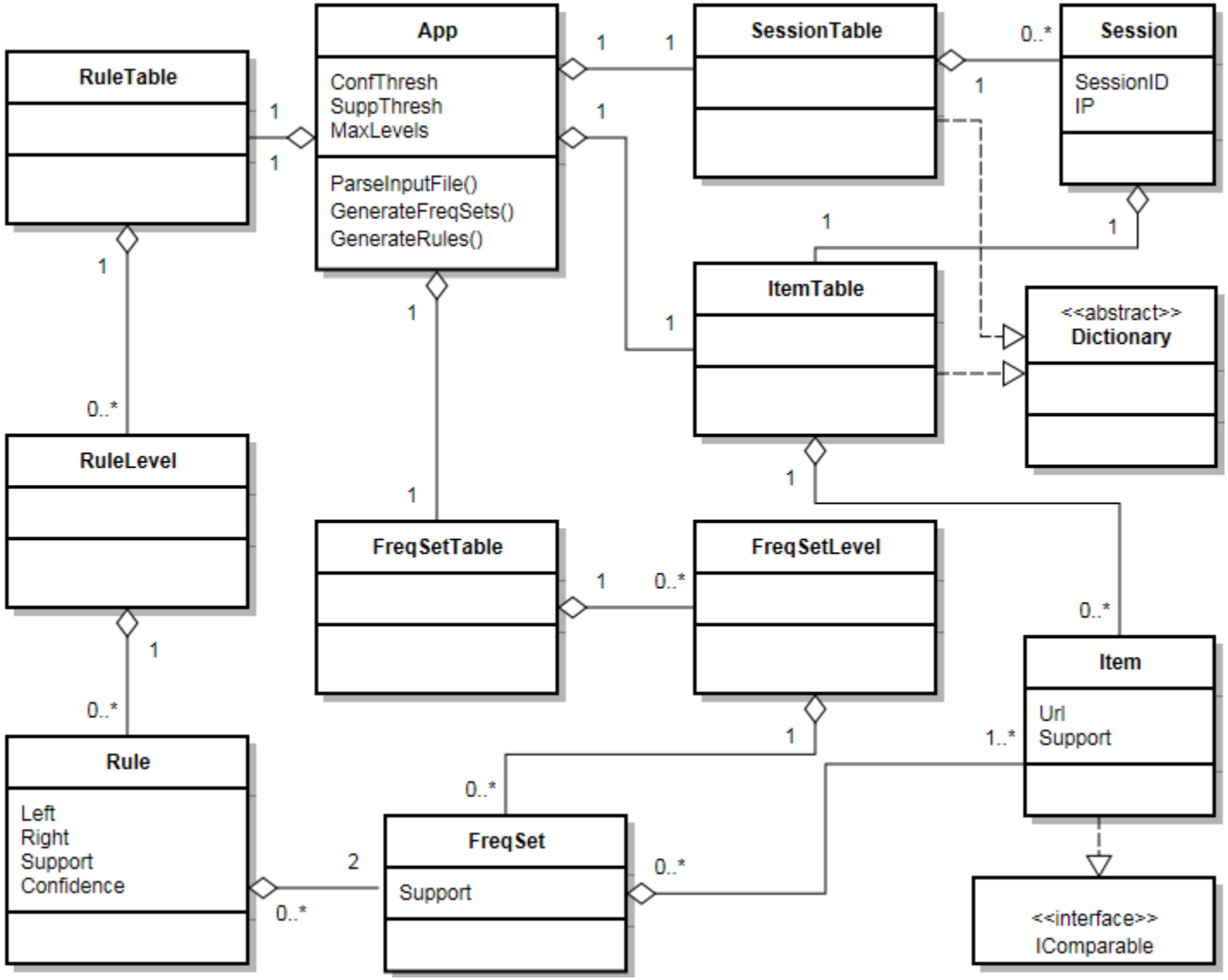
Klasa FreqSetTable sadrži listu svih generisanih frekventnih skupova, raspoređenih u nivoe dužine. Tako objekat klase FreqSetTable sadrži listu svih objekata klase FreqSetLevel, kreiranih tokom izvršavanja Apriori algoritma.

Svaki objekat klase Rule odnosi se na po jedno asocijativno pravilo. Objekat klase Rule sadrži referencu na dva objekta klase FreqSet, koji čine levu, odnosno desnu stranu asocijativnog pravila. U okviru objekta klase Rule čuvaju se vrednosti support i confidence mere izračunate za asocijativno pravilo tokom izvršavanja Apriori algoritma.

Svaki objekat klase RuleLevel sadrži listu asocijativnih pravila jednake dužine, pri čemu je dužina asocijativnog pravila definisana kao ukupni broj web stranica koje čine asocijativno pravilo.

Objekat klase RuleTable sadrži listu svih generisanih asocijativnih pravila, raspoređenih u nivoe dužine. Dakle, objekat klase RuleTable sadrži listu svih objekata klase RuleLevel.

Klasa ItemTable koristi se dvojako. Objekat klase ItemTable sadrži listu svih stavki (web stranica) koje se pojavljuju bar jednom u nekoj web sesiji, i kao takva je član klase App. Sa druge strane, svaka web sesija (objekat klase Session) sadrži referencu na po jedan objekat klase ItemTable, koji predstavlja listu svih web stranica koje čine web sesiju.



Slika 5.8. Dijagram klasa koje učestvuju u procesu pronalaženja asocijativnih pravila

5.5 Izbor parametara i mera interesantnosti

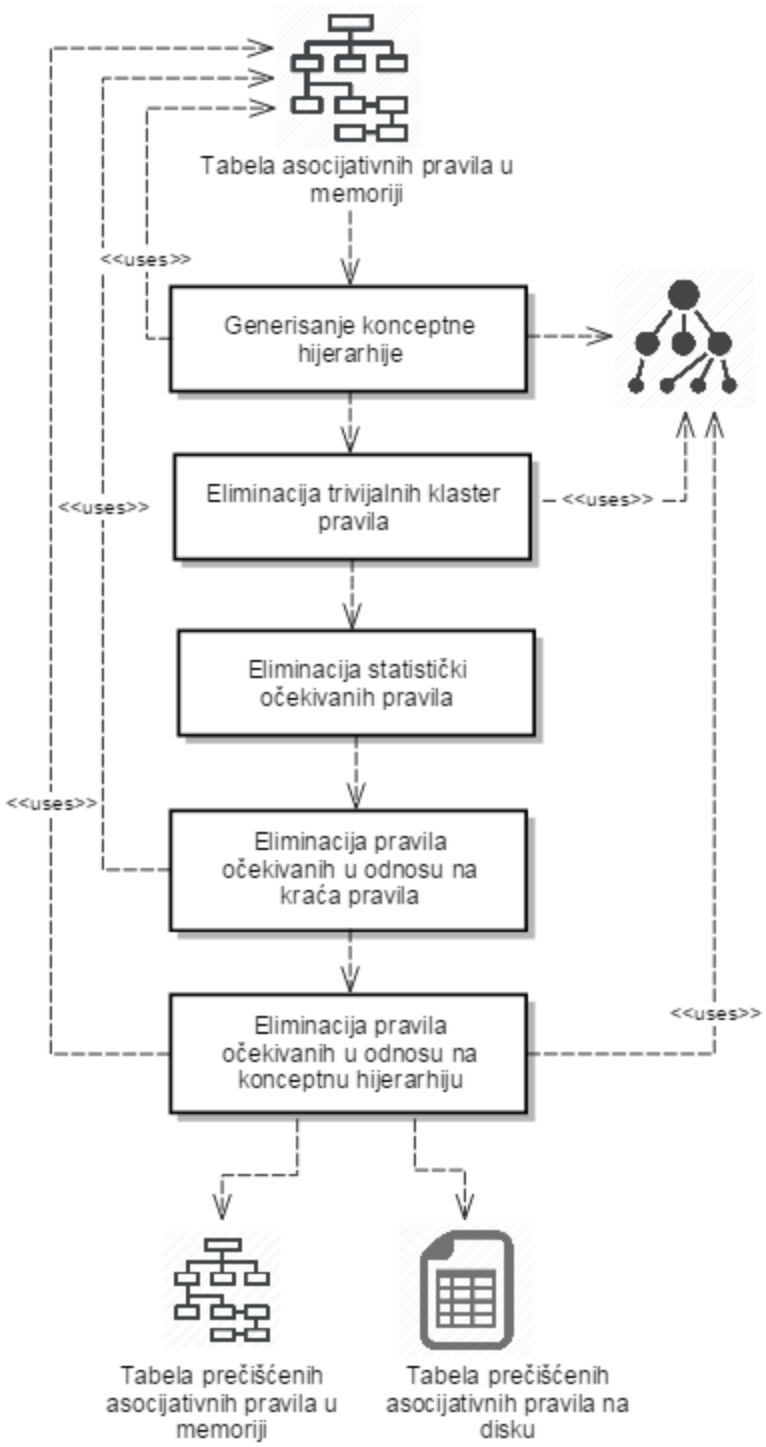
Kao rezultat procesa pronalaženja asocijativnih pravila opisanog u prethodnom poglavlju generiše se tabela asocijativnih pravila. Pri tome je za svako pravilo izračunata vrednost support i confidence mera interesantnosti.

U okviru procesa pronalaženja asocijativnih pravila i implementiranog Apriori algoritma omogućen je odabir vrednosti support i confidence parametara, na osnovu kojih se delimično vrši eliminacija neinteresantnih pravila. Međutim, kao što je objašnjeno u prethodnim poglavljima, i pored pažljivog odabira ovih parametara, preostaje preveliki broj pravila koja imaju visoke vrednosti statističkih mera interesantnosti, a ipak nisu stvarno interesantna analitičarima podataka. Stoga su u cilju povećanja kvaliteta otkrivenih asocijativnih pravila, u okviru sistema za otkrivanje pravila o korišćenju web sajtova implementirane metode eliminacije neinteresantnih asocijativnih pravila.

Na slici 5.9 predstavljen je proces eliminacije neinteresantnih asocijativnih pravila na osnovu zadatih parametara. Na početku procesa vrši se generisanje konceptne hijerarhije web stranica na osnovu asocijativnih pravila čija confidence vrednost prelazi maksimalnu vrednost zadatu putem parametra. Drugi korak procesa je eliminacija trivijalnih klaster pravila, pri čemu se koristi prethodno generisana konceptna hijerarhija. Treći korak procesa je eliminacija statistički očekivanih asocijativnih pravila. Ovim se eliminišu pravila čija vrednost ZScore mere interesantnosti ne prelazi zadati parametar (minimalni ZScore).

Četvrti korak procesa je eliminacija asocijativnih pravila koja su očekivana u odnosu na kraća pravila koja takođe postoje u tabeli otkrivenih asocijativnih pravila. Eliminišu se ona pravila za koja u tabeli asocijativnih pravila postoji kraće pravilo u odnosu na koje je lokalna ZScore vrednost manja od zadatog parametra (minimalni lokalni ZScore).

Peti korak procesa je eliminacija pravila koja su očekivana u odnosu na opštije pravilo iste dužine. Pri tome je relacija „opštije pravilo“ definisana u odnosu na prethodno generisanu konceptnu hijerarhiju. Eliminišu se ona pravila za koja u tabeli asocijativnih pravila postoji opštije pravilo jednake dužine, u odnosu na koje je lokalna ZScore vrednost manja od zadatog parametra (minimalni lokalni ZScore).



Slika 5.9. Proces eliminacije neinteresantnih asocijativnih pravila na osnovu zadatih parametara

Dijagram klasa i njihovih ključnih elemenata koji učestvuju u procesu eliminacije neinteresantnih asocijativnih pravila na osnovu zadatih parametara prikazan je na slici 5.10.

Klasa App sadrži objekat klase RuleTable, koji sadrži sva otkrivena asocijativna pravila, kao što je opisano u prethodnom poglavlju. Pored toga, klasa App sadrži objekat klase KonceptTree, koji predstavlja konceptno stablo korišćeno u procesu prečišćavanja skupa asocijativnih pravila.

Konceptno stablo koje se koristi u okviru algoritma za eliminaciju neinteresantnih asocijativnih pravila formira se na osnovu kratkih asocijativnih pravila (čija leva i desna strane sadrže samo po jednu web stranicu), a čija confidence vrednost prelazi zadati parametar ConfMax.

Klasa App sadrži objekat klase ConceptTree, koji predstavlja generisano konceptno stablo. Objekat klase ConceptTree sadrži listu objekata klase ConceptEdge, koji predstavljaju grane konceptnog stabla.

Svaki objekat klase ConceptEdge predstavlja po jednu granu konceptnog stabla. Svakoj grani konceptnog stabla odgovara po jedno asocijativno pravilo čija je confidence vrednost veća od zadatog parametra (ConfMax). Objekat klase ConceptEdge sadrži reference na gornji i donji čvor te grane, koji su predstavljeni kao objekti klase ConceptNode. Gornji čvor grane odnosi se na levu stranu odgovarajućeg asocijativnog pravila, dok se donji čvor grane odnosi na desnu stranu istog asocijativnog pravila.

Svaki čvor konceptnog stabla predstavljen je kao objekat klase ConceptNode. On sadrži referencu na objekat klase Item, koji predstavlja odgovarajuću web stranicu. Svaki čvor konceptnog stabla može pripadati jednoj ili više grana, koje mogu sadržati referencu na njega.

Klasa RulePruning sadrži implementirane algoritme za prečišćavanje skupa otkrivenih asocijativnih pravila.

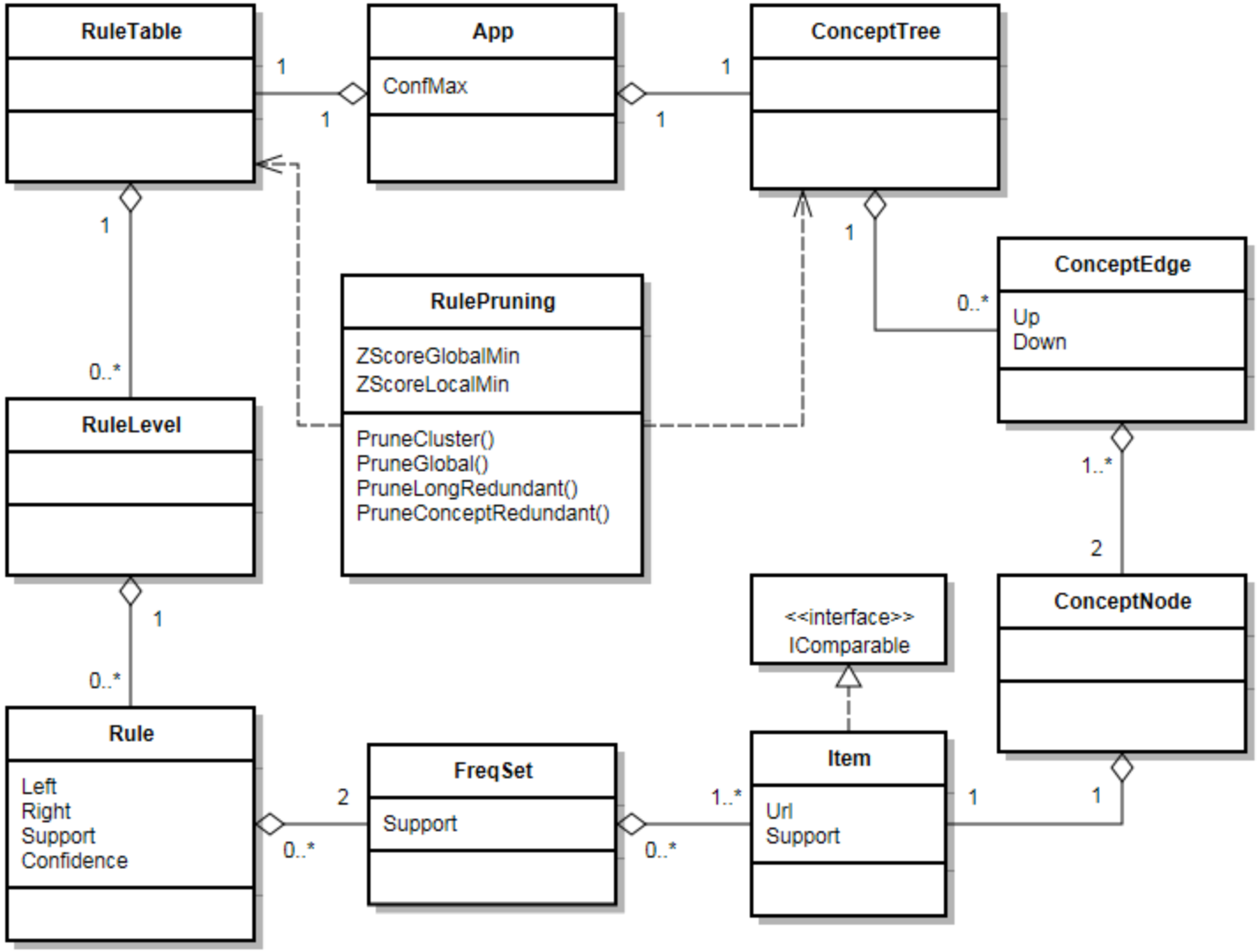
Metodom PruneCluster eliminišu se klaster asocijativna pravila. U okviru ovog algoritma pronalaze se klasteri u konceptnom stablu, takvi da postoji grana u oba pravca između svaka dva čvora klastera.

Metodom PruneGlobal eliminišu se asocijativna pravila statistički očekivana u skupu svih web sesija. Eliminišu se pravila imaju vrednost minimalne ZScore mere manju od zadatog parametra (ZScoreGlobalMin).

Metodom PruneLongRedundant eliminišu se asocijativna pravila koja su očekivana u odnosu na neko kraće i opštije pravilo. Pri tome se koristi tabela svih otkrivenih asocijativnih pravila (RuleTable) kako bi se za dato pravilo pronašla kraća i opštija pravila.

Metodom PruneConceptRedundant eliminišu se asocijativna pravila koja su očekivana u odnosu na neko opštije pravilo jednake dužine. Pri tome se koristi generisano konceptno stablo prilikom određivanja potencijalnih opštijih pravila.

U okviru algoritama implementiranih u metodama PruneLongRedundant i PruneConceptRedundant vrednost lokalne ZScore mere u odnosu na kraće i opštije pravilo (poglavlje 4.2.2) se izračunava na osnovu support vrednosti odgovarajućih frekventnih skupova. Pri tome se eliminišu pravila čija je vrednost lokalne ZScore mere u odnosu na neko opštije pravilo manja od zadatog parametra (ZScoreLocalMin).

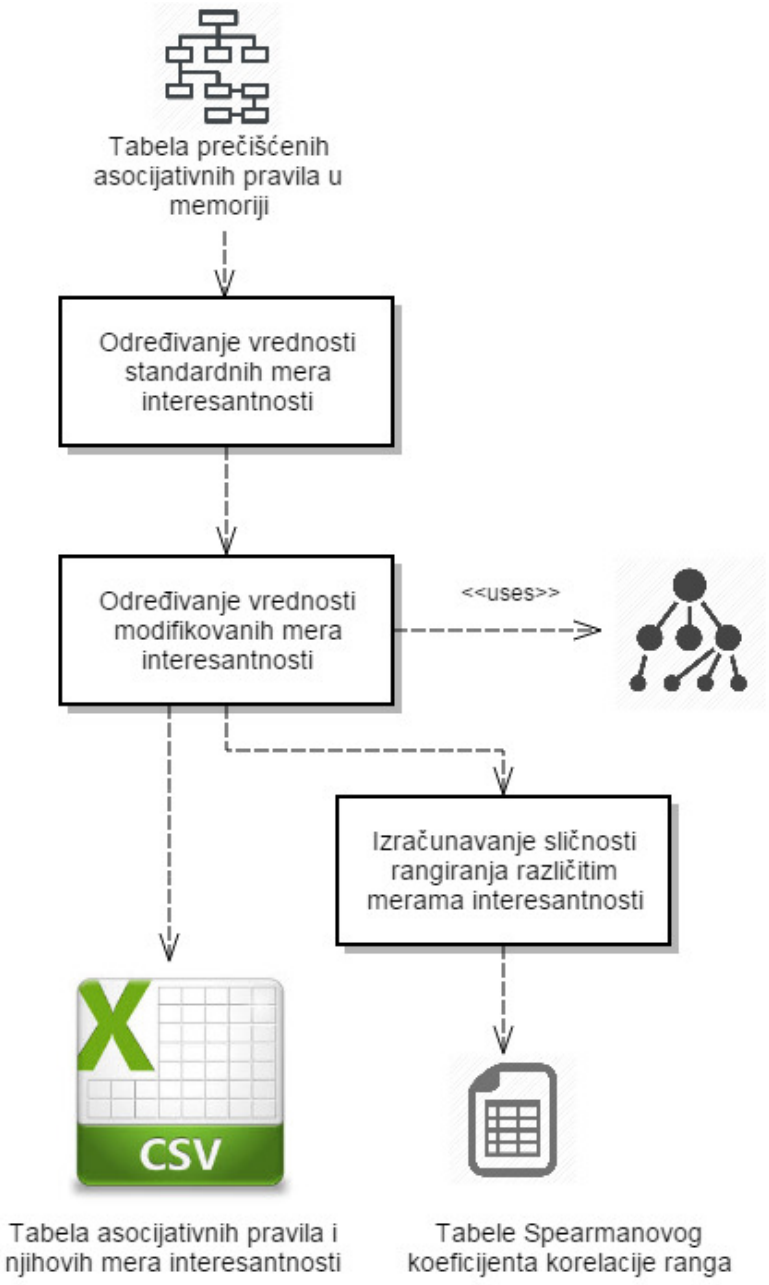


Slika 5.10. Dijagram klasa koje učestvuju u procesu eliminacije neinteresantnih asocijativnih pravila

5.6 Implementacija i ugradnja novih mera interesantnosti

Na slici 5.11 prikazan je proces implementacije novih mera interesantnosti ugrađenih u sistem. Tabela skupa prečišćenih asocijativnih pravila, iz koga su prethodno eliminisana neinteresantna, statistički očekivana asocijativna pravila koristi se u okviru ovog procesa. Vrši se izračunavanje vrednosti standardnih i modifikovanih mera interesantnosti, pri čemu se koristi prethodno generisana konceptna hijerarhija. Tabela asocijativnih pravila i njihovih izračunatih vrednosti mera interesantnosti se upisuje na disk u Excel formatu, kako bi se mogla koristiti u eksperimentalnom istraživanju.

Takođe se izračunava sličnost rangiranja asocijativnih pravila kada se koriste različite implementirane mere interesantnosti. Kao rezultat, na disku se snima datoteka koja sadrži vrednosti Spearman-ovog koeficijenta korelacije ranga za različite parove mera interesantnosti. Rezultati eksperimenata u kojima su prikazane vrednosti ovih koeficijenata za različite parove mera interesantnosti prikazani su u poglavlju 6.3.



Slika 5.11. Proces implementacije novih mera interesantnosti

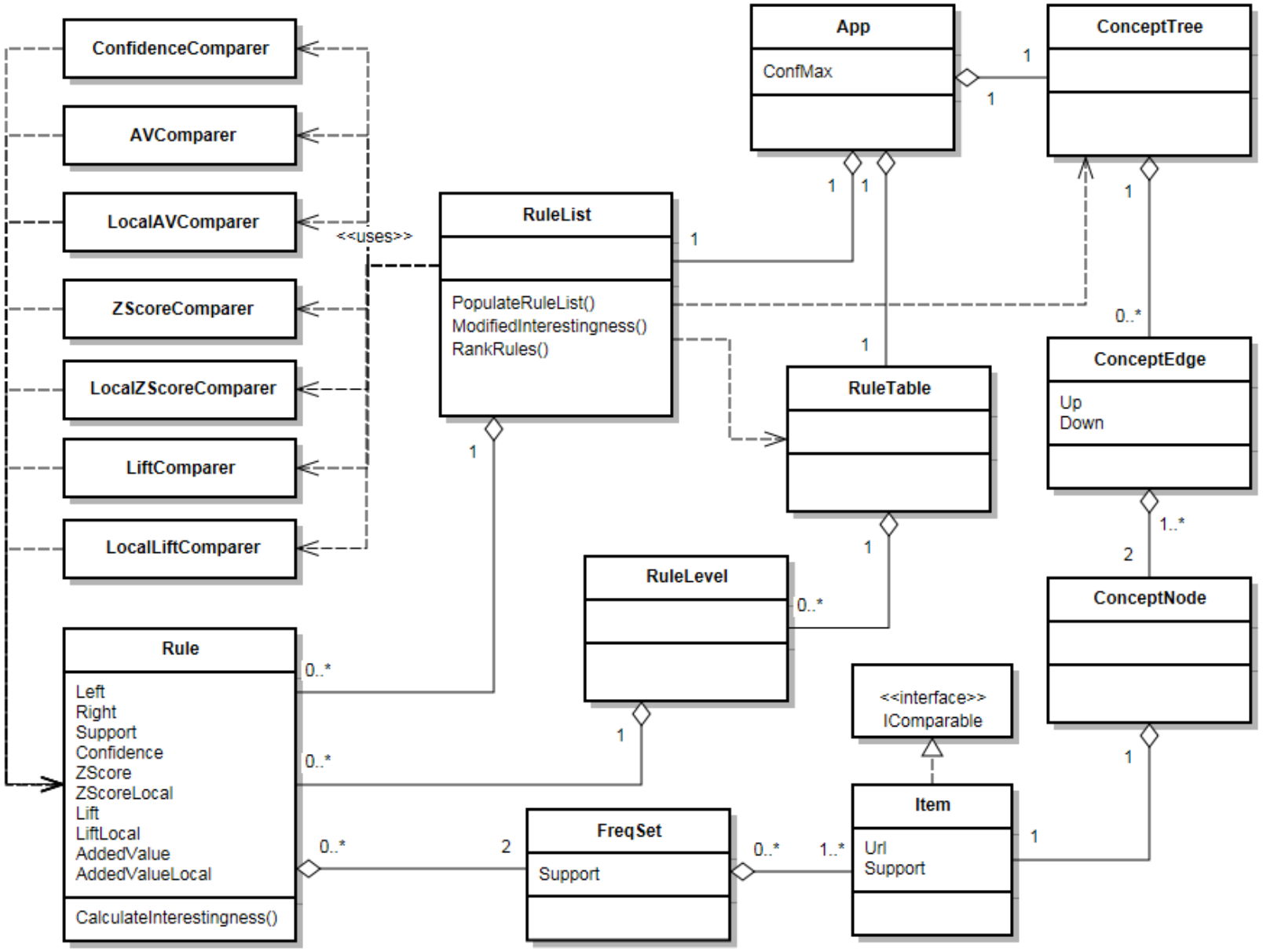
Na slici 5.12 prikazan je dijagram klasa i njihovih ključnih elemenata koje učestvuju u procesu implementacije i ugradnje novih mera interesantnosti.

U okviru objekta klase Rule čuvaju se vrednosti raznih statističkih mera interesantnosti odgovarajućeg asocijativnog pravila. Formule za izračunavanje standardnih statističkih mera interesantnosti koje se zasnivaju na tabelama kontigencije asocijativnog pravila (poglavlje 4.1.1) implementirane su u samoj klasi Rule. U okviru ovih formula koriste se Support vrednosti samog asocijativnog pravila, kao i frekventnih skupova leve i desne strane asocijativnog pravila.

Izračunavanje vrednosti modifikovanih mera interesantnosti za dato asocijativno pravilo oslanja se na prethodno generisanu konceptnu hijerarhiju (ConceptTree). U okviru klase RuleList implementiran je algoritam za izračunavanje vrednosti modifikovanih mera interesantnosti (ModifiedInteresingness). Pri tome se modifikuje interesantnost onih asocijativnih pravila čija su leva i desna strana potkoncepti zajedničkog natkoncepta.

Tokom prethodno realizovanog procesa generisanja svih pravila, formirana je tabela svih pravila (RuleTable) u okviru koje su pravila raspoređena prema nivoima dužine (lista objekata klase RuleLevel). Sada se formira lista svih pravila svih dužina (RuleList), koja sadrži niz referenci na sve objekte klase Rule.

Niz asocijativnih pravila u okviru objekta klase RuleList može se sortirati prema različitim kriterijumima, odnosno prema različitim merama interesantnosti. Prilikom sortiranja koriste se implementirane klase za poređenje asocijativnih pravila prema različitim kriterijumima (ConfidenceComparer, AVComparer, LocalAVComparer, ZScoreComparer, LocalZScoreComparer, LiftComparer, LocalLiftComparer). Tako se metodom RankRules klase RuleList vrši rangiranje asocijativnih pravila prema raznim merama interesantnosti i potom poredi koristeći Spearmanov koeficijent korelacije ranga. Rezultujuće tabela koeficijenta korelacije ranga se upisuju na disk i koriste u eksperimentalnom istraživanju.



Slika 5.12. Dijagram klasa koje učestvuju u procesu implementacije i ugradnje novih mera interesantnosti

5.7 Prikaz korisničkog interfejsa

Korisnički interfejs sistema za otkrivanje asocijativnih pravila o korišćenju web sajtova razvijen je za potrebe ovog eksperimentalnog istraživanja. Njime je implementirana osnovna interakcija sa korisnikom vezana za pretprocesiranje web log podataka, generisanje asocijativnih pravila, eliminisanje neinteresantnih pravila i rangiranje pravila prema različitim merama interesantnosti.

5.7.1 Korisnička forma za pretprocesiranje web log podataka

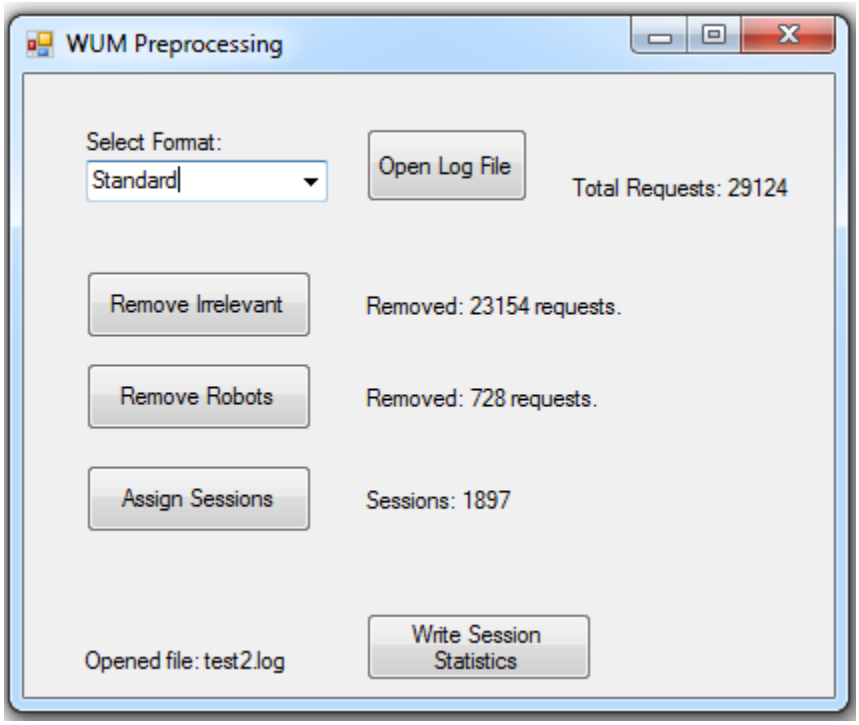
Na slici 5.13 prikazana je forma za pretprocesiranje web log podataka. Moguće je odabrati jedan od dva ponuđena formata web log podataka (opcija „Standard/Extended“) i otvoriti web log datoteku sa diska („Open log file“). Pri tome se vrši parsiranje web zahteva i prikazuje njihov ukupan broj („Total requests“).

Eliminacija irelevantnih web zahteva vrši se akcijom „Remove irrelevant“, posle čega se prikazuje broj eliminisanih irelevantnih web zahteva.

Akcijom „Remove robots“ eliminišu se robotski zahtevi i prikazuje njihov broj.

Akcijom „Assign sessions“ vrši se raspodela web zahteva u web sesije, prikazuje ukupan broj web sesija i kreira datoteka na disku, koja sadrži relevantne web zahteve raspoređene u web sesije. Ovako pripremljena datoteka može se koristiti za pronalaženje asocijativnih pravila.

Akcijom „Write session statistics“ formira se kontrolna datoteka na disku koja sadrži statističke podatke o formiranim web sesijama, kao što je ukupan broj web zahteva u svakoj web sesiji.



Slika 5.13. Korisnička forma za pretprocesiranje web log datoteteka

5.7.2 Korisnička forma za generisanje asocijativnih pravila

Korisnička forma za generisanje asocijativnih pravila (slika 5.14) omogućuje odabir parametara potrebnih za generisanje frekventnih skupova i asocijativnih pravila o korišćenju web sajtova.

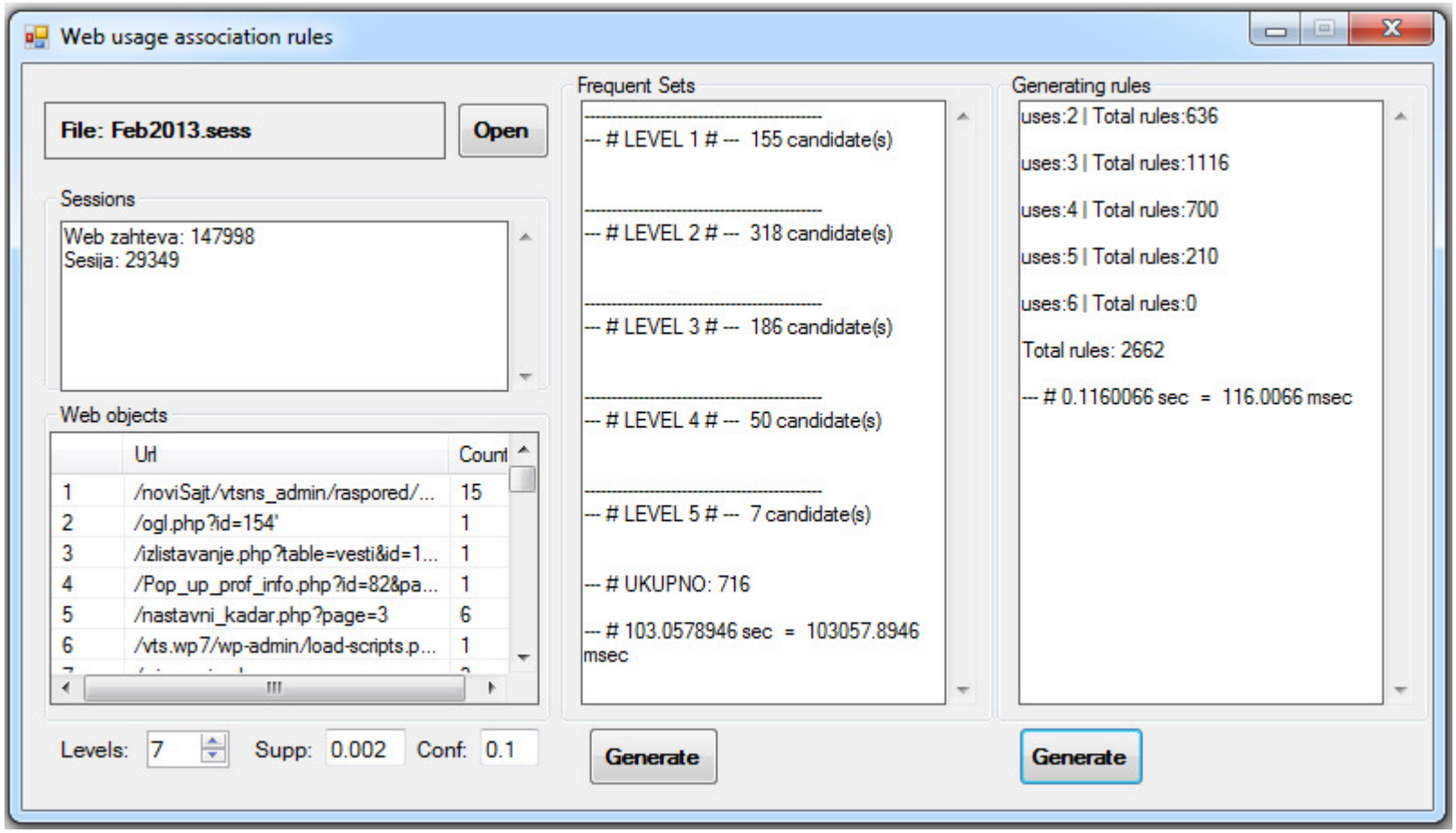
Akcija „Open“ omogućuje odabir datoteke koja sadrži skup web zahteva raspoređenih u web sesije, prethodno kreirane kao rezultat pretprocesiranja web log podataka. Nakon parsiranja datoteke u odeljku „Sessions“ prikazuje se ukupan broj web sesija i ukupan broj web zahteva u datoteci. Istovremeno, popunjava se lista „Web objects“ koja prikazuje sve web objekte koji se pojavljuju u bar jednoj web sesiji. Pri tome se prikazuje Url svakog web objetka i ukupan broj njegovog pojavljivanja u web sesijama.

Za generisanje frekventnih skupova koristi se parametar „Supp“, koji predstavlja minimalni support prag frekventnih skupova u web sesijama odabrane datoteke.

Parametar „Levels“ predstavlja maksimalni broj nivoa (maksimalnu dužinu) generisanih frekventnih skupova.

Akcija „Generate“ ispod odeljka „Frequents sets“ inicira proces generisanja frekventnih skupova. Kontrolni log prikazuje broj frekventnih skupova (candidates) pronađenih na svakom nivou dužine. Na kraju se prikazuje ukupno vreme potrebno za generisanje svih frekventnih skupova.

Akcija „Generate“ ispod odeljka „Generating rules“ inicira generisanje asocijativnih pravila na osnovu otkrivenih frekventnih skupova. Pri tome se koristi parametar „Conf“, koji predstavlja minimalni confidence prag koji svako otkriveno asocijativno pravilo mora preći. Kontrolni log prikazuje broj asocijativnih pravila raznih dužina. Na kraju se prikazuje ukupno vreme potrebno za generisanje svih asocijativnih pravila.



Slika 5.14. Korisnička forma za generisanje asocijativnih pravila

5.7.3 Korisnička forma za prečišćavanje i rangiranje asocijativnih pravila

Proces prečišćavanja i rangiranja asocijativnih pravila podržan je opcijama implementiranim na korisničkoj formi prikazanoj na slici 5.15.

Akcijom „Create“ u odeljku „Hierarchy“ inicira se kreiranje konceptne hijerarhije, na osnovu otkrivenih asocijativnih pravila. Pri tome se odabira vrednost parametra „Conf“, koji predstavlja minimalni confidence asocijativnih pravila koja učestvuju u kreiranju konceptne hijerarhije.

Odeljak „Limit“ služi za ograničavanje eksperimenata na kratka i/ili pozitivna asocijativna pravila. Akcija „Target“ briše sva non-target pravila, odnosno pravila koja imaju više od jedne web stranice sa desne strane pravila. Akcija „Positive“ briše sva negativna pravila. Pri tome se u odeljku „Pruning log“ prikazuje broj target/non-target, pozitivnih/negativnih i kratkih/dugih pravila.

Odeljak „Pruning parameters“ služi za eliminaciju neinteresantnih asocijativnih pravila.

Akcija „Cluster“ eliminiše sva „klaster“ asocijativna pravila.

Akcija „GlobalZ“ eliminiše sva asocijativna pravila statistički očekivana u skupu svih web sesija. Pri tome se eliminišu sva pravila čiji je Z-score manji od zadatog parametra „Z-global“.

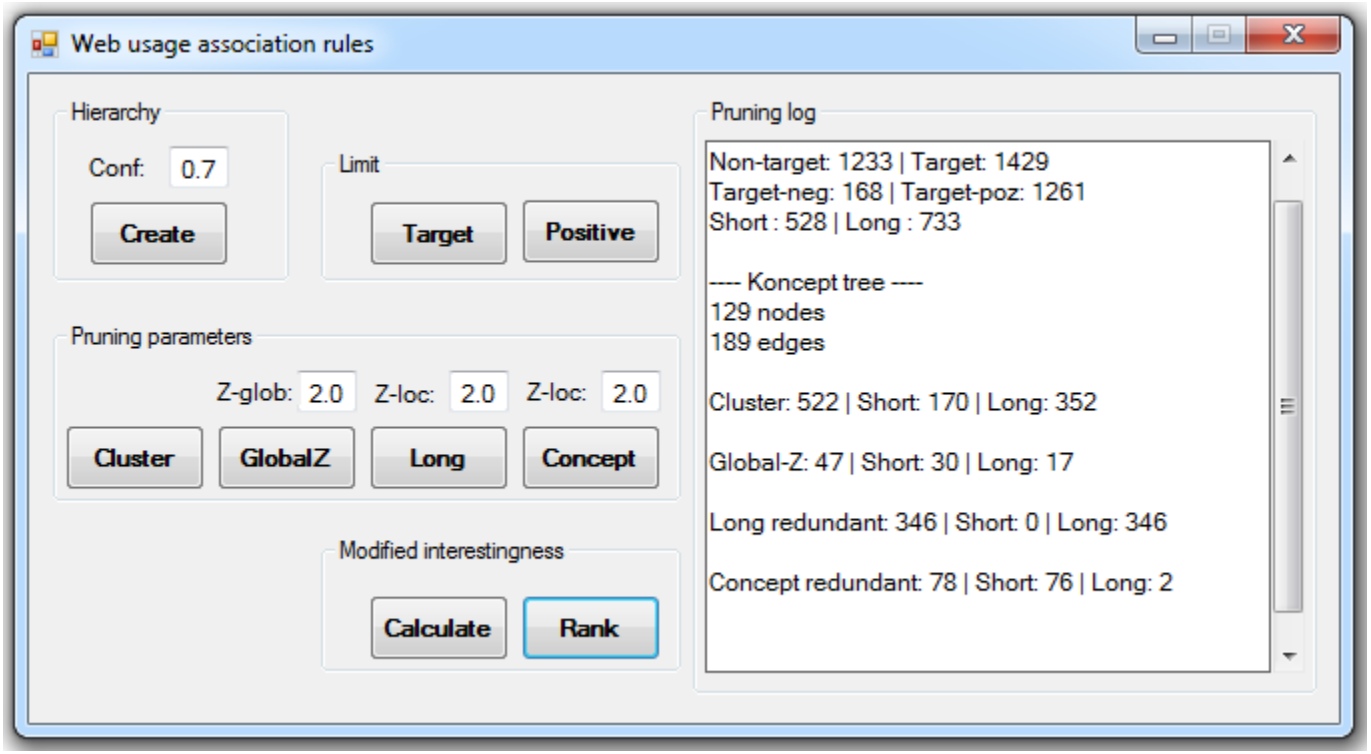
Akcija „Long“ eliminiše sva duga asocijativna pravila koja su očekivana u odnosu na kraća i opštija pravila. Pri tome se eliminišu sva pravila čiji je lokalni Z-score manji od zadatog parametra „Z-loc“.

Akcija „Concept“ eliminiše sva asocijativna pravila koja su očekivana u odnosu na opštija pravila u prisustvu konceptne hijerarhije. Pri tome se eliminišu sva pravila čiji je lokalni Z-score manji od zadatog parametra „Z-loc“.

Posle svake akcije eliminacije asocijativnih pravila u odeljku „Log“ se ispisuje broj eliminisanih kratkih i dugih pravila.

Odeljak „Modified interestingness“ odnosi se na izračunavanje modifikovanih mera interesantnosti za asocijativna pravila koja nisu prethodno eliminisana kao

neinteresantna. Akcija „Calculate“ inicira izračunavanje vrednosti modifikovanih mera interesantnosti za sva pravila čija su leva i desna strana potkoncept zajedničkog natkoncepta. Akcija „Rank“ inicira rangiranje asocijativnih pravila prema raznim izračunatim merama interesantnosti. Pri tome se tabele Spearman-ovog koeficijenta korelacije ranga upisuju u datoteku na disku. Eksperimenti u kojima su ove tabele prikazane za stvarne podatke o korišćenju web sajtova dati su u poglavlju 6.3.



Slika 5.15. Korisnička forma za prečišćavanje i rangiranje asocijativnih pravila

5.8 Testiranje sistema na eksperimentalnim podacima

Eksperimenti u kojima se ispituje efikasnost implementiranog softverskog sistema sprovedeni su na dva skupa realnih podataka o posetama web sajtovima dve visokoškolske ustanove – Visoke tehničke škole strukovnih studija u Novom Sadu (u daljem tekstu „VTŠ“) i Fakulteta organizacionih nauka u Beogradu (u daljem tekstu „FON“). Eksperimenti su izvršeni na laptop računaru sledeće konfiguracije: Intel Core(TM) i7-4500U CPU 1.8GHz, 8GB osnovne memorije, operativni sistem Windows 7, 64-bit, SP1.

5.8.1 Osnovne karakteristike eksperimentalnih skupova podataka

Eksperimentalni skupovi podataka preuzeti su sa web servera Visoke tehničke škole strukovnih studija u Novom Sadu i Fakulteta organizacionih nauka u Beogradu u obliku tekstualnih web log datoteka u standardnom „*W3C extended log file*“ formatu.

U tabeli 5.1 prikazane su osnovne karakteristike skupova podataka, uključujući period u kome su web zahtevi načinjeni, veličinu web log fajlova, kao i ukupni broj web zahteva pre njihovog prečišćavanja i obrade.

Oznaka	Web adresa	Period	Veličina	Web zahteva
FON	www.fon.bg.ac.rs	15.06.2013 – 15.07.2013	712,400 kB	2,867,700
VTŠ	www.vtsns.edu.rs	01.02.2013 – 31.04.2013	404,976 kB	1,871,267

Tabela 5.1. Karakteristike eksperimentalnih skupova podataka

U trenutku preuzimanja web log podataka, oba web sajta sadrže hijerarhijsku meni strukturu.

VTŠ web sajt sadrži jedan padajući hijerarhijski meni. Karakteristično je postojanje većeg broja pomoćnih web stranica koje služe samo da bi se sa njih pristupilo dokumentima kao što su podaci o polaganju ispita, rasporedu predavanja i ispita, kao i pojedinim vestima.

FON web sajt sadrži dva menija koja se ne menjaju tokom pretraživanja sajta – osnovni meni i meni u zaglavlju. Pored toga sajt sadrži levi i desni podmeni koji se menjaju u zavisnosti od web stranice koja je trenutno učitana.

VTŠ skup podataka obuhvata sve web zahteve upućene web serveru za VTŠ web sajt u periodu od tri meseca. U tom periodu u Visokoj tehničkoj školi strukovnih studija u Novom Sadu je bio održavan ispitni rok.

FON skup podataka obuhvata sve web zahteve upućene web serveru za FON web sajt u periodu od jednog meseca. U tom periodu na Fakultetu organizacionih nauka u Beogradu je bio održavan ispitni rok, kao i upis studenata u prvu godinu studija.

5.8.2 Priprema i prečišćavanje web log podataka

Tabela 5.2 sadrži rezultate pripreme i prečišćavanja FON i VTŠ skupova podataka. U tabeli je prikazan ukupan broj web zahteva („Web zahtevi ukupno“), broj web zahteva za irelevantnim web objektima kao što su slike („Irelevantni zahtevi“), broj web zahteva postavljanih od strane mašina za indeksiranje („Automatski zahtevi“), kao i broj relevantnih web zahteva koji preostaju posle eliminacije automatskih i irelevantnih web zahteva („Relevantni zahtevi“). Takođe je prikazan broj web sesija dobijenih spajanjem web zahteva u sesije („Ukupno sesija“), kao i broj relevantnih sesija („Relevantne sesije“) koje preostaju u skupu svih sesija posle eliminacije irelevantnih sesija (kratkih sesija, dugih sesija i sesija koje ne sadrže osnovnu stranicu web sajta.

Skup	Web zahtevi ukupno	Irelevantni zahtevi	Automatski zahtevi	Relevantni zahtevi	Ukupno sesija	Relevantne sesije
FON	2,867,700	2,314,876	67,681	485,143	203,380	74,841
VTŠ	1,871,267	1,331,476	135,876	403,915	66,746	32,091

Tabela 5.2. Rezultati prečišćavanja skupova podataka

Iako VTŠ skup podataka obuhvata period od tri meseca, dok FON skup podataka obuhvata period od samo jednog meseca, ukupan broj web sesija za FON skup podataka je znatno veći u odnosu na VTŠ skup podataka, jer je FON web sajt bio češće posećivan u odnosu na VTŠ web sajt u datom vremenskom periodu.

Udeo irelevantnih web zahteva u ukupnom broju web zahteva je nešto veći za FON web sajt u odnosu na VTŠ web sajt. Taj odnos zavisi od broja slika i drugih irelevantnih web objekata, kojih je nešto više na FON web sajtu u odnosu na VTŠ web sajt.

Ukupan broj automatskih web zahteva (generisanih od strane mašina za indeksiranje) je gotovo trostruko veći za VTŠ web sajt, upravo iz razloga što VTŠ skup podataka obuhvata trostruko duži vremenski period u odnosu na FON skup podataka, u toku koga su relativno ravnomerno pristizali automatski web zahtevi.

Udeo relevantnih sesija u odnosu na irelevantne je nešto veći za VTŠ web sajt u odnosu na FON web sajt. Na ovu razliku najviše utiče činjenica da je na VTŠ web sajtu veći broj sesija koje sadrže samo jednu web stranicu. Razlog za to je ponašanje korisnika web sajta, koji su u datom vremenskom periodu preko bookmark-a tražili samo određeni podatak na web sajtu (najčešće rezultate polaganja ispita). Takve web sesije ne utiču na formiranje asocijativnih pravila, te su eliminisane iz skupova podataka kao irelevantne.

Na ovako pripremljene podatke o posetama web sajtovima dve visokoškolske ustanove primenjeni su algoritmi za generisanje, prečišćavanje i rangiranje asocijativnih pravila i analiziran je kvalitet znanja dobijenog proširenim softverskim sistemom, što je prikazano u poglavlju 6.

6 Rezultati istraživanja

U ovom poglavlju analizira se efikasnost implementiranih metoda za povećanje kvaliteta otkrivenih asocijativnih pravila o korišćenju web sajtova.

6.1 Uticaj mera interesantnosti AP pri analizi web log podataka

Na prethodno pripremljenim i prečišćenim web log podacima izvršeno je generisanje asocijativnih pravila primenom Apriori algoritma implementiranog u okviru sistema za otkrivanje asocijativnih pravila i analiziran uticaj mera interesantnosti na kvalitet rezultata.

6.1.1 Uticaj support mere pri generisanju frekventnih skupova web stranica

U tabelama 6.1 i 6.2 dati su rezultati generisanja frekventnih skupova web stranica za različite vrednosti minimalnog support praga za skupove podataka VTŠ, odnosno FON. U koloni „Web stranice“ dat je broj frekventnih skupova web stranica, u koloni „Frekventni skupovi“ dat je ukupan broj svih frekventnih skupova za dati minimalni support prag, a vreme potrebno za njihovo generisanje dato je u koloni „Vreme“ izraženo u sekundama.

Support	Web stranice	Frekventni skupovi	Vreme (s)
0.0001	816	19805	8764.3
0.0005	476	3313	2323.8
0.001	330	1615	1095.7
0.002	183	720	405.1
0.005	75	217	157.5
0.01	43	95	121.6
0.1	10	13	102.4

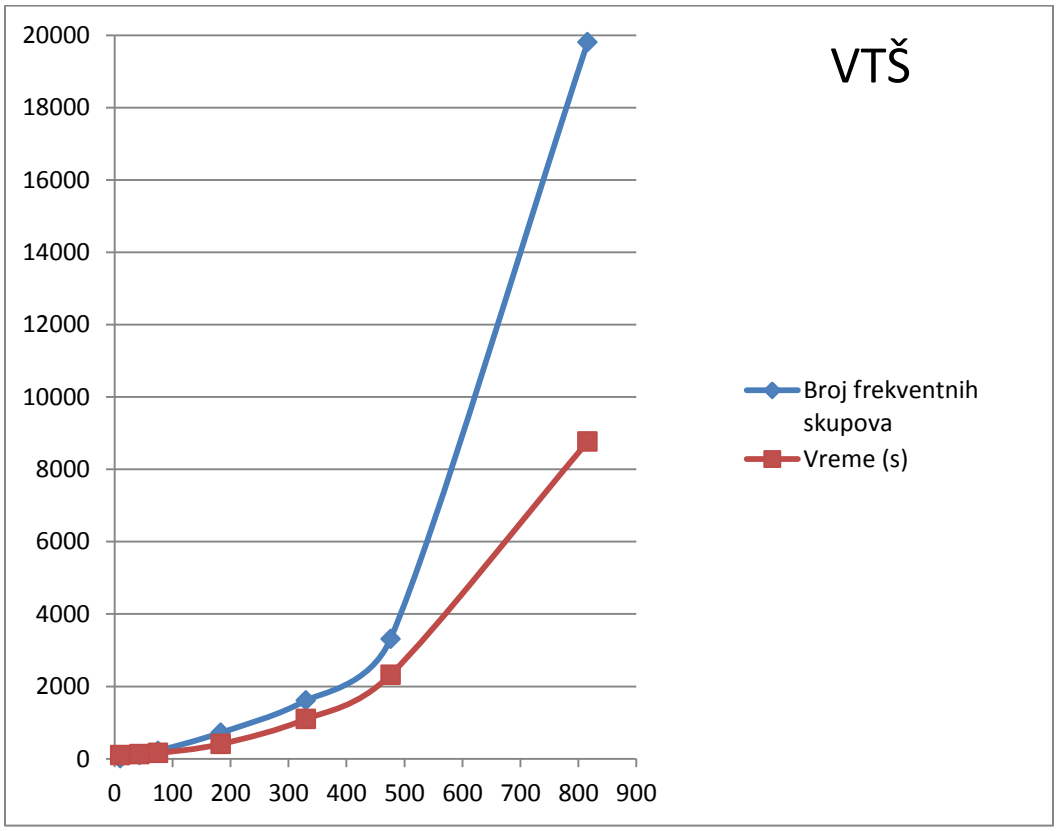
Tabela 6.1. Frekventni skupovi web stranica: VTŠ

Support	Web stranice	Frekventni skupovi	Vreme (s)
0.0001	281	7776	3447.1
0.0005	111	1498	558.5
0.001	80	769	284.7
0.002	56	360	148.5
0.005	38	153	89.8
0.01	28	70	69.0
0.1	5	0	52.9

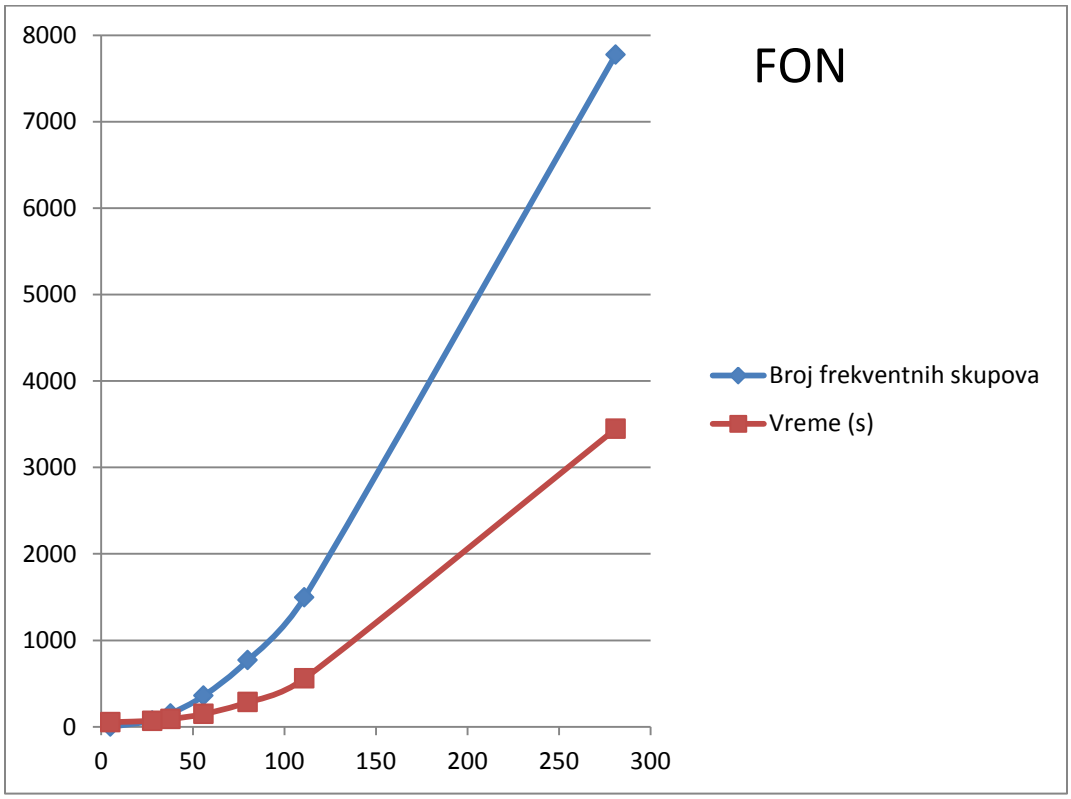
Tabela 6.2. Frekventni skupovi web stranica: FON

Prema podacima datim u tabelama 6.1 i 6.2, za VTŠ web sajt postoji veći broj frekventnih web stranica, kao i frekventnih skupova web stranica nego za FON web sajt, za sve vrednosti minimalnog support praga.

Grafikoni na slikama 6.1 i 6.2 generisani na osnovu podataka datih u tabelama 6.1 i 6.2, prikazuju trend rasta broja frekventnih skupova web stranica, kao i vremena potrebnog za njihovo generisanje, u odnosu na broj frekventnih web stranica, generisanih za vrednosti minimalnog support praga, koje su date u tabelama 6.3 i 6.4.



Slika 6.1. Broj frekventnih skupova web stranica i vreme njihovog generisanja u zavisnosti od broja frekventnih web stranica: VTŠ



Slika 6.2. Broj frekventnih skupova web stranica i vreme njihovog generisanja u zavisnosti od broja frekventnih web stranica: FON

Kao što je očekivano, za oba skupa podataka ukupan broj frekventnih skupova, kao i vreme potrebno za njihovo generisanje raste eksponencijalno sa brojem frekventnih web stranica, s tim što raste brže za FON skup podataka nego za VTŠ skup podataka.

U narednom poglavlju opisujemo rezultate generisanja svih asocijativnih pravila o korišćenju VTŠ i FON web sajtova, na osnovu frekventnih skupova web stranica za zadate vrednosti minimalnog support i confidence praga.

6.1.2 Uticaj confidence mere na generisanje asocijativnih pravila

Generisanje asocijativnih pravila na osnovu frekventnih skupova web stranica izvršeno je primenom algoritma implementiranog u okviru sistema za otkrivanje asocijativnih pravila. Minimalni confidence prag u svim eksperimentima je konstantan i iznosi 0.1, dok su odabrane tri vrednosti minimalnog support praga: 0.005, 0.01 i 0.02.

Pravila čija je confidence vrednost manja od 0.1 najčešće nisu interesantna analitičarima podataka, te su postavljanjem ovakvog minimalnog confidence praga eliminisana iz daljeg razmatranja. Smatramo da pravila čija je confidence vrednost veća od 0.1 mogu potencijalno biti interesantna analitičarima podataka, te ostaju u skupu svih pravila, a na njima se dalje primenjuju predložene metode prečišćavanja i rangiranja asocijativnih pravila.

Za tri odabrane vrednosti minimalnog support praga veličine skupova generisanih pravila, kao i veličine skupova sesija koje sadrže odgovarajuće frekventne skupove web stranica su dovoljno visoki, tako da su dalji eksperimenti statistički validni.

Iako su metode predložene u okviru ovog istraživanja teoretski primenljive na opšti oblik asocijativnih pravila, proveru njihovih efikasnosti u cilju unapređenja kvaliteta otkrivenih pravila vršimo na podskupu skupa svih pravila – skupu *pozitivnih target* pravila.

Pozitivna asocijativna pravila definisana su koristeći Z-score meru između leve i desne strane asocijativnog pravila. Pozitivna asocijativna pravila su ona pravila čija je vrednost Z-score mere veća ili jednaka sa -2. U protivnom smatramo da se leva i desna strana asocijativnog pravila „odbijaju“, odnosno da postoji negativna statistička korelacija među njima, te je takvo pravilo definisano kao „negativno“. Otkrivanje negativnih asocijativnih pravila je posebna oblast istraživanja, koja izlazi van okvira ove disertacije.

Target asocijativna pravila, koja sadrže samo jednu web stranicu sa desne strane, su jednostavnija za razumevanje i najčešće korišćena od strane analitičara podataka. Obzirom da je cilj ovog eksperimentalnog istraživanja povećanje upotrebljivosti skupa otkrivenih asocijativnih pravila od strane analitičara podataka, efikasnost predloženih

metoda za otkrivanje i eliminaciju asocijativnih pravila o korišćenju web sajtova merimo upravo na skupu pozitivnih target asocijativnih pravila.

U tabelama 6.3 i 6.4 dati su ukupni brojevi otkrivenih asocijativnih pravila za različite vrednosti minimalnog support praga za skupove podataka VTŠ, odnosno FON. U koloni „Ukupno pravila“ dat je ukupan broj generisanih asocijativnih pravila za datu vrednost minimalnog support praga. Zatim je dat broj target pravila („Target pravila“), broj pozitivnih target pravila („Target pozitivna“), kao i broj pozitivnih target pravila čija je confidence vrednost veća od 0.1. Vreme potrebno za generisanje skupova asocijativnih pravila za dati minimalni support prag dato je u koloni „Vreme“, izraženo u milisekundama.

Support	Ukupno pravila	Target pravila	Target pozitivna	Conf > 0.1	Vreme (ms)
0.0005	15658	7586	6834	5267	889.0
0.001	6078	3270	2897	2317	271.0
0.002	2254	1316	1121	915	93.6

Tabela 6.3. Generisanje skupova asocijativnih pravila: VTŠ

Support	Ukupno pravila	Target pravila	Target pozitivna	Conf > 0.1	Vreme (ms)
0.0005	10278	4113	3507	2764	309.0
0.001	4130	1909	1573	1276	224.0
0.002	1404	777	612	505	110.0

Tabela 6.4. Generisanje skupova asocijativnih pravila: FON

Kao što je poznato, broj otkrivenih asocijativnih pravila za različite vrednosti minimalnog support praga (tabele 6.3 i 6.4) ponaša se kao broj frekventnih skupova za date vrednosti minimalnog support praga (tabele 6.1 i 6.2). Dakle, broj otkrivenih asocijativnih pravila raste eksponencijalno sa brojem frekventnih web stranica, odnosno sa smanjenjem minimalnog support praga.

6.2 Implementacija i proširenje funkcionalnosti softverskog sistema za pronalaženje asocijativnih pravila dodavanjem novih mera interesantnosti

U ovom poglavlju dajemo rezultate eliminacije statistički očekivanih asocijativnih pravila primenom metoda predloženih u poglavljima 4.2 i 4.3, i to sledećim redosledom:

1. Eliminacija klaster-asocijativnih pravila
2. Eliminacija statistički očekivanih pravila u skupu svih sesija
3. Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila manje dužine
4. Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila jednake dužine, i u prisustvu konceptne hijerarhije web stranica

Metode eliminacije primenjene u koracima 1. i 2. su algoritamski manje zahtevne u odnosu na metode primenjene u koracima 3. i 4., a njihovom primenom se eliminišu očigledno trivijalna asocijativna pravila. Metode primenjene u koracima 3. i 4. se u eksperimentima primenjuju na skupu asocijativnih pravila prethodno prečišćenim eliminacijom očigledno trivijalnih asocijativnih pravila u koracima 1. i 2. Na ovaj način meri se efikasnost metoda 3. i 4. u cilju dodatnog smanjenja veličine skupa asocijativnih pravila kada se na njega prethodno primene elementarnije metode eliminacije u koracima 1. i 2.

Rezultati primene svih metoda eliminacije dati su u tabelama 6.5 i 6.6 za VTŠ skup podataka, i u tabelama 6.7 i 6.8 za FON skup podataka. U eksperimentima čiji su rezultati dati u tabelama 6.5 i 6.7, vrednost minimalnog Z-score praga postavljena je na 2.0, dok je u eksperimentima čiji su rezultati dati u tabelama 6.6 i 6.8, ova vrednost postavljena na 4.0.

Conf	Kr	Dug	Kr %	Kla-ster kr	Kla-ster dug	Kla-ster uk %	SO kr	SO dug	SO uk %	Opšt dug	Opšt dug %	Opšt konc kr	Opšt konc kr %	Opšt konc dug	Opšt konc dug %	Elim dug uk	Elim dug uk %	Elim kr uk	Elim kr uk %	Elim uk %
Support threshold = 0.0005																				
0.1-0.2	220	399	35.5	38	90	20.7	3	8	2.2	274	91.0	59	33.0	6	22.2	378	94.7	100	45.5	77.2
0.2-0.3	158	402	28.2	41	129	30.4	21	49	17.9	199	88.8	28	29.2	3	12.0	380	94.5	90	57.0	83.9
0.3-0.4	98	300	24.6	13	67	20.1	30	50	25.2	164	89.6	14	25.5	0	0.0	281	93.7	57	58.2	84.9
0.4-0.5	70	248	22.0	18	56	23.3	8	29	15.2	139	85.3	7	15.9	0	0.0	224	90.3	33	47.1	80.8
0.5-0.6	34	138	19.8	4	33	21.5	3	7	7.4	86	87.8	6	22.2	0	0.0	126	91.3	13	38.2	80.8
0.6-0.7	35	143	19.7	3	22	14.0	1	4	3.3	102	87.2	3	9.7	0	0.0	128	89.5	7	20.0	75.8
0.7-0.8	38	160	19.2	27	71	49.5	0	0	0.0	84	94.4	1	9.1	0	0.0	155	96.9	28	73.7	92.4
0.8-0.9	53	181	22.6	25	111	58.1	0	0	0.0	67	95.7	3	10.7	0	0.0	178	98.3	28	52.8	88.0
0.9-1.0	598	1992	23.1	265	1084	52.1	0	0	0.0	898	98.9	76	22.8	0	0.0	1982	99.5	341	57.0	89.7
Ukupno	1304	3963	24.8	434	1663	39.8	66	147	6.7	2013	93.5	197	24.5	9	6.4	3832	96.7	697	53.5	86.0
Support threshold = 0.001																				
0.1-0.2	112	127	46.9	14	20	14.2	1	2	1.5	92	87.6	23	23.7	3	23.1	117	92.1	38	33.9	64.9
0.2-0.3	67	120	35.8	10	29	20.9	4	14	12.2	65	84.4	13	24.5	2	16.7	110	91.7	27	40.3	73.3
0.3-0.4	56	134	29.5	3	17	10.5	17	26	25.3	79	86.8	8	22.2	0	0.0	122	91.0	28	50.0	78.9
0.4-0.5	31	89	25.8	5	14	15.8	3	5	7.9	55	78.6	5	21.7	0	0.0	74	83.1	13	41.9	72.5
0.5-0.6	25	70	26.3	3	15	18.9	1	4	6.5	44	86.3	4	19.0	0	0.0	63	90.0	8	32.0	74.7
0.6-0.7	24	78	23.5	2	13	14.7	0	0	0.0	53	81.5	2	9.1	0	0.0	66	84.6	4	16.7	68.6
0.7-0.8	25	59	29.8	20	28	57.1	0	0	0.0	28	90.3	0	0.0	0	0.0	56	94.9	20	80.0	90.5
0.8-0.9	27	48	36.0	13	29	56.0	0	0	0.0	17	89.5	3	21.4	0	0.0	46	95.8	16	59.3	82.7
0.9-1.0	409	816	33.4	194	384	47.2	0	0	0.0	428	99.1	46	21.4	0	0.0	812	99.5	240	58.7	85.9
Ukupno	776	1541	33.5	264	549	35.1	26	51	5.1	861	91.5	104	21.4	5	6.3	1466	95.1	394	50.8	80.3
Support threshold = 0.002																				
0.1-0.2	65	39	62.5	6	7	12.5	1	1	2.2	26	83.9	9	15.5	1	20.0	35	89.7	16	24.6	49.0
0.2-0.3	41	32	56.2	6	10	21.9	1	2	5.3	15	75.0	3	8.8	2	40.0	29	90.6	10	24.4	53.4
0.3-0.4	27	40	40.3	2	5	10.4	8	4	20.0	25	80.6	1	5.9	0	0.0	34	85.0	11	40.7	67.2
0.4-0.5	15	36	29.4	2	8	19.6	1	3	9.8	20	80.0	1	8.3	0	0.0	31	86.1	4	26.7	68.6
0.5-0.6	12	20	37.5	0	2	6.3	0	1	3.3	16	94.1	4	33.3	0	0.0	19	95.0	4	33.3	71.9
0.6-0.7	16	38	29.6	0	7	13.0	0	0	0.0	28	90.3	2	12.5	0	0.0	35	92.1	2	12.5	68.5
0.7-0.8	5	17	22.7	4	5	40.9	0	0	0.0	10	83.3	0	0.0	0	0.0	15	88.2	4	80.0	86.4
0.8-0.9	10	17	37.0	6	12	66.7	0	0	0.0	3	60.0	0	0.0	0	0.0	15	88.2	6	60.0	77.8
0.9-1.0	184	301	37.9	78	142	45.4	0	0	0.0	157	98.7	16	15.1	0	0.0	299	99.3	94	51.1	81.0
Ukupno	375	540	41.0	104	198	33.0	11	11	3.6	300	90.6	36	13.8	3	9.7	512	94.8	151	40.3	72.5

Tabela 6.5. Rezultati primene metoda eliminacije: VTŠ, Z=2.0

Conf	Kr	Dug	Kr %	Kla-ster kr	Kla-ster dug	Kla-ster uk %	SO kr	SO dug	SO uk %	Opšt dug	Opšt dug %	Opšt konc kr	Opšt konc kr %	Opšt konc dug	Opšt konc dug %	Elim dug uk	Elim dug uk %	Elim kr uk	Elim kr uk %	Elim uk %
Support threshold = 0.0005																				
0.1-0.2	220	399	35.5	38	90	20.7	11	30	8.4	273	97.8	60	35.1	1	16.7	394	98.7	109	49.5	81.3
0.2-0.3	158	402	28.2	41	129	30.4	23	54	19.7	212	96.8	28	29.8	1	14.3	396	98.5	92	58.2	87.1
0.3-0.4	98	300	24.6	13	67	20.1	32	50	25.8	174	95.1	18	34.0	0	0.0	291	97.0	63	64.3	88.9
0.4-0.5	70	248	22.0	18	56	23.3	10	42	21.3	143	95.3	10	23.8	0	0.0	241	97.2	38	54.3	87.7
0.5-0.6	34	138	19.8	4	33	21.5	5	26	23.0	76	96.2	5	20.0	0	0.0	135	97.8	14	41.2	86.6
0.6-0.7	35	143	19.7	3	22	14.0	6	21	17.6	97	97.0	1	3.8	0	0.0	140	97.9	10	28.6	84.3
0.7-0.8	38	160	19.2	27	71	49.5	1	21	22.0	68	100.0	2	20.0	0	0.0	160	100.0	30	78.9	96.0
0.8-0.9	53	181	22.6	25	111	58.1	1	2	3.1	65	95.6	5	18.5	0	0.0	178	98.3	31	58.5	89.3
0.9-1.0	598	1992	23.1	265	1084	52.1	4	51	4.4	853	99.5	74	22.5	0	0.0	1988	99.8	343	57.4	90.0
Ukupno	1304	3963	24.8	434	1663	39.8	93	297	12.3	1961	97.9	203	26.1	2	4.8	3923	99.0	730	56.0	88.3
Support threshold = 0.001																				
0.1-0.2	112	127	46.9	14	20	14.2	4	5	4.4	97	95.1	22	23.4	0	0.0	122	96.1	40	35.7	67.8
0.2-0.3	67	120	35.8	10	29	20.9	4	14	12.2	72	93.5	13	24.5	1	20.0	116	96.7	27	40.3	76.5
0.3-0.4	56	134	29.5	3	17	10.5	19	26	26.5	86	94.5	10	29.4	0	0.0	129	96.3	32	57.1	84.7
0.4-0.5	31	89	25.8	5	14	15.8	4	17	20.8	53	91.4	5	22.7	0	0.0	84	94.4	14	45.2	81.7
0.5-0.6	25	70	26.3	3	15	18.9	2	11	16.9	42	95.5	3	15.0	0	0.0	68	97.1	8	32.0	80.0
0.6-0.7	24	78	23.5	2	13	14.7	4	15	21.8	49	98.0	1	5.6	0	0.0	77	98.7	7	29.2	82.4
0.7-0.8	25	59	29.8	20	28	57.1	0	1	2.8	30	100.0	0	0.0	0	0.0	59	100.0	20	80.0	94.0
0.8-0.9	27	48	36.0	13	29	56.0	0	0	0.0	17	89.5	4	28.6	0	0.0	46	95.8	17	63.0	84.0
0.9-1.0	409	816	33.4	194	384	47.2	0	0	0.0	429	99.3	46	21.4	0	0.0	813	99.6	240	58.7	86.0
Ukupno	776	1541	33.5	264	549	35.1	37	89	8.4	875	96.9	104	21.9	1	3.6	1514	98.2	405	52.2	82.8
Support threshold = 0.002																				
0.1-0.2	65	39	62.5	6	7	12.5	3	1	4.4	27	87.1	8	14.3	0	0.0	35	89.7	17	26.2	50.0
0.2-0.3	41	32	56.2	6	10	21.9	1	2	5.3	17	85.0	3	8.8	1	33.3	30	93.8	10	24.4	54.8
0.3-0.4	27	40	40.3	2	5	10.4	10	4	23.3	29	93.5	1	6.7	0	0.0	38	95.0	13	48.1	76.1
0.4-0.5	15	36	29.4	2	8	19.6	2	8	24.4	17	85.0	0	0.0	0	0.0	33	91.7	4	26.7	72.5
0.5-0.6	12	20	37.5	0	2	6.3	1	2	10.0	16	100.0	3	27.3	0	0.0	20	100.0	4	33.3	75.0
0.6-0.7	16	38	29.6	0	7	13.0	4	4	17.0	26	96.3	1	8.3	0	0.0	37	97.4	5	31.3	77.8
0.7-0.8	5	17	22.7	4	5	40.9	0	0	0.0	12	100.0	0	0.0	0	0.0	17	100.0	4	80.0	95.5
0.8-0.9	10	17	37.0	6	12	66.7	0	0	0.0	3	60.0	0	0.0	0	0.0	15	88.2	6	60.0	77.8
0.9-1.0	184	301	37.9	78	142	45.4	0	0	0.0	157	98.7	16	15.1	0	0.0	299	99.3	94	51.1	81.0
Ukupno	375	540	41.0	104	198	33.0	21	21	6.9	304	94.7	32	12.8	1	5.9	524	97.0	157	41.9	74.4

Tabela 6.6. Rezultati primene metoda eliminacije: VTŠ, Z=4.0

Conf	Kr	Dug	Kr %	Kla-ster kr	Kla-ster dug	Kla-ster uk %	SO kr	SO dug	SO uk %	Opšt dug	Opšt dug %	Opšt konc kr	Opšt konc kr %	Opšt konc dug	Opšt konc dug %	Elim dug uk	Elim dug uk %	Elim kr uk	Elim kr uk %	Elim uk %
Support threshold = 0.0005																				
0.1-0.2	106	525	16.8	0	0	0.0	3	38	6.5	394	80.9	30	29.1	6	6.5	438	83.4	33	31.1	74.6
0.2-0.3	42	327	11.4	0	0	0.0	5	43	13.0	215	75.7	3	8.1	5	7.2	263	80.4	8	19.0	73.4
0.3-0.4	33	268	11.0	1	1	0.7	2	31	11.0	175	74.2	4	13.3	1	1.6	208	77.6	7	21.2	71.4
0.4-0.5	31	265	10.5	0	0	0.0	4	47	17.2	172	78.9	2	7.4	0	0.0	219	82.6	6	19.4	76.0
0.5-0.6	12	179	6.3	1	0	0.5	0	12	6.3	134	80.2	0	0.0	0	0.0	146	81.6	1	8.3	77.0
0.6-0.7	9	193	4.5	0	0	0.0	0	0	0.0	161	83.4	0	0.0	0	0.0	161	83.4	0	0.0	79.7
0.7-0.8	14	153	8.4	3	2	3.0	0	0	0.0	134	88.7	1	9.1	0	0.0	136	88.9	4	28.6	83.8
0.8-0.9	20	156	11.4	2	2	2.3	0	0	0.0	153	99.4	2	11.1	0	0.0	155	99.4	4	20.0	90.3
0.9-1.0	44	387	10.2	5	14	4.4	0	0	0.0	370	99.2	4	10.3	0	0.0	384	99.2	9	20.5	91.2
Ukupno	311	2453	11.3	12	19	1.1	14	171	6.8	1908	84.3	46	16.1	12	3.4	2110	86.0	72	23.2	78.9
Support threshold = 0.001																				
0.1-0.2	73	222	24.7	0	0	0.0	1	13	4.7	154	73.7	17	23.6	3	5.5	170	76.6	18	24.7	63.7
0.2-0.3	31	142	17.9	0	0	0.0	4	17	12.1	91	72.8	1	3.7	2	5.9	110	77.5	5	16.1	66.5
0.3-0.4	23	100	18.2	1	1	1.6	2	10	9.9	52	59.1	2	10.5	0	0.0	63	63.6	5	22.7	55.3
0.4-0.5	22	105	17.3	0	0	0.0	3	16	15.0	64	71.9	1	5.3	0	0.0	80	76.2	4	18.2	66.1
0.5-0.6	9	87	9.4	0	0	0.0	0	3	3.1	61	72.6	0	0.0	0	0.0	64	73.6	0	0.0	66.7
0.6-0.7	8	88	8.3	0	0	0.0	0	0	0.0	61	69.3	0	0.0	0	0.0	61	69.3	0	0.0	63.5
0.7-0.8	8	53	13.1	0	0	0.0	0	0	0.0	44	83.0	1	12.5	0	0.0	44	83.0	1	12.5	73.8
0.8-0.9	14	69	16.9	0	0	0.0	0	0	0.0	68	98.6	2	14.3	0	0.0	68	98.6	2	14.3	84.3
0.9-1.0	32	190	13.6	4	12	7.2	0	0	0.0	175	98.3	4	14.3	0	0.0	187	98.4	8	25.0	87.8
Ukupno	220	1056	17.2	5	13	1.4	10	59	5.5	770	78.3	28	13.7	5	2.3	847	80.2	43	19.5	69.7
Support threshold = 0.002																				
0.1-0.2	49	61	44.5	0	0	0.0	1	6	6.4	38	69.1	8	16.7	1	5.9	45	73.8	9	18.4	49.1
0.2-0.3	23	51	31.1	0	0	0.0	3	5	10.8	31	67.4	1	5.0	1	6.7	37	72.5	4	17.4	55.4
0.3-0.4	14	37	27.5	0	0	0.0	1	3	7.8	16	47.1	2	15.4	0	0.0	19	51.4	3	21.4	43.1
0.4-0.5	16	39	29.1	0	0	0.0	1	6	12.7	23	69.7	0	0.0	0	0.0	29	74.4	1	6.3	54.5
0.5-0.6	7	28	20.0	0	0	0.0	0	0	0.0	17	60.7	0	0.0	0	0.0	17	60.7	0	0.0	48.6
0.6-0.7	6	28	17.6	0	0	0.0	0	0	0.0	16	57.1	0	0.0	0	0.0	16	57.1	0	0.0	47.1
0.7-0.8	4	14	22.2	0	0	0.0	0	0	0.0	9	64.3	1	25.0	0	0.0	9	64.3	1	25.0	55.6
0.8-0.9	11	27	28.9	0	0	0.0	0	0	0.0	26	96.3	1	9.1	0	0.0	26	96.3	1	9.1	71.1
0.9-1.0	14	76	15.6	0	0	0.0	0	0	0.0	73	96.1	2	14.3	0	0.0	73	96.1	2	14.3	83.3
Ukupno	144	361	28.5	0	0	0.0	6	20	5.1	249	73.0	15	10.9	2	2.2	271	75.1	21	14.6	57.8

Tabela 6.7. Rezultati primene metoda eliminacije: FON, Z=2.0

Conf	Kr	Dug	Kr %	Kla-ster kr	Kla-ster dug	Kla-ster uk %	SO kr	SO dug	SO uk %	Opšt dug	Opšt dug %	Opšt konc kr	Opšt konc kr %	Opšt konc dug	Opšt konc dug %	Elim dug uk	Elim dug uk %	Elim kr uk	Elim kr uk %	Elim uk %
Support threshold = 0.0005																				
0.1-0.2	106	525	16.8	0	0	0.0	6	52	9.2	434	91.8	37	37.0	1	2.6	487	92.8	43	40.6	84.0
0.2-0.3	42	327	11.4	0	0	0.0	9	67	20.6	230	88.5	3	9.1	2	6.7	299	91.4	12	28.6	84.3
0.3-0.4	33	268	11.0	1	1	0.7	3	48	17.1	195	89.0	4	13.8	0	0.0	244	91.0	8	24.2	83.7
0.4-0.5	31	265	10.5	0	0	0.0	4	48	17.6	199	91.7	6	22.2	0	0.0	247	93.2	10	32.3	86.8
0.5-0.6	12	179	6.3	1	0	0.5	0	22	11.6	149	94.9	0	0.0	0	0.0	171	95.5	1	8.3	90.1
0.6-0.7	9	193	4.5	0	0	0.0	0	17	8.4	164	93.2	1	11.1	0	0.0	181	93.8	1	11.1	90.1
0.7-0.8	14	153	8.4	3	2	3.0	0	2	1.2	140	94.0	1	9.1	0	0.0	144	94.1	4	28.6	88.6
0.8-0.9	20	156	11.4	2	2	2.3	0	0	0.0	154	100.0	2	11.1	0	0.0	156	100.0	4	20.0	90.9
0.9-1.0	44	387	10.2	5	14	4.4	0	0	0.0	372	99.7	4	10.3	0	0.0	386	99.7	9	20.5	91.6
Ukupno	311	2453	11.3	12	19	1.1	22	256	10.2	2037	93.5	58	20.9	3	2.1	2315	94.4	92	29.6	87.1
Support threshold = 0.001																				
0.1-0.2	73	222	24.7	0	0	0.0	2	19	7.1	181	89.2	25	35.2	1	4.5	201	90.5	27	37.0	77.3
0.2-0.3	31	142	17.9	0	0	0.0	6	28	19.7	95	83.3	1	4.0	1	5.3	124	87.3	7	22.6	75.7
0.3-0.4	23	100	18.2	1	1	1.6	2	11	10.7	71	80.7	2	10.0	0	0.0	83	83.0	5	21.7	71.5
0.4-0.5	22	105	17.3	0	0	0.0	3	17	15.7	77	87.5	3	15.8	0	0.0	94	89.5	6	27.3	78.7
0.5-0.6	9	87	9.4	0	0	0.0	0	11	11.5	70	92.1	0	0.0	0	0.0	81	93.1	0	0.0	84.4
0.6-0.7	8	88	8.3	0	0	0.0	0	3	3.1	73	85.9	1	12.5	0	0.0	76	86.4	1	12.5	80.2
0.7-0.8	8	53	13.1	0	0	0.0	0	0	0.0	47	88.7	1	12.5	0	0.0	47	88.7	1	12.5	78.7
0.8-0.9	14	69	16.9	0	0	0.0	0	0	0.0	69	100.0	2	14.3	0	0.0	69	100.0	2	14.3	85.5
0.9-1.0	32	190	13.6	4	12	7.2	0	0	0.0	177	99.4	4	14.3	0	0.0	189	99.5	8	25.0	88.7
Ukupno	220	1056	17.2	5	13	1.4	13	89	8.1	860	90.1	39	19.3	2	2.1	964	91.3	57	25.9	80.0
Support threshold = 0.002																				
0.1-0.2	49	61	44.5	0	0	0.0	2	8	9.1	47	88.7	13	27.7	1	16.7	56	91.8	15	30.6	64.5
0.2-0.3	23	51	31.1	0	0	0.0	5	11	21.6	34	85.0	1	5.6	0	0.0	45	88.2	6	26.1	68.9
0.3-0.4	14	37	27.5	0	0	0.0	1	3	7.8	23	67.6	2	15.4	0	0.0	26	70.3	3	21.4	56.9
0.4-0.5	16	39	29.1	0	0	0.0	1	7	14.5	26	81.3	1	6.7	0	0.0	33	84.6	2	12.5	63.6
0.5-0.6	7	28	20.0	0	0	0.0	0	3	8.6	22	88.0	0	0.0	0	0.0	25	89.3	0	0.0	71.4
0.6-0.7	6	28	17.6	0	0	0.0	0	0	0.0	20	71.4	1	16.7	0	0.0	20	71.4	1	16.7	61.8
0.7-0.8	4	14	22.2	0	0	0.0	0	0	0.0	10	71.4	1	25.0	0	0.0	10	71.4	1	25.0	61.1
0.8-0.9	11	27	28.9	0	0	0.0	0	0	0.0	27	100.0	1	9.1	0	0.0	27	100.0	1	9.1	73.7
0.9-1.0	14	76	15.6	0	0	0.0	0	0	0.0	75	98.7	2	14.3	0	0.0	75	98.7	2	14.3	85.6
Ukupno	144	361	28.5	0	0	0.0	9	32	8.1	284	86.3	22	16.3	1	2.2	317	87.8	31	21.5	68.9

Tabela 6.8. Rezultati primene metoda eliminacije: FON, Z=4.0

Sva pravila generisana u eksperimentima podeljena su prema intervalima confidence vrednosti kojima pripadaju, a koji su u tabelama 6.5, 6.6, 6.7 i 6.8 prikazani u koloni „Conf“. Rezultati eliminacije skupa svih pravila generisanih za određeni minimalni support prag prikazani su u redu „Ukupno“. Rezultati eliminacije skupa pravila koja pripadaju određenom confidence intervalu prikazani su u onom redu u tabeli, čiji je confidence interval označen u koloni „Conf“.

Kolona „Kr“ odnosi se na broj target pozitivnih kratkih pravila, dok se kolona „Dug“ odnosi na broj target pozitivnih dugih pravila, generisanih za odgovarajući minimalni support prag i confidence interval. Kolona „Kr %“ odnosi se na procenat kratkih pravila, u odnosu na ukupni broj pravila.

Kolone „Klaster kr“ i „Klaster dug“ odnose se na broj eliminisanih kratkih, odnosno dugih, klaster-asocijativnih pravila. Kolona „Klaster uk %“ odnosi se na procenat ukupno eliminisanih klaster-pravila, u odnosu na ukupni broj generisanih pravila (kratkih i dugih zajedno).

Kolone „SO kr“ i „SO dug“ odnose se na broj eliminisanih kratkih, odnosno dugih, statistički očekivanih pravila u skupu svih sesija. Kolona „SO uk %“ odnosi se na procenat ukupno eliminisanih statistički očekivanih pravila u skupu svih sesija, u odnosu na ukupni broj pravila (kratkih i dugih zajedno), koja preostaju u skupu posle eliminacije klaster-asocijativnih pravila.

Kolona „Opšt dug“ odnosi se na broj eliminisanih dugih pravila, koja su statistički očekivana u prisustvu opštijih pravila manje dužine. Kolona „Opšt dug %“ odnosi se na procenat eliminisanih dugih pravila (kolona „Opšt dug“), u odnosu na broj dugih pravila, koja preostaju u skupu posle eliminacije klaster-asocijativnih pravila („Klaster dug“) i eliminacije statistički očekivanih pravila u skupu svih sesija („SO dug“).

Kolone „Opšt konc kr“ i „Opšt konc dug“ odnose se na broj eliminisanih kratkih, odnosno dugih, statistički očekivanih pravila u prisustvu opštijih pravila jednake dužine, koja postoje u prisustvu konceptne hijerarhije. Kolona „Opšt konc kr %“ odnosi se na procenat eliminisanih kratkih pravila („Opšt konc kr“), u odnosu na broj kratkih pravila, koja preostaju u skupu posle eliminacije kratkih klaster-asocijativnih pravila („Klaster kr“) i eliminacije kratkih statistički očekivanih pravila u skupu svih sesija („SO

kr“). Kolona „Opšt konc dug %“ odnosi se na procenat eliminisanih dugih pravila („Opšt konc dug“), u odnosu na broj dugih pravila, koja preostaju u skupu posle eliminacije dugih klaster-asocijativnih pravila („Klaster dug“), eliminacije dugih statistički očekivanih pravila u skupu svih sesija („SO dug“), i eliminacije dugih pravila koja su statistički očekivana u prisustvu opštijih pravila manje dužine („Opšt dug“).

Kolone „Elim dug uk“ i „Elim kr uk“ odnose se na ukupni broj eliminisanih dugih, odnosno kratkih asocijativnih pravila primenom svih metoda eliminacije. Kolone „Elim dug uk %“ i „Elim kr uk %“ odnose se na procenat eliminacije dugih, odnosno kratkih pravila, u odnosu na ukupni broj generisanih dugih, odnosno kratkih pravila za date vrednosti minimalnog support praga i za dati confidence interval. Kolona „Elim uk %“ odnosi se na procenat eliminacije svih pravila (dugih i kratkih zajedno), u odnosu na ukupni broj generisanih pravila za date vrednosti minimalnog support praga i za dati confidence interval.

Rezultati eliminacije prikazani u tabelama 6.5, 6.6, 6.7 i 6.8, detaljno su analizirani u narednim poglavljima.

6.2.1 Eliminacija klaster-asocijativnih pravila

U ovom poglavlju analiziramo rezultate eliminacije klaster-pravila, koja su rezultat postojanja cikličnih struktura u konceptnoj hijerarhiji (poglavlje 4.2.3.2).

Eliminacija klaster-pravila predstavlja prvi korak prečišćavanja skupa asocijativnih pravila. Efikasnost ostalih metoda za eliminaciju statistički očekivanih asocijativnih pravila ispitujemo na skupu otkrivenih asocijativnih pravila, iz koga su prethodno eliminisana klaster-pravila.

Ukupan udeo klaster pravila u VTŠ skupu podataka je visok za sve vrednosti minimalnog support parametra i kreće se između 33% i 40%. Za razliku od toga, u FON skupu podataka udeo klaster pravila je vrlo nizak i kreće se od 0 do 1.4%. Klaster-pravila postoje u VTŠ skupu podataka usled strukture web sajta, koja uključuje pomoćne web stranice, koje se moraju posetiti da bi se došlo do traženih dokumenata, što je ujedno i jedini način da se dođe do tih dokumenata. Pri tome, često jedna pomoćna *php* web stranica odgovara jednom traženom *doc* dokumentu, te se takva

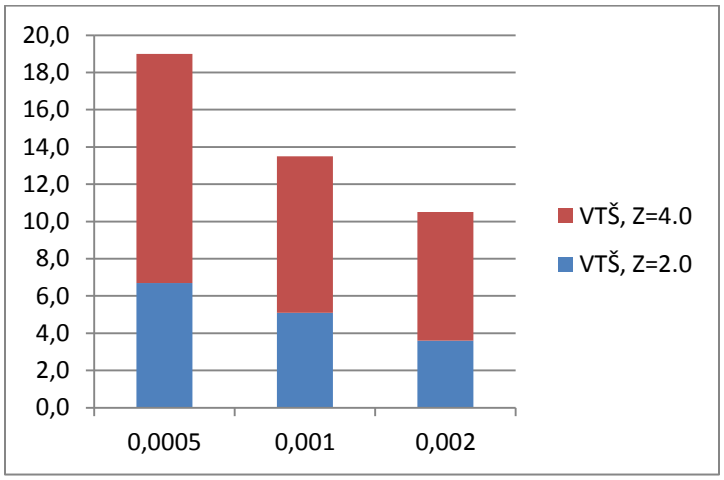
dva web objekta gotovo uvek pojavljuju u paru u web sesijama i čine cikličnu potkoncept/natkoncept strukturu. Čak i samo jedna ovakva ciklična struktura uzrokuje postojanje mnoštva klaster-asocijativnih pravila, te je ukupno smanjenje veličine skupa asocijativnih pravila eliminacijom klaster-pravila za VTŠ skup podataka značajno.

Eliminaciju klaster pravila bilo bi moguće ugraditi u fazu generisanja frekventnih skupova web stranica, čime bi se povećala efikasnost generisanja frekventnih skupova, i izbeglo formiranje klaster-pravila, ali ovu optimizaciju ostavljamo kao mogućnost budućeg istraživanja u cilju unapređenja performansi softverskog sistema za otkrivanje asocijativnih pravila o ponašanju korisnika web sajtova.

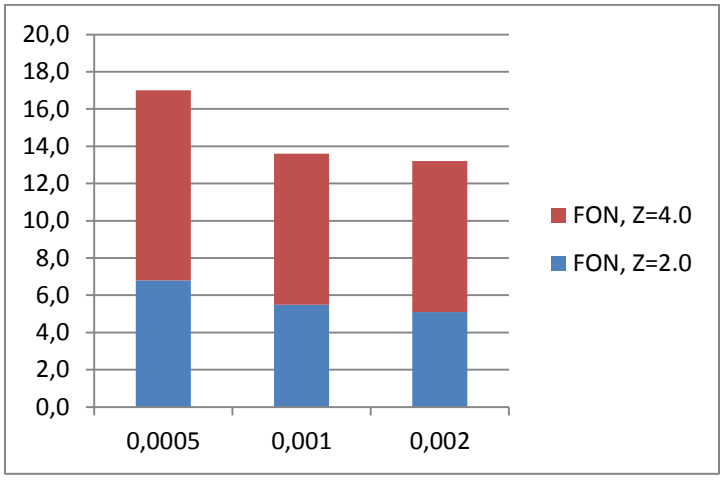
6.2.2 Eliminacija statistički očekivanih pravila u skupu svih sesija

U ovom poglavlju razmatramo učinak eliminacije statistički očekivanih pravila u skupu svih sesija primenom uslova baziranog na minimalnom Z-score pragu (poglavlje 4.2.1). Ovu metodu eliminacije primenjujemo na skup asocijativnih pravila prethodno prečišćen eliminisanjem klaster asocijativnih pravila, što je opisano u prethodnom poglavlju.

Grafikoni na slikama 6.3 i 6.4, generisani na osnovu podataka datih u tabelama 6.5, 6.6, 6.7 i 6.8, prikazuju učinak eliminacije statistički očekivanih pravila u skupu svih sesija za skupove podataka VTŠ, odnosno FON. Na Y-osi prikazan je procenat smanjenja veličine skupa asocijativnih pravila usled eliminacije statistički očekivanih pravila za vrednosti minimalnog Z-score praga 2.0 i 4.0, i to za tri različite vrednosti minimalnog support praga, koje su prikazane na X-osi.



Slika 6.3. Eliminacija statistički očekivanih pravila u skupu svih sesija: VTŠ



Slika 6.4. Eliminacija statistički očekivanih pravila u skupu svih sesija: FON

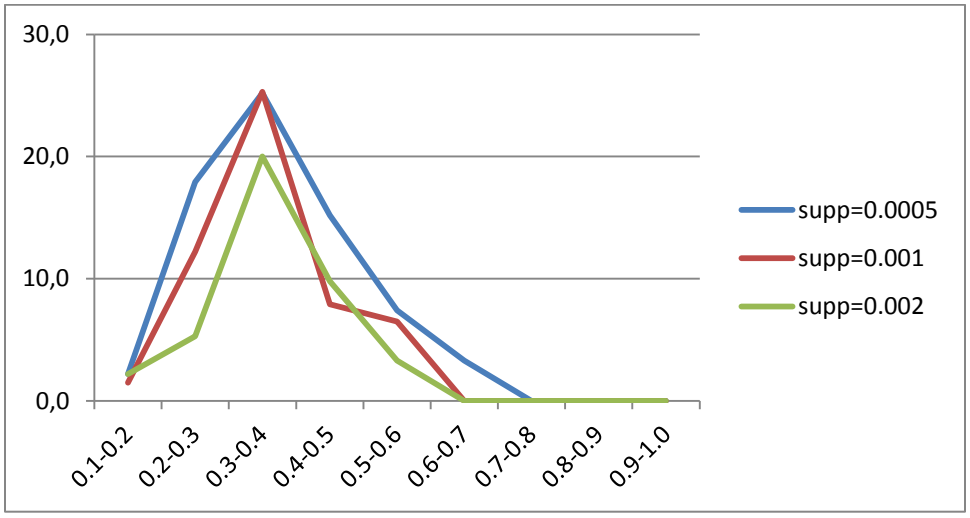
Kao što je očekivano, učinak eliminacije asocijativnih pravila blago opada sa povećanjem minimalnog support praga. Eksperimenti potvrđuju da je u skupu pravila koja imaju niže support vrednosti veći udeo statistički očekivanih pravila, u odnosu na skup pravila koja imaju više support vrednosti.

Ukupan učinak eliminacije statistički očekivanih pravila je relativno nizak i nalazi se u intervalu od 3% do 7% za minimalni Z-score prag od 2.0, i u intervalu od 7% do 13%

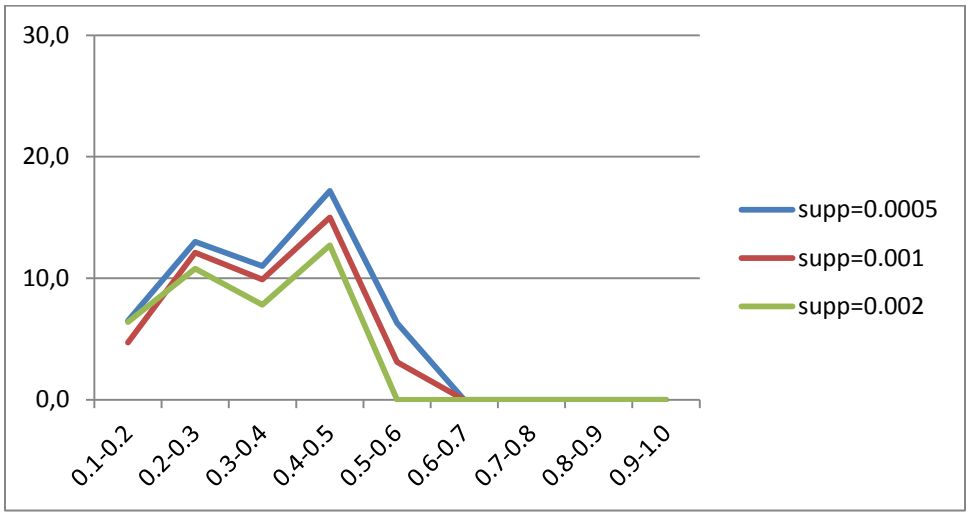
kada se minimalni Z-score prag poveća na 4.0. Za više vrednosti Z-score praga ne bi bilo statistički opravdano eliminisati asocijativna pravila, te takvi eksperimenti nisu rađeni.

S obzirom da je confidence jedna od najčešće korišćenih mera interesantnosti, izvršili smo eksperimente u kojima ispitujemo učinak eliminacije statistički očekivanih pravila za različite vrednosti confidence mere asocijativnih pravila.

Grafikoni na slikama 6.5 i 6.6 prikazuju učinak eliminacije statistički očekivanih pravila u skupu svih sesija za skupove podataka VTŠ i FON, pri čemu je vrednost minimalnog Z-score praga jednaka 2.0. Dati su rezultati za tri vrednosti minimalnog support praga (0.0005, 0.001 i 0.002). Pri tome su sva pravila podeljena u grupe prema intervalima vrednosti confidence mere kojima pripadaju, a koji su dati na X-osi. Na Y-osi prikazan je procenat smanjenja veličine skupa pravila čija confidence vrednost pripada određenom intervalu, koji je obeležen na X-osi.



Slika 6.5: Eliminacija statistički očekivanih pravila u skupu svih sesija prema confidence intervalima: VTŠ



Slika 6.6: Eliminacija statistički očekivanih pravila u skupu svih sesija prema confidence intervalima: FON

U skladu sa grafikonima prikazanim na slikama 6.3 i 6.4, na grafikonima prikazanim na slikama 6.5 i 6.6 primećuje se da je učinak eliminacije nešto veći za niže vrednosti minimalnog support praga.

Posebno je interesantno što je u gotovo svim eksperimentima najveći udeo statistički očekivanih asocijativnih pravila među pravilima koja imaju relativno povišene confidence vrednosti (u intervalu od 0.3 do 0.5). Analizom konkretnih pravila generisanih na skupovima podataka VTŠ i FON utvrdili smo da je razlog ove pojave često u tome što među pravilima čija je confidence mera u intervalu od 0.3 do 0.5, postoji veliki broj pravila čija je desna strana visoko frekventna (support 20-30%). Ovakva pravila prelaze minimalni support prag i imaju povišenu confidence vrednost isključivo kao posledicu visoke frekventnosti desne strane, pri čemu zapravo ne postoji statistički značajna korelacija leve i desne strane pravila. Dakle, ovakva pravila nepotrebno opterećuju skup otkrivenih pravila i pri tom zbunjuju analitičara podataka, pa ih je potpuno opravdano eliminisati.

Sa druge strane, među asocijativnim pravilima čija je vrednost confidence mere niska (u intervalu od 0.1 do 0.2) nalazi se veliki broj pravila čija su leva i desna strana povezane više nego što je to statistički očekivano. Analizom konkretnih pravila

generisanih na skupovima podataka VTŠ i FON utvrdili smo da ova pravila često povezuju web stranice koje nisu direktno povezane hiperlink strukturom web sajta, a koje su istovremeno posećivane češće nego što je to statistički očekivano. Smatramo da su neka od ovih pravila neočekivana, otkrivajući novo znanje o ponašanju korisnika web sajta, te su i pored niskih confidence vrednosti potencijalno interesantna analitičarima podataka.

Eksperimenti potvrđuju da pravila koja imaju ekstremno visoke confidence vrednosti (više od 0.7) nisu statistički očekivana u skupu svih sesija, te ne mogu biti eliminisana ovom metodom. Međutim, analizom konkretnih pravila generisanih na skupovima podataka VTŠ i FON utvrdili smo da pravila sa ekstremno visokim confidence vrednostima gotovo uvek sadrže web stranice između kojih postoji direktan hiperlink u strukturi web sajta. Visoka statistička korelacija leve i desne strane ovih pravila je samo posledica poznate hiperlink strukture web sajta, i stoga smatramo da ona nisu potencijalno interesantna analitičarima podataka. Ipak, ovakva pravila ne mogu biti eliminisana kao statistički očekivana, i ostaju u skupu otkrivenih asocijativnih pravila, nepotrebno zbunjujući analitičare podataka visokim vrednostima statističkih mera interesantnosti.

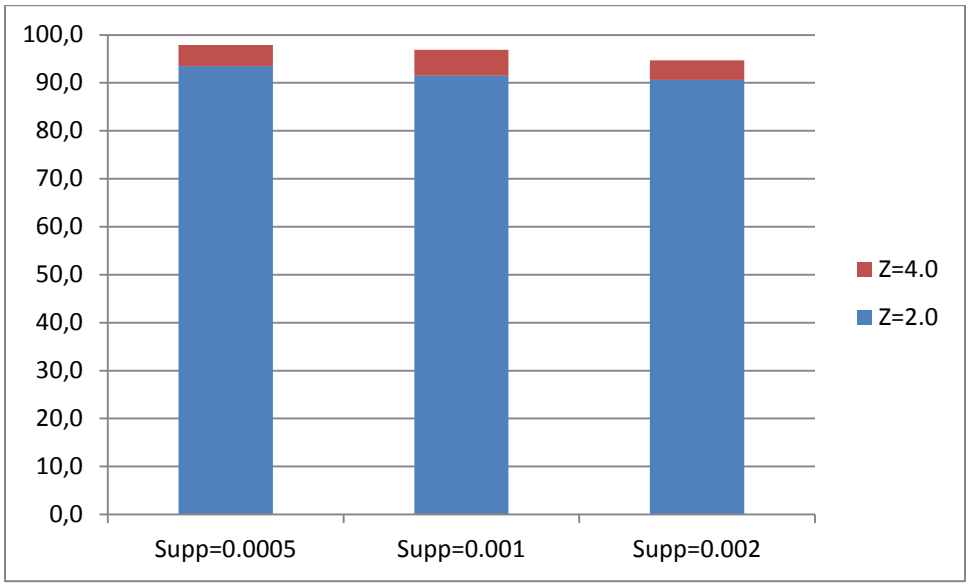
Generalno, eksperimenti opisani u ovom poglavlju potvrđuju poznatu činjenicu da confidence nije adekvatna mera interesantnosti asocijativnih pravila, te da postoji veliki broj pravila sa povišenim confidence vrednostima koja su apirori neinteresantna kao statistički očekivana, a koja mogu biti eliminisana u ovoj fazi prečišćavanja skupa otkrivenih pravila. Eksperimenti takođe pokazuju da je ukupni učinak eliminacije statistički očekivanih pravila u skupu svih sesija nizak. I posle primene ove metode eliminacije, preostaje veliki broj pravila koja imaju povišene vrednosti confidence i Z-score mere, a koja vrlo verovatno nisu interesantna analitičarima podataka.

6.2.3 Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila manje dužine

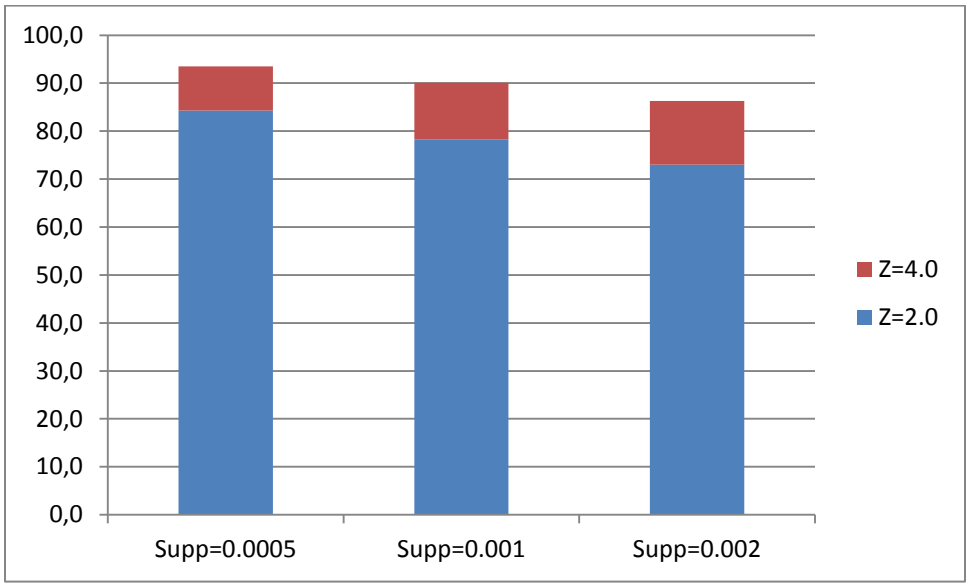
U ovom poglavlju razmatramo učinak eliminacije statistički očekivanih pravila u prisustvu opštijih asocijativnih pravila manje dužine, baziranog na minimalnom Z-score parametru računatom lokalno, na skupu svih sesija koje se odnose na opštije

asocijativno pravilo (poglavlje 4.2.3.1). Obzirom da se ovom metodom mogu eliminisati samo pravila koja sadrže bar tri web stranice („duga pravila“ u daljem tekstu), učinak ove metode eliminacije merimo samo na skupu target pozitivnih dugih pravila, koji je prethodno prečišćen eliminisanjem klaster asocijativnih pravila i statistički očekivanih asocijativnih pravila.

Grafikoni na slikama 6.7 i 6.8, generisani na osnovu podataka datih u tabelama 6.5, 6.6, 6.7 i 6.8, prikazuju učinak ove metode eliminacije u eksperimentima na skupovima podataka VTŠ, odnosno FON. Na Y-osi prikazan je procenat smanjenja veličine skupa target pozitivnih dugih pravila za vrednosti minimalnog Z-score praga 2.0 i 4.0, i to za tri različite vrednosti minimalnog support praga, koje su prikazane na X-osi.



Slika 6.7. Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila manje dužine: VTŠ



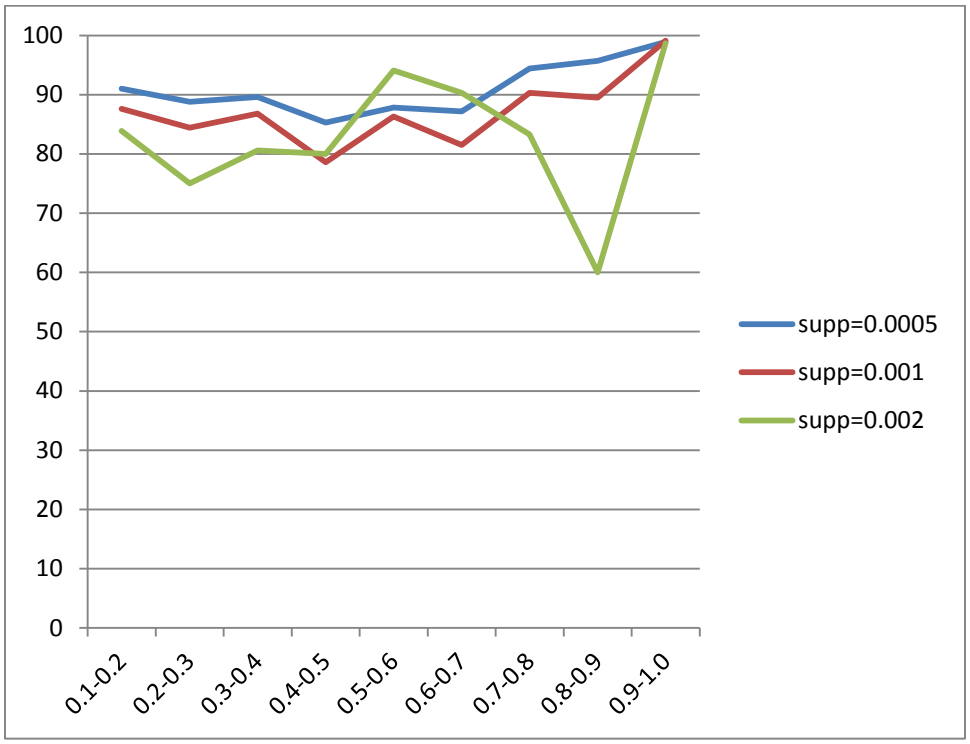
Slika 6.8. Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila manje dužine: FON

Učinak eliminacije statistički očekivanih pravila u prisustvu opštijih pravila manje dužine u opisanim eksperimentima je izuzetno visok. Za vrednost Z-score praga od 2.0 učinak eliminacije za VTŠ skup podataka se kreće od 90% do 93%, dok se za FON skup podataka kreće od 73% do 84%. Sa povećanjem minimalnog Z-score praga na 4.0 povećava se učinak eliminacije u proseku za oko 5% za VTŠ, i za oko 10% za FON skup podataka.

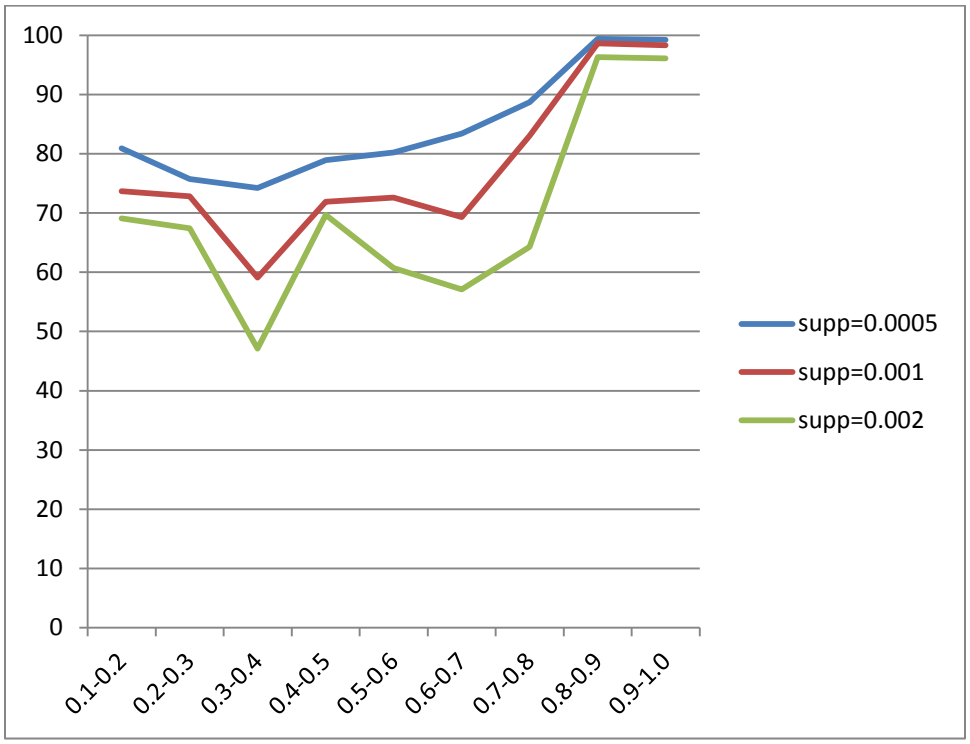
Učinak eliminacije statistički očekivanih pravila u prisustvu opštijih pravila manje dužine blago opada sa povećanjem minimalnog support praga, što je u skladu sa trendom učinka eliminacije statistički očekivanih pravila u skupu svih sesija. Dakle, među pravilima sa nižim support vrednostima nešto je veći udeo pravila koja su statistički očekivana u odnosu na kraća i opštija pravila.

Grafikoni na slikama 6.9 i 6.10 prikazuju učinak eliminacije statistički očekivanih pravila u prisustvu opštijih pravila manje dužine za skupove podataka VTŠ i FON, za vrednost minimalnog Z-score praga od 2.0. Dati su rezultati za tri vrednosti minimalnog support praga (0.0005, 0.001 i 0.002). Pri tome su sva pravila podeljena u grupe prema

intervalima vrednosti confidence mere, koji su dati na X-osi. Na Y-osi prikazan je procenat smanjenja veličine skupa pravila čija confidence vrednost pripada određenom intervalu obeleženom na X-osi.



Slika 6.9. Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila manje dužine prema confidence intervalima: VTŠ



Slika 6.10. Eliminacija statistički očekivanih pravila u prisustvu opštijih pravila manje dužine prema confidence intervalima: FON

Interesantno je da je učinak eliminacije najveći za pravila koja imaju ekstremno visoke confidence vrednosti. Za pravila čije su confidence vrednosti blizu maksimalne (u intervalu od 0.9 do 1.0) učinak eliminacije je blizu 100%. Ovakva pravila su trivijalna, i ona su gotovo uvek uzrokovana postojanjem kratkih pravila koja sadrže samo jednu web stranicu sa obe strane, i čija je confidence vrednost takođe u blizini maksimalne.

Napominjemo da oscilacije u procentima eliminacije mogu biti rezultat greške u merenju zbog malog broja pravila koja ulaze u dati confidence interval, što je naročito izraženo za minimalni support prag 0.002, kada u nekim confidence intervalima nema statistički značajnog broja pravila.

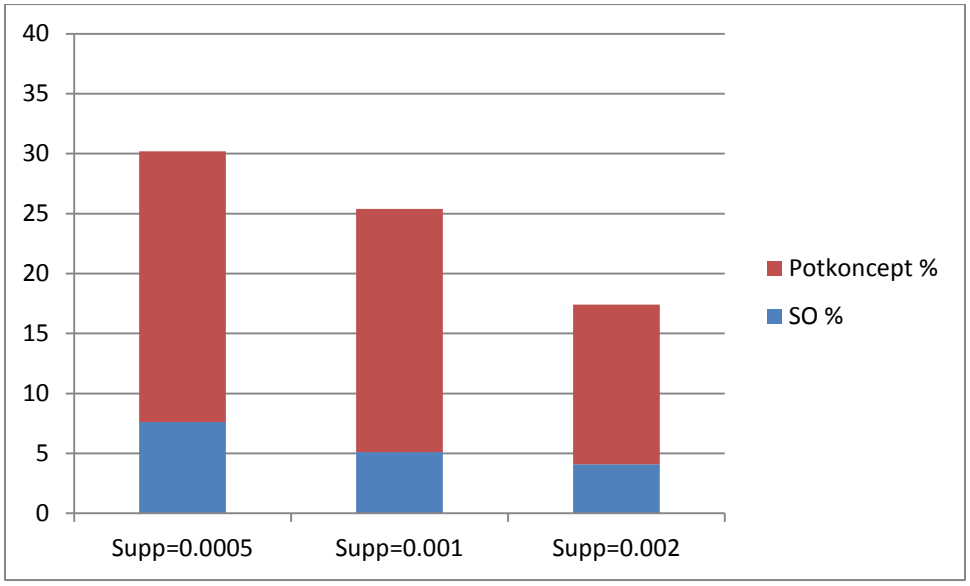
6.2.4 Eliminacija statistički očekivanih pravila u prisustvu konceptne hijerarhije

U ovom poglavlju razmatramo učinak eliminacije statistički očekivanih pravila u prisustvu opštijih asocijativnih pravila jednake dužine, pri čemu se minimalni Z-score parametar računa lokalno, na skupu svih sesija koje se odnose na opštije asocijativno pravilo (poglavlje 4.2.3.1). Ova metoda eliminacije podrazumeva prethodno formiranje konceptne hijerarhije u odnosu na koju se utvrđuje relacija „opštije/specifičnije asocijativno pravilo“. Ovu metodu primenjujemo na skup target pozitivnih pravila posle njegovog prečišćavanja eliminisanjem klaster asocijativnih pravila, statistički očekivanih asocijativnih pravila i dugih asocijativnih pravila koja su statistički očekivana u odnosu na kraća i opštija asocijativna pravila, čija je eliminacija opisana u poglavljima (6.2.1, 6.2.2 i 6.2.3). Na taj način ispitujemo značaj primene ove metode za eliminaciju onih asocijativnih pravila koja ne mogu biti prečišćena drugim, elementarnijim metodama.

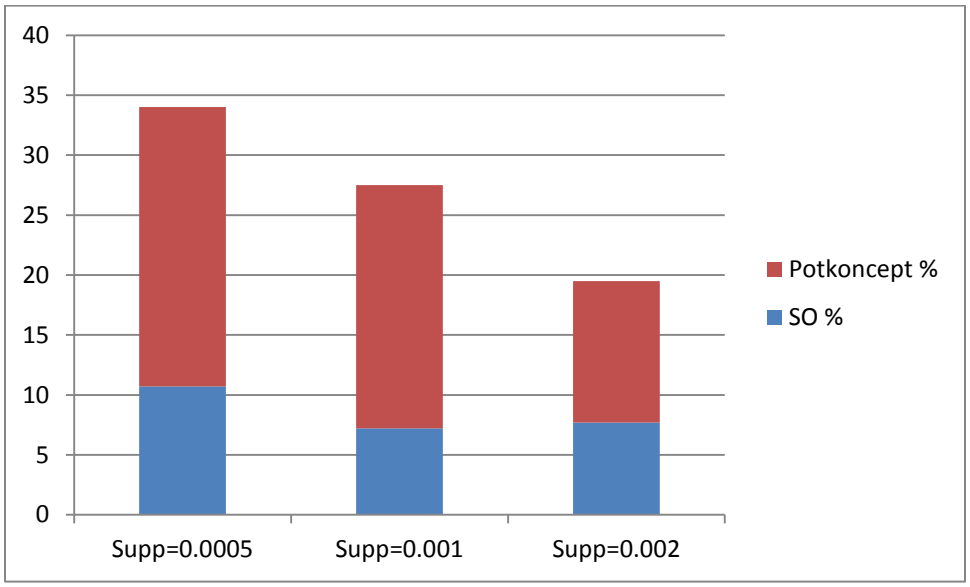
Eliminacija statistički očekivanih asocijativnih pravila u prisustvu konceptne hijerarhije utiče kako na duga, tako i na kratka asocijativna pravila. Međutim, posle prethodno primenjene eliminacije dugih asocijativnih pravila u prethodnom poglavlju, veoma je nizak broj dugih pravila koja preostaju u skupu, te eksperimenti u kojima bi se skupovi dugih pravila dalje prečišćavali nisu statistički validni. Sa druge strane, veličina skupa kratkih pravila se ne smanjuje znatno prethodno primenjenim metodama eliminacije. Pri tome, kratka pravila su posebno značajna zbog svoje jednostavnosti i lakoće razumevanja. Iz ovih razloga, efikasnost eliminacije statistički očekivanih pravila u prisustvu konceptne hijerarhije merimo na skupu target pozitivnih kratkih pravila.

Grafikoni na slikama 6.11, 6.12, 6.13 i 6.14, generisani na osnovu podataka datih u tabelama 6.7, 6.8, 6.9 i 6.10, prikazuju učinak prečišćavanja skupa target pozitivnih kratkih pravila za skupove podataka VTŠ, odnosno FON. Na Y-osi prikazan je procenat smanjenja veličine skupa target pozitivnih kratkih pravila kumulativno za dve metode eliminacije – statistički očekivanih pravila globalno u skupu svih sesija („SO kr“) i statistički očekivanih pravila u prisustvu konceptne hijerarhije („Opšt konc kr“), jer promena minimalnog Z-score praga utiče na učinak eliminacije za obe ove metode.

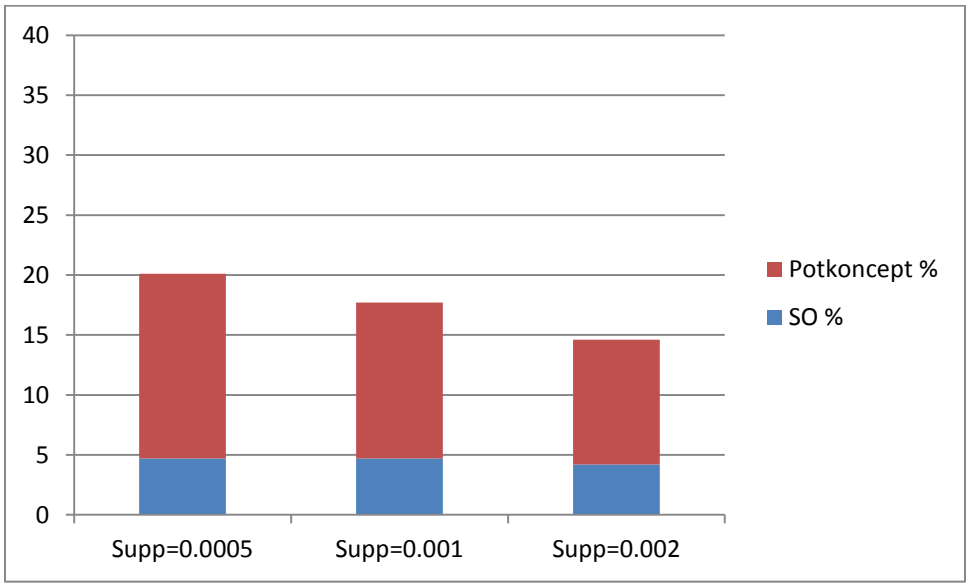
Grafikoni na slikama 6.11 i 6.13 prikazuju rezultate eksperimenata u kojima je vrednost minimalnog Z-score praga jednaka 2.0 za skupove podataka VTŠ, odnosno FON. U eksperimentima čiji je rezultat prikazan na grafikonima 6.12 i 6.14 za skupove podataka VTŠ, odnosno FON vrednost minimalnog Z-score praga je jednaka 4.0. Svi eksperimenti izvršeni su za tri različite vrednosti minimalnog support praga, koje su obeležene na X-osi.



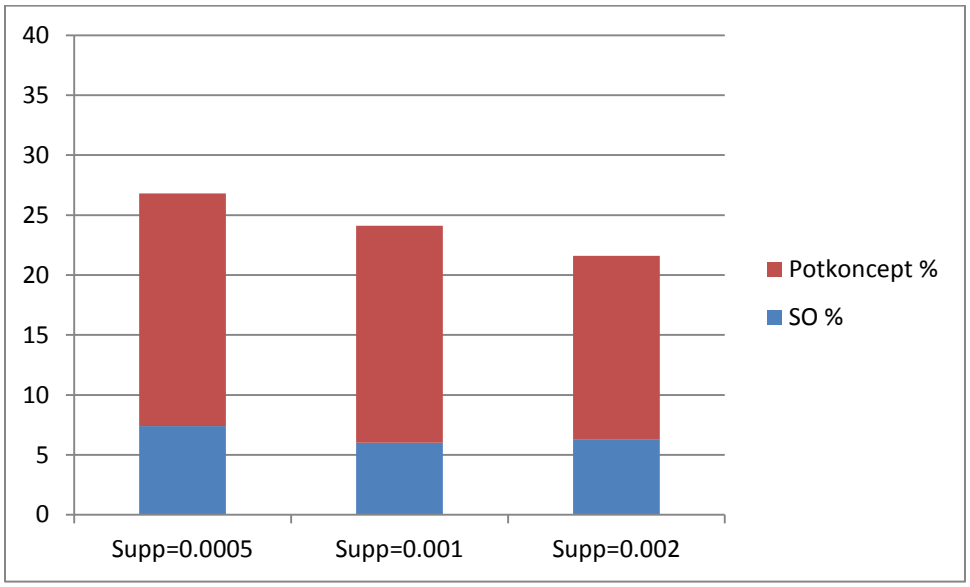
Slika 6.11. Eliminacija statistički očekivanih kratkih pravila u prisustvu opštijih pravila jednake dužine: VTŠ, Z=2.0



Slika 6.12. Eliminacija statistički očekivanih kratkih pravila u prisustvu opštijih pravila jednake dužine: VTŠ, Z=4.0



Slika 6.13. Eliminacija statistički očekivanih kratkih pravila u prisustvu opštijih pravila jednake dužine: FON, Z=2.0



Slika 6.14. Eliminacija statistički očekivanih kratkih pravila u prisustvu opštijih pravila jednake dužine: FON, Z=4.0

Rezultati eksperimenata pokazuju da je učinak eliminacije target pozitivnih kratkih pravila u proseku veći za VTŠ skup podataka nego za FON skup podataka. Razlog leži u tome što je veličina konceptne hijerarhije znatno veća za VTŠ skup podataka nego za FON skup podataka, usled velikog broja „jakih“ pravila, koja su posledica strukture VTŠ web sajta.

Učinak eliminacije blago opada sa povećanjem minimalnog support praga, što je u skladu sa trendom učinka eliminacije statistički očekivanih pravila u skupu svih sesija, kao i statistički očekivanih pravila u odnosu na kraća i opštija pravila. Dakle, među pravilima sa nižim support vrednostima nešto je veći udeo specifičnijih pravila koja se mogu eliminisati kao statistički očekivana u odnosu na opštija pravila.

Sa povećanjem minimalnog Z-score praga povećava se ukupni učinak eliminacije za između 3% i 7% u zavisnosti od skupa podataka i minimalnog support praga. Ukupni učinak eliminacije kreće se u intervalu od 14% do 34%, u zavisnosti od skupa podataka, minimalnog support praga i minimalnog Z-score praga.

6.3 Analiza kvaliteta znanja dobijenog proširenim softverskim sistemom

Kvalitet asocijativnih pravila otkrivenih primenom implementiranog softverskog sistema znatno je povećan samom činjenicom da su eliminisana neinteresantna pravila, kao što je detaljno opisano u prethodnom poglavlju. Poslednji korak u procesu otkrivanja asocijativnih pravila je rangiranje ne-eliminisanih pravila primenom odgovarajućih mera interesantnosti. U okviru ovog poglavlja ispitujemo primenu implementiranih standardnih i modifikovanih statističkih mera interesantnosti na rangiranje asocijativnih pravila otkrivenih u stvarnim skupovima podataka.

Predložene modifikovane statističke mere interesantnosti se mogu primenjivati kako na kratka, tako i na duga pravila. Međutim, prečišćavanje skupa dugih pravila primenjeno u prethodnom poglavlju svelo je broj dugih pravila na veoma mali broj pravila u većini eksperimenata. Kvalitet rangiranja ovih pravila bilo bi teško validno izmeriti zbog njihovog malog broja. Pored toga, smatramo da je generalno teško izvršiti pravilno rangiranje dugih pravila tako da ona budu zaista korisna analitičarima podataka, te se stoga ona retko i upotrebljavaju u praksi. Sa druge strane, kratka pravila su zbog lakoće razumevanja najčešće upotrebljavana od strane analitičara podataka, te je razvoj metoda koje ih pravilno rangiraju posebno značajan (Liu, 2007; Kazienko, 2009). Iz ovih razloga se u eksperimentima koji se odnose na rangiranje asocijativnih pravila o korišćenju web sajtova u okviru ovog istraživanja ograničavamo na skup kratkih asocijativnih pravila, a rangiranje dugih pravila ostavljamo kao mogućnost budućeg istraživanja.

6.3.1 Rangiranje kratkih asocijativnih pravila

Prilikom rangiranja kratkih asocijativnih pravila, predlažemo da se ona podele u grupe, u zavisnosti od toga da li sadrže web stranice koje su u relaciji potkoncept/natkoncept u okviru prisutne konceptne hijerarhije, i to na sledeći način:

1. **Grupa „P-N“**

   Grupu „P-N“ čine pravila čija je leva strana potkoncept desne strane. Confidence vrednost ovih pravila je prema definiciji relacije potkoncept/natkonept veća od zadatog praga (0.7 u našim eksperimentima). Pri tome je besmisleno koristiti standardne statističke mere interesantnosti za ova pravila jer je skup sesija koje sadrže levu stranu pravila skoro pravi podskup skupa sesija koje sadrže desnu stranu pravila. Smatramo da pravila iz ove grupe ne otkrivaju novo znanje, već potvrđuju vezu potkoncept/natkoncept koja je ugrađena u hiperlink strukturu web sajta. Stoga predlažemo da se ova pravila izdvoje u posebnu grupu čija uloga je da opiše konceptnu hijerarhiju web stranica.

2. **Grupa „N-P“**

   Grupu „N-P“ čine pravila čija je desna strana potkoncept leve strane. Ova pravila se mogu tumačiti u smislu: „Posetioci web sajta koji su zainteresovani za koncept N zainteresovani su i za njegov potkoncept P sa određenim stepenom poverenja“. Confidence je dovoljno relevantna mera interesantnosti za pravila iz ove grupe, dok je besmisleno koristiti lift, added value i Z-score kao mere interesantnosti za ova pravila. Pravila iz ove grupe mogu se koristiti u specifične svrhe, kao što je ocenjivanje upotrebljivosti linkova. Zbog specifičnosti značenja ovih pravila, ona se ne mogu svrstati u istu grupu kao pravila opšteg oblika.

3. **Grupa „P1-P2“**

   Grupu „P1-P2“ čine pravila čija su leva i desna strana potkoncepti zajedničkog natkoncepta. Ova pravila se mogu tumačiti u smislu: „Posetioci web sajta koji su zainteresovani za koncept N i njegov potkoncept P1, zainteresovani su i za njegov potkoncept P2, sa određenim stepenom poverenja“. Za rangiranje ovih pravila predlažemo korišćenje modifikovanih statističkih mera interesantnosti računatih lokalno, na skupu sesija koje sadrže zajednički natkoncept (poglavlje 4.7).

4. **Grupa „GEN“**

   Grupu „GEN“ čine pravila opšteg oblika, čija leva i desna strana nisu povezane vezom potkoncept/natkoncept, niti su one potkoncepti zajedničkog natkoncepta. Za rangiranje pravila u ovoj grupi može se koristiti bilo koja standardna statistička mera interesantnosti.

Dok pravila iz grupa „P-N i „N-P“ imaju specifično značenje i posmatraju se odvojeno, pravila iz grupa „P1-P2“ i „GEN“ tumače se na sličan način. Njihovo značenje ima smisao: „Posetioci web sajta koji su zainteresovani za neki koncept, takođe su zainteresovani i za drugi koncept, sa određenim stepenom poverenja“. Ova pravila potencijalno otkrivaju novo znanje o interesovanjima korisnika web sajta, što je upravo fokus našeg istraživanja.

U eksperimentima prikazanim u ovom poglavlju grupe „P1-P2“ i „GEN“ su spojene i izvršeno je zajedničko rangiranje pravila iz ove dve grupe. Korišćene su tri modifikovane statističke mere interesantnosti: modifikovani added value, modifikovani lift i modifikovani Z-score.

Vrednost modifikovane statističke mere interesantnosti $M(R)$ za neko asocijativno pravilo R formira se prema sledećem kriterijumu. Ako R pripada grupi pravila „GEN“, $M(R)$ dobija vrednost standardne statističke mere interesantnosti. U protivnom, ako R pripada grupi pravila „P1-P2“, $M(R)$ dobija vrednost odgovarajuće lokalne mere interesantnosti (poglavlje 4.3).

U eksperimentima u okviru kojih je izvršeno rangiranje pravila iz grupa „P1-P2“ i „GEN“ korišćeno je sedam statističkih mera interesantnosti:

1. confidence
2. added value
3. lift
4. Z-score
5. modifikovani added value
6. modifikovani lift
7. modifikovani Z-score

6.3.2 Primeri iz stvarnog skupa podataka

Kao primer koji ilustruje razliku u rangiranju standardnom i modifikovanom merom interesantnosti, u tabeli 6.9 data su tri kratka pravila otkrivena u skupu podataka FON, u eksperimentu gde je minimalni support prag bio 0.002, a minimalni Z-score prag 4.0. Pri tome unija kratkih pravila iz grupa „GEN“ i „P1-P2“ generisanih u ovom eksperimentu sadrži ukupno 77 pravila.

	Levo	Desno	Grupa	AV	Rang AV	Mod AV	Rang MAV
1	/osnovnestudije/om/index.html	/osnovnestudije/uk/index.html	P1-P2	0.43	11	0.40	9
2	/istrazivanjeirazvoj/index.html	/ofakultetu/index.html	GEN	0.35	18	0.35	13
3	/osnovnestudije/om/index.html	/osnovnestudije/isit/index.html	P1-P2	0.43	10	0.24	23

Tabela 6.9. Primeri rangiranih pravila – modifikovana added value mera interesantnosti

Kolona „Levo“ u tabeli 6.9 sadrži url web stranice na levoj strani pravila, a kolona „Desno“ sadrži url web stranice na desnoj strani pravila. Kolona „Grupa“ označava grupu kojoj pravilo pripada prema podeli navedenoj u prethodnom poglavlju. Kolona „AV“ sadrži vrednost added value mere interesantnosti, dok kolona „Rang AV“ sadrži redni broj pravila kada su ona rangirana prema added value meri interesantnosti. Kolona „Mod.AV“ sadrži vrednost modifikovane added value mere interesantnosti, dok kolona „Rang MAV“ sadrži redni broj pravila kada su ona rangirana prema modifikovanoj added value meri interesantnosti.

Web stranica */osnovnestudije/om/index.html* odnosi se na studijski program „Operacioni menadžment“, web stranica */osnovnestudije/uk/index.html* odnosi se na studijski program „Upravljanje kvalitetom“, a web stranica */osnovnestudije/isit/index.html* odnosi se na studijski program „Informacione tehnologije“ Fakulteta organizacionih nauka u Beogradu. Sve tri web stranice su potkoncepti zajedničkog natkoncepta „Osnovne studije“ na koji se odnosi web stranica */osnovnestudije/index.html.* Web stranice */istrazivanjeirazvoj/index.html* i */ofakultetu/index.html* nemaju natkoncept i obe se nalaze na osnovnom top-meniju web sajta.

Ukoliko bi se rangiranje vršilo standardnom added value merom interesantnosti, pravila sa rednim brojevima 1 i 3 bila bi rangirana na visokom 10. i 11. mestu („Rang AV“) i imala bi jednaku vrednost mere interesantnosti („AV“), dok bi pravilo sa rednim brojem 2 bilo niže rangirano (na 18. mestu). Korišćenjem modifikovane added value mere interesantnosti primećujemo da postoji razlika u rangiranju pravila sa rednim brojevima 1 i 3 („Rang M.AV“). Pravilo sa rednim brojem 1 ima višu vrednost modifikovane added value mere i prema njoj je rangirano na 9. mestu, dok je pravilo sa rednim brojem 3 rangirano tek na 23. mestu. Dakle, modifikovana added value mera pravila 1 i 3 govori o tome da su posetioci web sajta koji su zainteresovani za osnovne studije na studijskom programu „Operacioni menadžment“ mnogo više zainteresovani za studijski program „Upravljanje kvalitetom“ nego za studijski program „Informacione tehnologije“ (što se može očekivati s obzirom na srodnost ovih studijskih programa). Sa druge strane, razlika u vrednostima standardne added value mere između pravila 1 i 3 toliko je mala da se ova informacija gubi i pravila su podjednako rangirana.

Pravilo 2, koje povezuje web stranice između kojih ne postoji hiperlink na web sajtu i koje nisu povezane vezom potkoncept/natkoncept, može biti značajno analitičarima podataka i nepoželjno bi bilo rangirati ga previše nisko u odnosu na pravila 1 i 3, što bi bio slučaj pri rangiranju standardnom added value merom. Smatramo da je njegovo rangiranje korišćenjem modifikovane added value mere („Rang MAV“) bliže njegovoj stvarnoj interesantnosti, u odnosu na njegovo rangiranje korišćenjem standardne added value mere („Rang AV“).

6.3.3 Poređenje rangiranja standardnim statističkim merama interesantnosti

U ovom poglavlju poredimo rangiranje kratkih asocijativnih pravila korišćenjem četiri standardne mere interesantnosti: confidence, added value, lift i Z-score. U eksperimentima poredimo sličnost rangiranja putem ovih mera interesantnosti korišćenjem Spearmanovog koeficijenta korelacije ranga. Vrednosti minimalnog support praga su kao i u prethodnom poglavlju varirane između 0.0005, 0.001 i 0.002, dok je vrednost minimalnog Z-score praga konstantna i iznosi 4.0.

U tabelama od 6.10 do 6.15 date su vrednosti Spearmanovog koeficijenta korelacije ranga za svaki par mera interesantnosti. Skraćenica „A.value“ u tabelama odnosi se na added value, a skraćenica „Conf“ na confidence. Oznaka skupa podataka i vrednost minimalnog support praga data je u zagradama ispod svake tabele.

	A.value	Lift	Z-score
Conf	0.9001	0.2414	0.3021
A.value	-	0.5446	0.5730
Lift	-	-	0.8926

Tabela 6.10. Poređenje rangiranja standardnim merama: VTŠ, Supp=0.0005

	A.value	Lift	Z-score
Conf	0.8598	0.1320	0.2325
A.value	-	0.5211	0.5977
Lift	-	-	0.9212

Tabela 6.11. Poređenje rangiranja standardnim merama: VTŠ, Supp=0.001

	A.value	Lift	Z-score
Conf	0.818	-0.065	0.0653
A.value	-	0.3900	0.4971
Lift	-	-	0.8941

Tabela 6.12. Poređenje rangiranja standardnim merama: VTŠ, Supp=0.002

	A.value	Lift	Z-score
Conf	0.9107	0.1956	0.1991
A.value	-	0.4858	0.4348
Lift	-	-	0.7234

Tabela 6.13. Poređenje rangiranja standardnim merama: FON, Supp=0.0005

	A.value	Lift	Z-score
Conf	0.89308	0.2390	0.2139
A.value	-	0.5401	0.4783
Lift	-	-	0.7816

Tabela 6.14. Poređenje rangiranja standardnim merama: FON, Supp=0.001

	A.value	Lift	Z-score
Conf	0.8967	0.214	0.1818
A.value	-	0.5109	0.453
Lift	-	-	0.8239

Tabela 6.15. Poređenje rangiranja standardnim merama: FON, Supp=0.002

U svim prikazanim eksperimentima Spearmanov koeficijent korelacije ranga između confidence i added value mere je visok, što znači da ove dve mere daju slično rangiranje. Koeficijent korelacije ranga između confidence i lift, kao i između confidence i Z-score mere je nizak, što ukazuje na različitost rangiranja putem ovih mera interesantnosti. Ipak, koeficijent u većini slučajeva zadržava pozitivnu vrednost, što znači da postoji pozitivna korelacija u rangiranju, ali je niska. Sličnost u rangiranju između added value i lift, kao i added value i Z-score mere postoji, ali nije izrazito visoka, dok je sličnost u rangiranju između lift i Z-score mere izrazito visoka u svim eksperimentima.

6.3.4 Poređenje rangiranja standardnim i modifikovanim statističkim merama interesantnosti

U ovom poglavlju poredimo rangiranje kratkih asocijativnih pravila korišćenjem standardne i modifikovane added value, standardne i modifikovane lift, kao i standardne i modifikovane Z-score mere interesantnosti. Kao i u prethodnom poglavlju, sličnost rangiranja merimo korišćenjem Spearmanovog koeficijenta korelacije ranga. Vrednosti minimalnog support praga, kao i u prethodnom poglavlju, varirane su između 0.0005, 0.001 i 0.002, dok je vrednost minimalnog Z-score praga konstantna i iznosi 4.0.

U tabelama 6.16 i 6.17 date su vrednosti Spearmanovog koeficijenta korelacije ranga između parova standardna/modifikovana mera interesantnosti. Kolona „Support“ odnosi se na vrednost minimalnog support praga, kolona „Mod AV“ odnosi se na korelaciju između modifikovane i standardne added value mere, kolona „Mod Lift“ na korelaciju između modifikone i standardne lift mere, a kolona „Mod Z-score“ na korelaciju između modifikone i standardne Z-score mere. Oznaka skupa podataka data je u zagradama ispod obe tabele.

Support	Mod AV	Mod Lift	Mod Z-score
0.0005	0.8948	0.6787	0.6022
0.001	0.8770	0.6381	0.5539
0.002	0.8553	0.5565	0.4733

Tabela 6.16. Rangiranje modifikovanim merama: FON

Support	Mod AV	Mod Lift	Mod Z-score
0.0005	0.8893	0.6971	0.5642
0.001	0.8512	0.6127	0.5288
0.002	0.8654	0.5876	0.5234

Tabela 6.17. Rangiranje modifikovanim merama: VTŠ

Eksperimenti pokazuju da je sličnost u rangiranju između modifikovane i standardne added value mere interesantnosti nešto veća u odnosu na sličnost između modifikovane i standardne lift, odnosno Z-score mere interesantnosti. Razlog za ovu pojavu je verovatno u tome što vrednosti lift i Z-score mere više variraju kada se smanji veličina skupa sesija u odnosu na koji se one računaju. Kada se lift, odnosno Z-score vrednost računaju lokalno u nekom podskupu skupa svih sesija, dolazi do većih oscilacija u odnosu na njihove vrednosti računate globalno u skupu svih sesija.

7 Zaključak

Otkrivanje asocijativnih pravila u web log podacima jedna je od popularnih metoda za automatsku ekstrakciju potencijalno interesantnih informacija o korišćenju web sajtova. Znanje otkriveno u formi asocijativnih pravila primenjuje se u raznovrsnim domenima, kao što su aplikacije elektronskog poslovanja, razni sistemi za preporučivanje, personalizacija korišćenja web sajtova, sistemi za povećanje performansi web servera keširanjem web stranica, kao i za unapređenje web sajt dizajna.

Analiza dosadašnjih istraživanja u oblasti otkrivanja asocijativnih pravila ukazuje da je osnovni faktor koji negativno utiče na upotrebljivost asocijativnih pravila tendencija generisanja prevelikog broja pravila u kojima se analitičari podataka teško snalaze pri odabiru stvarno korisnih pravila. U slučaju asocijativnih pravila o korišćenju web sajtova ovaj problem je pogoršan usled snažne korelacije između različitih web stranica, koja je rezultat hiperlink strukture web sajtova. Kao posledica toga generiše se preveliki broj asocijativnih pravila sa visokim vrednostima statističkih mera interesantnosti, koja su zapravo očekivana i samim tim neinteresantna analitičarima podataka. Postojanje ovakvih asocijativnih pravila vrlo negativno utiče na kvalitet znanja sadržan u skupu asocijativnih pravila otkrivenih u web podacima, umanjujući njegovu upotrebljivost.

Shodno tome, opšti cilj istraživanja sprovedenog u okviru ove disertacije bio je *razvoj teorijskog okvira i unapređenje metoda za pronalaženje i vrednovanje asocijativnih pravila u web server log podacima*. U okviru opšteg cilja definisan je potcilj *C1.1 – uporedna analiza primene različitih mera interesantnosti pravila kroz aspekt njihove upotrebljivosti za analizu web server log podataka,* kao i potcilj *C1.2 – formulisanje smernica pri odabiru mera interesantnosti asocijativnih pravila.*

Jedan od važnijih teoretskih doprinosa ove disertacije, u okviru ispunjenja opšteg cilja, je predlog metode za eliminaciju neinteresantnih asocijativnih pravila, kojom se povećava upotrebljivost skupa otkrivenih pravila. Predložena je formalna definicija uslova za eliminaciju koja se bazira na statističkoj Z-score meri, definisanoj lokalno, na skupu transakcija (web sesija) koje sadrže opštiji skup atributa (web stranica).

Ovakva definicija opšteg uslova za eliminaciju asocijativnih pravila obuhvata dvojako prečišćavanje skupa otkrivenih asocijativnih pravila. U prvom slučaju eliminišu se asocijativna pravila koja su statistički očekivana u odnosu na opštija i kraća asocijativna pravila. U drugom slučaju eliminišu se pravila koja su statistički očekivana u odnosu na opštija asocijativna pravila jednake dužine, koja postoje u prisustvu konceptne hijerarhije atributa.

U okviru ove metode predloženo je korišćenje konceptne hijerarhije generisane na osnovu asocijativnih pravila čija je confidence vrednost približna maksimalnoj, a koja često postoje u skupovima asocijativnih pravila o korišćenju web sajtova. Ograničenje ovakve konceptne hijerarhije je što ona ne obuhvata sve odnose potkoncept/natkoncept koje bi mogao definisati ekspert, ili koji bi eventualno mogli biti automatski ekstrahovani semantičkom analizom web sajta. Sa druge strane, prednost ovako definisane konceptne hijerarhije je što za njeno generisanje nije potrebno ekspertsko znanje, niti se zahteva dodatno računsko procesiranje, jer se ona zasniva na već otkrivenim asocijativnim pravilima.

Predloženi metod za eliminaciju neinteresantnih asocijativnih pravila baziran na primeni statističke Z-score mere interesantnosti je poređen sa drugim metodama (potcilj *C1.1*). Teoretski je dokazano da je predloženi metod opštiji i više statistički opravdan u odnosu na neke od metoda korišćenih u prethodnim istraživanjima. Posebno je pogodan za primenu u domenu otkrivanja asocijativnih pravila o ponašanju korisnika web sajtova, gde postoji visoka korelacija između web stranica kao posledica hiperlink strukture web sajta.

U okviru potcilja *C1.1*, izvršena je uporedna analiza osobina različitih statističkih mera interesantnosti. One su primenjene na rangiranje asocijativnih pravila o korišćenju web sajtova otkrivenih u dva stvarna skupa web log podataka. Rezultati rangiranja putem različitih mera interesantnosti asocijativnih pravila o korišćenju web sajtova poređeni su primenom Spearmanov-og koeficijenta korelacije ranga.

Još jedan teoretski doprinos, u okviru potcilja *C1.2,* je predlog modifikacije standardnih matematičkih mera interesantnosti, kojom se povećava kvalitet rangiranja otkrivenih asocijativnih pravila o korišćenju web sajtova. Ova modifikacija primenljiva je u

prisustvu prethodno definisane konceptne hijerarhije web objekata. Njome se modifikuje interesantnost asocijativnih pravila čija leva i desna strana sadrže potkoncept zajedničkog natkoncepta. Ovakva asocijativna pravila imaju izrazito povišene vrednosti standardnih statističkih mera interesantnosti, iako su ona najčešće očekivana, i samim tim neinteresantna analitičarima podataka. Eksperimentalno je pokazano da ovakva modifikacija mera interesantnosti daje kvalitetnije rangiranje asocijativnih pravila o korišćenju web sajtova koja preostaju u skupu posle njegovog prečišćavanja.

Praktični cilj ovog istraživanja bio je *razvoj softverskog sistema za analizu web log podataka sa mogućnošću odabira postojećih i generisanja novih mera interesantnosti asocijativnih pravila (C2).* U okviru praktičnog cilja definisani su potcilj *C2.1 – analiza funkcionalnosti određenog softverskog sistema za data mining i predlog njegovog proširenja,* kao i potcilj *C2.2 – proširenje funkcionalnosti postojećeg softverskog sistema u cilju poboljšanja kvaliteta otkrivenog znanja.*

Analiziran je popularni Weka data mining sistem i predložen softverski sistem koji proširuje njegove funkcionalnosti (potcilj *C2.1*). Predloženi softverski sistem je specijalizovan za otkrivanje asocijativnih pravila o korišćenju web sajtova i integriše sve faze procesa otkrivanja asocijativnih pravila. Obuhvaćena je priprema web log podataka, otkrivanje asocijativnih pravila, implementacija predloženih metoda za eliminisanje neinteresantnih asocijativnih pravila, kao i rangiranje preostalih asocijativnih pravila primenom standardnih i modifikovanih mera interesantnosti.

U okviru potcilja *C2.2*, implementiran je predloženi softverski sistem, kojim se u okviru eksperimentalnog istraživanja otkrivaju asocijativna pravila u dva stvarna skupa podataka o korišćenju web sajtova. Pri tome se eksperimentalno ispituje doprinos predloženih metoda za poboljšanje kvaliteta otkrivenih asocijativnih pravila o korišćenju web sajtova.

Najpre je izvršeno pretprocesiranje web log podataka, zatim su generisana asocijativna pravila o ponašanju korisnika web sajtova *Apriori* algoritmom, a potom su primenjene predložene metode za eliminaciju i rangiranje otkrivenih asocijativnih pravila.

U prvom delu eksperimenata je potvrđeno da ukupan broj frekventnih skupova i otkrivenih asocijativnih pravila koja prelaze zadati support/confidence prag raste eksponencijalno sa brojem frekventnih web stranica.

Eksperimenti u kojima se proverava efikasnost predloženih metoda za prečišćavanje skupa otkrivenih asocijativnih pravila vršeni su gradativno, u četiri koraka. U prvim koracima se primenjuju jednostavnije metode prečišćavanja, i meri njihova efikasnost pri smanjenju veličine skupa otkrivenih asocijativnih pravila. U narednim koracima meri se efikasnost kompleksnijih metoda prečišćavanja, i to na skupu asocijativnih pravila koja nisu eliminisana prethodno primenjenim, elementarnijim metodama prečišćavanja.

U prvom koraku eliminisana su takozvana „klaster“ pravila, obzirom da su takva pravila očigledno trivijalna, a računski ih je jednostavno identifikovati. Ukupan udeo klaster pravila u VTŠ eksperimentalnom skupu podataka je visok (od 33% do 40%), dok je u FON skupu podataka nizak (od 0% do 1.4%). Postojanje klaster pravila uslovljeno je hiperlink strukturom web sajta, u okviru koje se dve ili više web stranica moraju posetiti istovremeno kako bi se došlo do tražene informacije. Čak i samo jedan takav par web stranica uzrokuje postojanje mnoštva klaster asocijativnih pravila, te je ukupno smanjenje veličine skupa otkrivenih pravila značajno.

Eliminacija asocijativnih pravila koja su statistički očekivana u skupu svih web sesija je jedna od metoda često predlagana u prethodnim istraživanjima, pa je ona primenjena kao drugi korak procesa prečišćavanja skupa otkrivenih asocijativnih pravila. Pri tome je za meru statističke očekivanosti asocijativnih pravila korišćena statistička Z-score mera. Eksperimentima je potvrđeno da je učinak eliminacije statistički očekivanih pravila u skupu otkrivenih pravila nizak (od 3% do 13%, u zavisnosti od skupa podataka i vrednosti zadatog minimalnog Z-score praga). Potvrđeno je i da se učinak eliminacije blago povećava sa sniženjem minimalnog support praga, što je u skladu sa prethodnim istraživanjima i sa definicijom support mere. Eksperimenti takođe pokazuju da je učinak eliminacije veći za pravila koja imaju relativno visoke confidence vrednosti (u intervalu od 0.3 do 0.5), nego za pravila sa confidence vrednostima nižim od 0.3. Razlog ove pojave je što je za veliki broj pravila povišena confidence vrednost samo posledica visoke frekventnosti desne strane pravila, pri čemu zapravo ne postoji

statistički značajna korelacija leve i desne strane pravila. Visoke confidence vrednosti ovakvih pravila zbunjuju analitičare podataka, dajući lažnu sliku o njihovoj interesantnosti, te njihova eliminacija svakako povećava upotrebljivost skupa otkrivenih pravila. Međutim, eksperimenti potvrđuju da je učinak eliminacije statistički očekivanih pravila u skupu svih web sesija nizak, te veliki broj potencjalno neinteresantnih pravila ostaje u skupu posle primene ove metode.

Metod predložen u okviru ovog istraživanja, kojim se eliminišu duga pravila, statistički očekivana u odnosu na kraća pravila koja takođe postoje u skupu svih pravila, primenjen je kao treći korak u procesu prečišćavanja skupa otkrivenih asocijativnih pravila. Učinak eliminacije na smanjenje veličine skupa dugih asocijativnih pravila u sprovedenim eksperimentima je izuzetno visok (od 73% do 93%, u zavisnosti od skupa podataka i vrednosti parametara), što značajno olakšava snalaženje analitičarima podataka pri odabiru potencijalno korisnih pravila.

U poslednjem, četvrtom koraku prečišćavanja primenjen je predloženi metod za eliminaciju pravila koja su statistički očekivana u odnosu na opštija pravila, pri čemu je relacija opštije/specifičnije pravilo definisana konceptnom hijerarhijom. Konceptna hijerarhija je konstruisana na osnovu otkrivenih kratkih asocijativnih pravila čija je confidence vrednost u blizini maksimalne. Obzirom da je u prethodnom, trećem koraku procesa prečišćavanja eliminisana ogromna većina dugih pravila, efikasnost ovog, poslednjeg koraka prečišćavanja merena je na skupu kratkih asocijativnih pravila, koji nije mogao biti prečišćen prethodno primenjenim metodama. Ukupni učinak ove eliminacije je u intervalu od 14% do 34%, u zavisnosti od skupa podataka, minimalnog support praga i minimalnog Z-score praga. Iako je učinak eliminacije kratkih pravila niži u odnosu na eliminaciju dugih pravila primenjenu u prethodnom koraku, eliminacija kratkih asocijativnih pravila je posebno značajna jer upravo njih analitičari podataka najčešće koriste.

Obzirom da sve primenjene metode prečišćavanja skupa otkrivenih asocijativnih pravila eliminišu samo ona pravila koja su statistički očekivana u skupu svih web sesija, ili u odnosu na postojeća opštija pravila, one ne dovode do gubitka potencijalno interesantnih pravila. Dakle, učinak smanjenja veličine skupa otkrivenih asocijativnih

pravila primenom ovih metoda direktno je proporcionalan povećanju kvaliteta, odnosno upotrebljivosti skupa otkrivenih asocijativnih pravila.

U poslednjoj fazi eksperimentalnog istraživanja vrši se rangiranje asocijativnih pravila koja preostaju u skupu otkrivenih pravila posle njegovog prečišćavanja. Iako je predložena modifikacija standardnih mera interesantnosti asocijativnih pravila o korišćenju web sajtova primenljiva na opšti oblik asocijativnih pravila, eksperimentalno istraživanje je ograničeno na rangiranje kratkih asocijativnih pravila. Naime, posle prečišćavanja skupa otkrivenih pravila u njemu preostaje veoma mali broj dugih pravila, pa bi validnost eksperimenata kojima se ispituje njihovo rangiranje bila diskutabilna. Nasuprot toga, u skupu otkrivenih pravila preostaje veliki broj neprečišćenih kratkih pravila, čije ispravno rangiranje može znatno olakšati odabir stvarno korisnih pravila analitičarima podataka.

Pre samog rangiranja kratkih pravila primenom matematičkih mera interesantnosti, pravila su podeljena u četiri grupe, u zavisnosti od odnosa potkoncept/natkoncept između leve i desne strane pravila. Pravila iz prve i druge grupe otkrivaju znanje o interesovanjima posetioca web sajtova, dok pravila iz treće i četvrte grupe govore o upotrebljivosti direktnih hiperlinkova između web stranica koje čine levu, odnosno desnu stranu pravila. Obzirom da analiza upotrebljivosti direktnih hiperlinkova nije fokus ovog istraživanja, u eksperimentima su rangirana pravila iz prve dve grupe.

Rangiranje pravila je u eksperimentima vršeno prema različitim merama interesantnosti, a sličnost rangiranja je merena koristeći Spearmanov koeficijent korelacije ranga. Pokazano je da confidence i added value mera čine jedan par standardnih mera interesantnosti jer konzistentno daju slično rangiranje, za različite vrednosti minimalnog support praga i različite skupove podataka. Nasuprot toga, lift i Z-score mera čine drugi par mera interesantnosti jer takođe daju slično rangiranje, koje se razlikuje od rangiranja prema confidence, odnosno added value merama.

Postoji mnoštvo istraživanja koje se bavi razvojem kriterijuma za odabir adekvatnih mera interesantnosti za dati domen primene i dati skup podataka. Činjenica da dve ili više mera daju slično rangiranje za različit izbor parametara, ukazuje na to da je dovoljno pri rangiranju koristiti samo jednu od takvih mera kao predstavnika skupa

sličnih mera. Dakle, dobijeni rezultat se može koristiti u cilju smanjenja ukupnog broja ponuđenih mera, što analitičaru podataka olakšava odabir optimalne mere. Ovo istraživanje se u budućnosti može proširiti tako da obuhvati veći broj mera interesantnosti, koje se pri tome mogu klasterovati u odnosu na rezultate rangiranja.

Kroz primere rangiranja pravila iz stvarnih skupova podataka potvrđeno je da modifikacija standardnih statističkih mera interesantnosti smanjuje rang pravila iz druge grupe, čija su leva i desna strana potkoncept zajedničkog natkoncepta. Intuitivno, ovakva modifikacija bolje odgovara stvarnoj interesantnosti pravila sa stanovišta analitičara podataka. Neko buduće istraživanje u kojem bi se preciznije merila stvarna interesantnost asocijativnih pravila moralo bi uključiti poređenje rangiranja modifikovanim i standardnim merama interesantnosti sa rangiranjem od strane ljudskih eksperata. Međutim, problemi sa ovim pristupom su subjektivnost eksperata i raznovrsnost kriterijuma koje različiti eksperti koriste u različitim situacijama, zbog čega se u većini postojećih istraživanja ovaj pristup izbegava.

Eksperimenti u kojima se poredi rangiranje asocijativnih pravila standardnim i modifikovanim merama interesantnosti pokazuju da modikacija lift i Z-score mere više umanjuje interesantnost pravila čija leva i desna strana imaju zajednički natkoncept, nego što to čini modifikacija added value mere. Ova osobina modifikovanih mera interesantnosti može se uzeti u obzir prilikom definisanja smernica za odabir odgovarajućih mera interesantnosti od strane analitičara podataka na datom skupu podataka, što predstavlja aktuelnu oblast istraživanja.

Predložene metode za prečišćavanje i rangiranje asocijativnih pravila o korišćenju web sajtova moguće je primeniti i na opšti oblik asocijativnih pravila u različitim domenima. Metode koje se zasnivaju na postojanju konceptne hijerarhije definisane na način koji je u ovoj disertaciji predložen podrazumevaju da u skupu otkrivenih asocijativnih pravila postoji značajan broj asocijativnih pravila čija je confidence vrednost približna maksimalnoj. U opštem slučaju međutim, bilo bi moguće uvesti konceptnu hijerarhiju definisanu od strane eksperta, a potom vršiti eliminaciju i rangiranje otkrivenih asocijativnih pravila u odnosu na ekspertsku konceptnu hijerarhiju, što takođe predstavlja jedan od mogućih pravaca daljeg istraživanja.

8 Literatura

Abdullah, Z., Herawan, T., & Deris, M. M. (2014, January). Discovering Interesting Association Rules from Student Admission Dataset. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 135-142). Springer Singapore.

Aggarwal, C. C., Yu, P. S. (1998). A new framework for itemset generation, In PODS 98, Symposium on Principles of Database Systems, pages 18-24.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases, In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 207-216.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Agrawal, R., Mannila, H., Srikant, R., & Toivonen, H. (1995). Fast discovery of association rules. Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press.

Anand, S. S., Mulvenna, M., & Chevalier, K. (2004). On the deployment of web usage mining. In *Web mining: from web to semantic web* (pp. 23-42). Springer Berlin Heidelberg.

Ayad, A., El-Makky, N. M., & Taha, Y. I. (2001, April). Incremental Mining of Constrained Association Rules. In *SDM* (pp. 1-18).

Baralis, E., Cagliero, L., Cerquitelli, T., & Garza, P. (2012). Generalized association rule mining with constraints. *Information Sciences*, *194*, 68-84.

Becker, K., & Vanzin, M. (2010). O3R: Ontology-based mechanism for a human-centered environment targeted at the analysis of navigation patterns.*Knowledge-Based Systems*, *23*(5), 455-470.

Ben Said, Z., Guillet, F., Richard, P., Picarougne, F., & Blanchard, J. (2013, July). Visualisation of association rules based on a molecular representation. In*Information Visualisation (IV), 2013 17th International Conference* (pp. 577-581). IEEE.

Berendt, B. (2002). Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*, *6*(1), 37-59.

Berendt, B. (2004). *Web Mining: From Web to Semantic Web: First European Web Mining Forum, EWMF 2003, Cavtat-Dubrovnik, Croatia, September 22, 2003, Revised Selected and Invited Papers* (Vol. 1). Springer Science & Business Media.

Berendt, B., Hollink, L., Hollink, V., Luczak-Rösch, M., Möller, K., & Vallet, D. (2011, May). Usage analysis and the web of data. In *ACM SIGIR Forum*(Vol. 45, No. 1, pp. 63-69). ACM.

Bonchi, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D., ... & Ruggieri, S. (2001). Web log data warehousing and mining for intelligent web caching. *Data & Knowledge Engineering*, *39*(2), 165-189.

Borodin, A., Roberts, G. O., Rosenthal, J. S., & Tsaparas, P. (2001, April). Finding authorities and hubs from link structures on the world wide web. In*Proceedings of the 10th international conference on World Wide Web* (pp. 415-429). ACM.

Bodon, F. (2003). A fast apriori implementation. In Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations (FIMI'03)(Vol. 90).

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (1999, August). Using association rules for product assortment decisions: A case study. InProceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 254-260). ACM.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997, June). Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*(Vol. 26, No. 2, pp. 255-264). ACM.

Büchner, A. G., Baumgarten, M., Anand, S. S., Mulvenna, M. D., & Hughes, J. G. (2000). Navigation Pattern Discovery from Internet Data, B. Masand, M. Spiliopoulou (eds.) Advances in Web Usage Analysis and User Profiling, Lecturer Notes in Computer Science.

Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. *Expert Systems with Applications*,*39*(12), 11243-11249.

Carvalho, D. R., Freitas, A. A., & Ebecken, N. (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. In *Knowledge Discovery in Databases: PKDD 2005* (pp. 453-461). Springer Berlin Heidelberg.

Cercone, N., & An, A. (2002, November). Comparison of interestingness functions for learning web usage patterns. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 617-620). ACM.

Ceglar, A., & Roddick, J. F. (2006). Association mining. *ACM Computing Surveys (CSUR)*, *38*(2), 5.

Chakrabarti, S. (2000). Data mining for hypertext: A tutorial survey. *ACM SIGKDD Explorations Newsletter*, *1*(2), 1-11.

Chakrabati, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., ... & Tomkins, A. (1999). Mining the link structure of the World Wide Web. *IEEE Computer*, *32*(8), 60-67.

Cheung, W., & Zaiane, O. R. (2003, July). Incremental mining of frequent patterns without candidate generation or support constraint. In *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International* (pp. 111-116). IEEE.

Chitraa, V., Davamani, D., & Selvdoss, A. (2010). A survey on preprocessing methods for web usage data. IEEE, IJCSIS, Vol. 7 No. 3, March 2010.

Clifton, C., Cooley, R., & Rennie, J. (2004). TopCat: data mining for topic identification in a text corpus. *Knowledge and Data Engineering, IEEE Transactions on*, *16*(8), 949-964.

Cooley, R. (2003). The use of web structure and content to identify subjectively interesting web usage patterns. ACM Transactions on Internet Technology (TOIT), 3(2), 93-116.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, *1*(1), 5-32.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, *41*(6), 391-407.

Dettmar G. (2004). *Logfile preprocessing using WUMprep*. Talk given at the Web Mining Seminar in Winter semester 2003/04, School of Business and Economics, Humboldt University Berlin, Berlin.

Devi, B. N., Devi, Y. R., Rani, B. P., & Rao, R. R. (2012). Design and implementation of web usage mining intelligent system in the field of e-commerce. *Procedia Engineering*, *30*, 20-27.

Ding, C. H., Zha, H., He, X., Husbands, P., & Simon, H. D. (2004). Link analysis: hubs and authorities on the World Wide Web. *SIAM review*, *46*(2), 256-268.

Ding, C., He, X., Husbands, P., Zha, H., & Simon, H. D. (2002, August). PageRank, HITS and a unified framework for link analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 353-354). ACM.

Dimitrijević, M., & Bošnjak, Z. (2014). Pruning Statistically Insignificant Association Rules in the Presence of High-confidence Rules in Web Usage Data. *Procedia Computer Science*, *35*, 271-280.

Dimitrijević, M., & Bošnjak, Z. (2010). Discovering interesting association rules in the web log usage data. *Interdisciplinary Journal of Information, Knowledge, and Management*, *5*, 191-207.

Dimitrijević, M., & Bošnjak, Z. (2011). Web Usage Association Rule Mining System. *Interdisciplinary Journal of Information, Knowledge, and Management*, *6*, 137-150.

Dimitrijević M., & Krunić T. (2014). Privacy preserving association rule mining applied to web usage data, *Monitoring and Expertise in Safety Engineering*, 4-3, 1-7.

Dimitrijević M., Krunić T., & Bošnjak Z. (2013). Association Rules for Improving Website Effectiveness: Case Analysis, *The Online Journal of Applied Knowledge Management*, 1, 56-63.

Dimitrijević M., Subić N. & Bošnjak Z. (2014). Improving the Interestingness of Web Usage Association Rules Containing Common Web Site Menu Items*, The Online Journal of Applied Knowledge Management*, 2, 82-92.

Dong, D. (2009, May). Exploration on web usage mining and its application. *Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on* (pp. 1-4). IEEE.

El-Hajj, M., & Zaïane, O. R. (2003, November). COFI-tree mining: A new approach to pattern growth with reduced candidacy generation. In *Workshop on Frequent Itemset Mining Implementations (FIMI'03) in conjunction with IEEE-ICDM*.

Etzioni, O. (1996). The World-Wide Web: quagmire or gold mine?.*Communications of the ACM*, *39*(11), 65-68.

Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, *53*(3), 225-241.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine.*Bioinformatics*, *27*(13), 1741-1748.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, *13*(3), 57.

Fu, Y., Shih, M. Y., Creado, M., & Ju, C. (2002). Reorganizing web sites based on user access patterns. *Intelligent Systems in Accounting, Finance and Management*, *11*(1), 39-53.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1993). *Design patterns: Abstraction and reuse of object-oriented design* (pp. 406-431). Springer Berlin Heidelberg

Gavalas, D., & Kenteris, M. (2011). A web-based pervasive recommendation system for mobile tourist guides. *Personal and Ubiquitous Computing*, *15*(7), 759-770.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), 38(3), 9.

Getoor, L., & Diehl, C. P. (2005). Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, *7*(2), 3-12.

Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., & Flake, G. W. (2002, May). Using web structure for classifying and describing web pages. In *Proceedings of the 11th international conference on World Wide Web* (pp. 562-569). ACM.

Gopalan, R. P., & Sucahyo, Y. G. (2004, April). High performance frequent patterns extraction using compressed FP-Tree. In *Proceedings of SIAM International Workshop on High Performance and Distributed Mining (HPDM), Orlando, USA*.

Hall, M., Frank E., Holmes G., Pfahringer, B., Reutemann P., Witten, I. H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 .

Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In *ACM SIGMOD Record* (Vol. 29, No. 2, pp. 1-12). ACM.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, *8*(1), 53-87.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques:* Elsevier.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.

Hilderman, R., & Hamilton, H. J. (2013). *Knowledge discovery and measures of interest* (Vol. 638). Springer Science & Business Media.

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Mining association rules: Deriving a superior algorithm by analyzing today's approaches. In *Principles of Data Mining and Knowledge Discovery* (pp. 159-168). Springer Berlin Heidelberg.

Hofmann, H., Siebes, A. P., & Wilhelm, A. F. (2000, August). Visualizing association rules with interactive mosaic plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 227-235). ACM.

Hong, T. P., Lin, K. Y., & Wang, S. L. (2003). Fuzzy data mining for interesting generalized association rules. *Fuzzy sets and systems*, *138*(2), 255-269.

Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A Brief Survey of Text Mining. In *Ldv Forum* (Vol. 20, No. 1, pp. 19-62).

da Jiménez, A., Berzal, F., & Cubero, J. C. (2013). Interestingness measures for association rules within groups. *Intelligent Data Analysis*, *17*(2), 195-215.

Joshi, K. P., Joshi, A., & Yesha, Y. (2003). On using a warehouse to analyze web logs. *Distributed and Parallel Databases*, *13*(2), 161-180.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. *Expert Systems with Applications*, *40*(4), 1034-1045.

Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). WEBSOM–self-organizing maps of document collections. *Neurocomputing*, *21*(1), 101-117.

Kautz, H., Selman, B., & Shah, M. (1997). The hidden web. *AI magazine*,*18*(2), 27.

Kazienko, P., & Pilarczyk, M. (2008). Hyperlink recommendation based on positive and negative association rules. *New Generation Computing*, *26*(3), 227-244.

Kazienko, P. (2009). Mining indirect association rules for web recommendation.*International Journal of Applied Mathematics and Computer Science*, *19*(1), 165-186.

Kazienko, P. (2007, January). Filtering of web recommendation lists using positive and negative usage patterns. In *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 1016-1023). Springer Berlin Heidelberg.

Khalil, F., Li, J., & Wang, H. (2006, November). A framework of combining Markov model with association rules for predicting web page accesses. In*Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61* (pp. 177-184). Australian Computer Society, Inc..

Kontonasios, K. N., Spyropoulou, E., & De Bie, T. (2012). Knowledge discovery interestingness measures based on unexpectedness. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(5), 386-399.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, *2*(1), 1-15.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

Langville, A. N., & Meyer, C. D. (2011). *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

Lazcorreta, E., Botella, F., & Fernández-Caballero, A. (2008). Towards personalized recommendation by two-step modified Apriori data mining algorithm. *Expert Systems with Applications*, *35*(3), 1422-1429.

Lee, C. H., Lo, Y. L., & Fu, Y. H. (2011). A novel prediction model based on hierarchical characteristic of web site. *Expert Systems with Applications*, *38*(4), 3422-3430.

Li, Y., & Feng, B. Q. (2009, June). The construction of transactions for web usage mining. In *Computational Intelligence and Natural Computing, 2009. CINC'09. International Conference on* (Vol. 1, pp. 121-124). IEEE.

Liang, B., Li, J., & Wang, K. (2004). Web page recommendation model for the semantic web. *JOURNAL-TSINGHUA UNIVERSITY*, *44*(9), 1272-1276.

Liberty, J. (2005). Programming C#: Building. NET Applications with C#. " O'Reilly Media, Inc.".

Lin, W., Alvarez, S. A., & Ruiz, C. (2002). Efficient adaptive-support association rule mining for recommender systems. *Data mining and knowledge discovery*, *6*(1), 83-105.

Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media.

Liu, B., Hsu, W., & Ma, Y. (1999, August). Pruning and summarizing the discovered associations. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 125-134). ACM.

Mabroukeh, N. R., & Ezeife, C. I. (2009, November). Using domain ontology for semantic web usage mining and next page prediction. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1677-1680). ACM.

Madria, S. K., Bhowmick, S. S., Ng, W. K., & Lim, E. P. (1999, September). Research issues in web data mining. In *DaWaK* (pp. 303-312).

Manda, P., McCarthy, F., & Bridges, S. M. (2013). Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships. *Journal of biomedical informatics*, *46*(5), 849-856.

Miner, G. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Mobasher, B. (2007). Data mining for web personalization. In *The adaptive web*(pp. 90-135). Springer Berlin Heidelberg.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Discovery and evaluation of aggregate usage profiles for web personalization. *Data mining and knowledge discovery*, *6*(1), 61-82.

Nanopoulos, A., Katsaros, D., & Manolopoulos, Y. (2002). Exploiting web log mining for web cache enhancement. In *WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points* (pp. 68-87). Springer Berlin Heidelberg.

Nebot, V., & Berlanga, R. (2012). Finding association rules in semantic web data. *Knowledge-Based Systems*, *25*(1), 51-62.0

Ng, R. T., Lakshmanan, L. V., Han, J., & Pang, A. (1998, June). Exploratory mining and pruning optimizations of constrained associations rules. In *ACM SIGMOD Record* (Vol. 27, No. 2, pp. 13-24). ACM.

Nguyen, L. T., Vo, B., Hong, T. P., & Thanh, H. C. (2013). CAR-Miner: An efficient algorithm for mining class-association rules. *Expert Systems with Applications*, *40*(6), 2305-2311.

Nguyen, L. T., & Nguyen, N. T. (2015). An improved algorithm for mining class association rules using the difference of Obidsets. *Expert Systems with Applications*, *42*(9), 4361-4369.

Niwa, S., & Honiden, S. (2006, April). Web page recommender system based on folksonomy mining for ITNG'06 submissions. In *Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on* (pp. 388-393). IEEE.

Oh, J., Lee, J., Kote, S., & Bandi, B. (2003). Multimedia data mining framework for raw video sequences. In *Mining Multimedia and Complex Data*(pp. 18-35). Springer Berlin Heidelberg.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining, Pearson, 2006.

Padmanabhan, B., & Tuzhilin, A. (2002). Knowledge refinement based on the discovery of unexpected patterns in data mining. *Decision Support Systems*,*33*(3), 309-321.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. Knowledge discovery in databases, 229-238.

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User modeling and user-adapted interaction*, *13*(4), 311-372.

Pohle, C., & Spiliopoulou, M. (2002). Building and exploiting ad hoc concept hierarchies for web log analysis. In *Data Warehousing and Knowledge Discovery* (pp. 83-93). Springer Berlin Heidelberg.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., & Lorensen, W. E. (1991). Object-oriented modeling and design (Vol. 199, No. 1). Englewood Cliffs: Prentice-hall.

Sahar, S. (2010). Interestingness measures-On determining what is interesting. In *Data mining and knowledge discovery handbook* (pp. 603-612). Springer US.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000, October). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce* (pp. 158-167). ACM.

Savasere, A., Omiecinski, E. R., & Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce* (pp. 115-153). Springer US.

Shaikh, M., & Beyene, J. (2015). Testing genotypes-phenotype relationships using permutation tests on association rules. *Statistical applications in genetics and molecular biology*, *14*(1), 83-92.

Senkul, P., & Salin, S. (2012). Improving pattern quality in web usage mining by using semantic information. *Knowledge and information systems*, *30*(3), 527-541.

Singh, B., & Singh, H. K. (2010, December). Web data mining research: a survey. In *Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on* (pp. 1-10). IEEE.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs journal on computing*, *15*(2), 171-190.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, *1*(2), 12-23.

Subašić, I., & Berendt, B. (2010). Discovery of interactive graphs for understanding and searching time-indexed corpora. *Knowledge and Information Systems*, *23*(3), 293-319.

Sucahyo, Y. G., & Gopalan, R. P. (2003, January). CT-ITL: Efficient frequent item set mining using a compressed prefix tree with pattern growth. In*Proceedings of the 14th Australasian database conference-Volume 17* (pp. 95-104). Australian Computer Society, Inc..

Szomszor, M., Cattuto, C., Alani, H., O'Hara, K., Baldassarri, A., Loreto, V., & Servedio, V. D. (2007). Folksonomies, the semantic web, and movie recommendation.

Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. Information Systems, 29(4), 293-313.

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (Vol. 8, p. 65).

Tao, Y. H., Hong, T. P., & Su, Y. M. (2008). Web usage mining with intentional browsing data. *Expert Systems with Applications*, *34*(3), 1893-1904.

Tseng, M. C., & Lin, W. Y. (2007). Efficient mining of generalized association rules with non-uniform minimum support. *Data & Knowledge Engineering*, *62*(1), 41-64.

Ventura, S., Romero, C., & Hervás, C. (2008, June). Analyzing rule evaluation measures with educational datasets: A framework to help the teacher. In*Educational Data Mining 2008*.

Wang, Y., & Zheng, L. (2012). Endocrine Hormones Association Rules Mining Based on Improved Apriori Algorithm. *Journal of Convergence Information Technology*, *7*(7).

Wang, X. S., Balasubramanian, A., Krishnamurthy, A., & Wetherall, D. (2013, April). Demystifying Page Load Performance with WProf. In *NSDI* (pp. 473-485).

Wu, X., & Kumar, V. (Eds.). (2009). *The top ten algorithms in data mining*. CRC Press.

Yan, W., Jiajin, L., & Dongmei, H. (2010, October). A method for privacy preserving mining of association rules based on web usage mining. In *Web Information Systems and Mining (WISM), 2010 International Conference on*(Vol. 1, pp. 33-37). IEEE.

Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(1), 60-70.

Yang, Q., & Zhang, H. H. (2003). Web-log mining for predictive web caching.*Knowledge and Data Engineering, IEEE Transactions on*, *15*(4), 1050-1053.

Zaiane, O. R., Han, J., Li, Z. N., Chee, S. H., & Chiang, J. Y. (1998, June). MultiMediaMiner: a system prototype for multimedia data mining. In *ACM SIGMOD Record* (Vol. 27, No. 2, pp. 581-583). ACM.

Zaki, M. J. (2004). Mining non-redundant association rules. *Data mining and knowledge discovery*, *9*(3), 223-248.

Zhang, Y., & Jiao, J. R. (2007). An associative classification-based recommendation system for personalization in B2C e-commerce applications.*Expert Systems with Applications*, *33*(2), 357-367.

Zhang, Y., Zhang, L., Nie, G., & Shi, Y. (2009, July). A survey of interestingness measures for association rules. In *Business Intelligence and Financial Engineering, 2009. BIFE'09. International Conference on* (pp. 460-463). IEEE.

Zhao, Q., & Bhowmick, S. S. (2003). Association rule mining: A survey.*Nanyang Technological University, Singapore*.

Zhou, B., Hui, S. C., & Fong, A. (2006, December). An effective approach for periodic web personalization. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on* (pp. 284-292). IEEE.