

UNIVERZITET SINGIDUNUM

DEPARTMAN ZA POSLEDIPLOMSKE STUDIJE

Data mining za Web personalizaciju

- Master rad -

Mentor:

Prof. dr Milan Milosavljević

Student:

Slavko Stanisavljević

Br. indeksa: 410209/2010

Beograd, 2012.

DEPARTMAN ZA POSLEDIPLOMSKE STUDIJE

UNIVERZITET SINGIDUNUM

DEPARTMAN ZA POSTDIPLOMSKE STUDIJE

Beograd, Danijelova 32

Kandidat: Slavko Stanisavljević

Broj indeksa: 410209/2010

Smer: Savremene informacione tehnologije

Tema: Data mining za Web personalizaciju

Datum odobrenja rada:
Beograd, ___.___._____.

MENTOR

_____

Prof. dr Milan Milosavljević

DEKAN

_____

Prof. dr Mladen Veinović

SADRŽAJ

1 Metodologija naučnog istraživanja ..... - 4 -

1.1 Predmet istraživanja ..... - 4 -

1.2 Hipotetički okvir istraživanja ..... - 4 -

1.3 Ciljevi i zadaci istraživanja ..... - 5 -

1.4 Metodi i tok istraživanja rada ..... - 5 -

2 Uvod ..... - 6 -

3 Data mining ..... - 8 -

3.1 Faze u procesu data mining-a ..... - 9 -

4 Automatska personalizacija i data mining ..... - 11 -

4.1 Pristupi personalizaciji ..... - 11 -

4.1.1 Sistem baziran na pravilima ..... - 12 -

4.1.2 Sistem filtriranja na osnovu sadržaju ..... - 12 -

4.1.3 Filtriranje na osnovu kolaboracije ..... - 13 -

4.2 Pristup korisničkom profilisanju ..... - 14 -

4.3 Data mining pristup personalizaciji ..... - 16 -

5 Prikupljanje podataka i preprocesiranje ..... - 18 -

5.1 Izvori podataka za koriscenje Web mininga ..... - 18 -

5.2 Podaci preprocesiranja za Web mining ..... - 20 -

6 Tehnike i alati otkrivanja paterna u Web mining-u ..... - 22 -

6.1 Analiza sesija i korisnika ..... - 23 -

6.2 Asocijativna i korelativna analiza ..... - 23 -

6.3 Klasterovanje i segmentacija korisnika ..... - 26 -

6.4 Analiza sekvencijalnih i navigacionih paterna ..... - 27 -

6.5 Klasifikacija i modeliranje korisnika ..... - 28 -

7 Analiza i alati za analizu otkrivenih paterna ..... - 29 -

8 Yahoo! Personalizacija ..... - 30 -

8.1 Oblasti za personalizaciju širih razmera ..... - 32 -
8.2 Privatnost I bezbednost ..... - 32 -
8.3 Korisnički interfejs ..... - 33 -
8.4 Neka opšta zapažanja i naučne lekcije ..... - 33 -
8.5 Učenje od korisnika ..... - 35 -
8.6 Zaključak ..... - 35 -
9 Zaključak ..... - 36 -
10 Literatura ..... - 37 -

1 Metodologija naučnog istraživanja

1.1 Predmet istraživanja

U ovom radu predstavljen je sveobuhvatan pregled data mining pristupa personalizaciji. Fokusiramo je pre svega na upotrebu Web mining-a gde je cilj da se iskoriste podaci koji su prikupljeni kao rezultat korisničke interakcije sa Web-om, kako bi saznali model korisnika i koristili ove modele za personalizaciju. Pružena je detaljna diskusija aktivnosti hosta data mining-a neophodnih za ovaj proces, uključujući preprocesuiranje i integraciju podataka iz vise izvora, zajednicki patern otkrivanja tehnike koje se primenjuju na ove podatke kako bi se izvukao agregat korisničkog modela i preporuka algoritama za kombinovanje otkrivenog znanja trenutnog statusa korisnika aktivnosti na nekoj Web lokaciji uz lični sadržaj.

1.2 Hipotetički okvir istraživanja

Hipotetički okvir istraživanja sastoji se iz sledećih hipoteza:

Generalna hipoteza:

Upotrebom Web mining-a postižemo bolje razumevanje posetilaca Web sajtova, stvaranja efikasnije ili korisnije organizacije za Web sajtove i efikasnijeg marketinga.

Posebna hipoteza:

- Osnovni cilj data mininga jeste otkrivanje do sada nepoznatih odnosa između podataka.
- Faze u procesu data mining-a
- Sa arhitektonske i algoritamske tačke gledišta u personalizaciju sistema spadaju tri osnovne kategorije: sistem baziran na pravilima, sadržaj za filtriranje sistema i kolabarativni sistem za filtriranje.
- Tipovi i nivoi analize korišćeni na integrisanim podacima zavise od primarnih ciljeva analitičara i željenih izlaza.

1.3 Ciljevi i zadaci istraživanja

Naučni cilj ovog rada je upoznavanje sa data miningom za Web personalizaciju, fazama u data miningu, hostu aktivnosti i tehnikama koje se primenjuju u različitim fazama ovog ciklusa uključujući predobradu i integraciju podataka iz više izvora kao i tehnike otkrivanja paterna koji se obično primenjuju na te podatke.

Društveni cilj ovog rada je predstavljanje data mininga za Web personalizaciju, fazama u procesu data mining, kao i prikaz najčešćih tipova otkrivanja paterna i analiza koji se koriste kod Web mining-a.

1.4 Metodi i tok istraživanja rada

U ovom radu, da bi se postigli postavljeni ciljevi kao i što efikasniji rezultat koji se dobijaju povezivanjem i odgovarajućom kombinacijom više naučnih metoda, korištene su sledeće metode:

- Metoda deskripcije
- Komparativna metoda
- Metoda analize
- Metoda sinteze

Metoda deskripcije, definisana kao postupak jednostavnog opisivanja ili ocrtavanja činjenica, procesa ili predmeta, te njihovih empirijskih potvrdjivanja odnosa i veza, ali bez naučnog tumačenja i objašnjavanja, je korišćena kroz ceo rad.

Komparativna metoda, koja podrazumeva upoređivanje istih ili srodnih činjenica, pojava, procesa i odnosa, odnosno utvrđivanja njihove sličnosti u ponašanju i intezitetu kao i razlika među njima.

Metoda analize, je postupak naučnog istraživanja i objašnjavanja stvarnosti putem raščlanjivanja složenih misaonih tvorevina (pojmova, sudova i zaključaka) na njihove jednostavnije delove i elemente, kao i izračunavanje svakog dela (elementa) za sebe i u odnosu na druge delove, celine.

Metoda sinteze, je suprotna metoda od analize koja se takođe primenjivala u izradi ovog rada.



2 Uvod

Krajnji cilj svakog korisničkog sistema je da obezbedi korisnicima ono što im treba bez eksplicitnih pitanja. Automatska personalizacija je glavna tehnologija koja se koristi u takvim sistemima. U kontekstu Weba, personalizacija podrazumeva isporuku dinamičkih sadržaja kao sto su tekstualni elementi, linkovi, reklame, preporuke o proizvodu itd, koji su prilagođeni potrebama i interesima određenih korisnika ili grupe korisnika. “Automatska personalizacija” se razlikuje od “prilagođavanja”. I prilagođavanje i personalizacija se odnose na isporuku sadržaja koji je prilagođen određenom korisniku. Ono sto razdvaja ova dva pojma je ko kontroliše kreiranje korisničkih profila kao i prezentaciju interfesa korisnika. U prilagođavanju, korisnici samostalno (često ručno) određuju svoje preferencije ili zahteve, na osnovu čega stvaraju interfejs. Automatska personalizacija, sa druge strane, podrazumeva stvaranje korisiničkih profila i potencijalno ažuriranje automatskih sistema sa minimalnom eksplicitnom kontrolom od strane korisnika. Primer za automatsku personalizaciju u komercijalnim sistemima uključuju personalizovane preporuke www.amazon.com , muziku ili listu preporuka, i razne vrste filtriranja agenata koji su danas dostupni.

Tradicionalni pristupi automatskoj personalizaciji su zasnovani na sadržaju, saradnji i pravilima filtriranja sistema. Svaki od ovih pristupa se razlikuje po specifičnom tipu prikupljanja podataka za izradu profila drugih korisnika, kao i određenoj vrsti algoritamskih pristupa radi pružanja personalizovanih sadržaja. Generalno, proces personalizacije se sastoji od prikupljanja podataka, faze u kojoj se dobijaju informacije koje se odnose na interese korisnika i fazu učenja u kojoj su korisnički profili izgrađeni od prikupljenih podataka. Učenja na osnovu podataka se mogu svrstati u učenja na osnovu memorije (takođe poznat kao ”lazy”) i model baziran na učenju u zavisnosti od toga da li je učenje online dok sistem obavlja personalizaciju zadatka ili offline korišćenjem prikupljenih podataka. Standardno korisničko kolaborativno filtriranje i većina sadržaja zasnovana na sistemu za filtriranje koji koriste “lazy” algoritme učenja su primeri memorije zasnovane na personalizaciji.

Sistemi zasnovani na memoriji jednostavno memorišu sve podatke i obrađuju ih sve vreme generisanja preporuke.

Model učenja koja obavljaju računarska učenja u offline fazi generalno imaju bolji obim nego memorije sistema zasnovane tokom online primene. S druge strane, memorijski sistemi su generalno bolji u prilagođavanju promenama interesa korisnika u odnosu na bazne modele gde model baziran na tehnici u kojoj on mora biti ili inkrementalan ili obnovljen kako bi primio nove podatke. Ove prednosti i nedostaci su doveli do raznih istraživanja u praksi koje se sastoje od raznih personalizacija ili preporučenih sistema koji generalno spadaju u pomenute kategorije.



Cilj u ovom poglavlju nije da se pruži uvid u automatsko prilagođavanje uopšteno, umesto toga fokus ce biti na Web personalizaciji, gde preporučeni predmeti dolaze iz spremišta Web objekata (stvari ili strana).

Kohavi[1] ukazuje na pet poželjnih stvari u data mining aplikacijama:

- Podaci bogati opisima koji omogućuju potragu izvan jednostavne korelacije.
- Veliki obim podataka za izradu pouzdanog modela.
- Kontrolisano i pouzdano prikupljanje podataka.
- Sposobnost procene razultata
- Lakoća integracije sa postojećim procesima

Retko su ovi kriterijumi zadovoljeni u tipičnoj data mining aplikaciji. Personalizacija na Webu a posebno e-commerce se smatra “killer app” (ubica aplikacija) za data mining, delom zato sto su mnogi od ovih elemenata zaista prisutni. Međutim, biti u mogućnosti da se u potpunosti iskoristi fleksibilnost podataka, kao i da se efikasno koristi otkriven model u automatski personalizovanim sistemima, proces personalizacije treba posmatrati kao aplikaciju data mining-a koja zahteva podršku za sve tipične faze data mining ciklusa, uključujući i prikupljanje podataka, predobradu, otkrivanje paterna i evaluaciju, offline režim i na kraju raspoređivanje znanja u realnom vremenu između korisnika i Web-a.

Prednosti i fleksibilnosti koju pružaju data mining pristupi kao i pristupi personalizaciji dolaze upravo iz činjenice da se personalizacija posmatra kao celovit proces, a ne kao pojedinačni algoritmi ili posebni tipovi podataka. Mnogi od tradicionalnih algoritama za personalizaciju se takođe mogu postaviti u kontekstu ovog procesa.

[1] Kohavi, R., Provost, F. Applications *of data mining to electronic commerce. Data Mining and Knowledge Discovery, 2001*

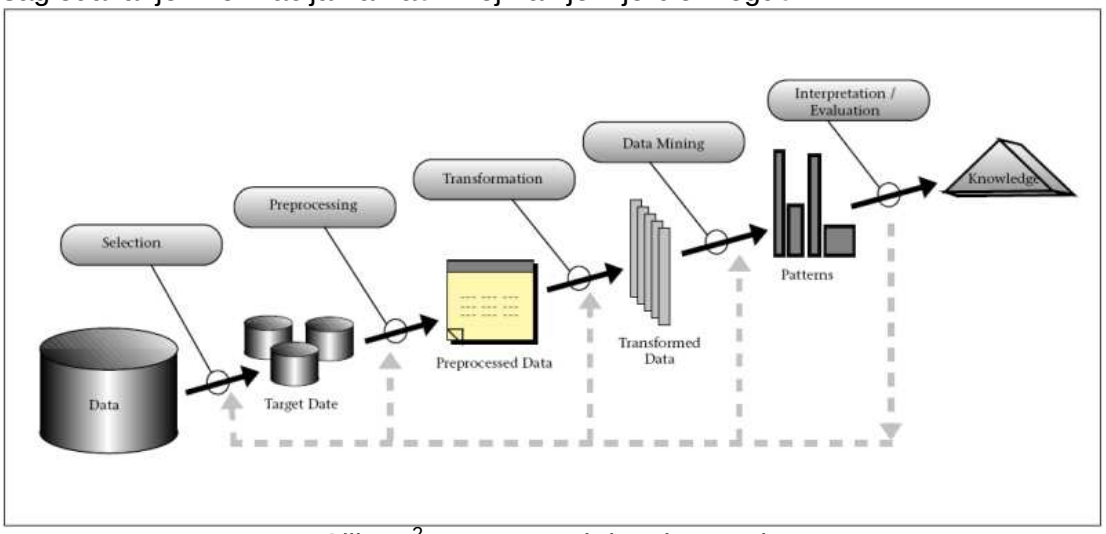
3 Data mining

Data mining je najvažniji proizvod iz familije Business Intelligence proizvoda, čija je svrha pronalaženje skrivenih obrazaca u podacima, povećavanje njihove upotrebljivosti i transformacija tih podataka u korisno znanje.

Postoji nekoliko definicija Data mininga. Data mining se može definisati kao proces pronalaženja skrivenih zakonitosti i veza među podacima. To je tehnika pretraživanja podataka u cilju identifikacije traženih uzoraka i njihovih međusobnih relacija. Jednostavno rečeno, data mining je postupak izdvajanja interesantnih, novih i potencijalno korisnih informacija ili uzoraka, sadržanih u velikim bazama podataka, a sve u cilju donošenja ispravnih poslovnih odluka.

Data mining je multidisiplinarno područje koje obuhvata: baze podataka, ekspertne sisteme, teoriju informacija, statistiku, matematiku, logiku i čitav niz drugih područja.

Data mining se zove i Knowledge Discovery in Databases (KDD) – otkrivanje znanja u bazama podataka. . To je proces analize koji omogućuje korisnicima da shvate sisteme i veze između njihovih podataka. On omogućava sagledavanje informacija na način koji ranije nije bio moguć.



Slika 1[2] – Proces otkrivanja znanja

Osnovni cilj data mininga jeste otkrivanje do sada nepoznatih odnosa između podataka. Analizom ogromnih baza podataka, upotrebom data mininga definišemo relacije, obrasce ili forme ponašanja, neophodne za odlučivanje i

[2] izvor - http://liris.cnrs.fr/abstract/

predviđanje. Prilikom same pretrage podataka data mining softver pomaže analitičaru da reši neke od sledećih problema[3] :

- Klasifikacija - analiziraju se skupovi podataka, otkrivaju skrivene veze i utvrđuju elementi (funkcije) za njihovo grupisanje u jednu od nekoliko klasa.

- Asocijacija podataka – utvrđuju se osobine koje se javljaju zajedno kod više uzoraka, odnosno veze među proizvoljnim atributima. Npr – koji se proizvodi kupuju zajedno u jednoj kupovini.

- Grupisanje (Klastering) – proces određivanja grupa podataka koji su međusobno slični, ali različiti od ostalih grupa podataka. Pri tome se indentifikuju i promenljive po kojima se vrši najbolje grupisanje.

- Predviđanje (Numeric prediction) – otkriva se ponašanje objekta posmatranja tokom vremena, i vrše se predviđanja. Utvrđuju se pravilnosti iz primera i na osnovu toga određuju očekivane numeričke vrednosti.

3.1 Faze u procesu data mining-a

Životni ciklus jednog data mining projekta se sastoji iz sledećih osam koraka :

1) Sakupljanje podataka – je obično prvi korak u data mining projektu. Poslovni podaci su uskladišteni u brojnim sistemima , internetu, bazama podataka kompanija, i prvi korak obično predstavlja prenos relevantnih podataka u bazu podataka gde se podaci analiziraju. Ponekad postoji i skladište podataka što olakšava dalji rad ali u velikom broju slučajeva podaci koji su sakupljeni mogu biti nedovoljno korisni za analizu te se zbog toga neophodni podaci moraju sakupiti iz drugih izvora. Nakon što se sakupe, podaci se mogu semplovati da bi se smanjila veličina trening skupa podataka. U mnogim slučajevima, obrasci koji su pronađeni na skupu od 50 000 kupaca su isti kao i oni pronađeni na trening skupu od 1 000 000 kupaca.

2) Filtiriranje podataka i transformacija – je najintenzivniji korak u data mining projektu kad su resursi u pitanju. Cilj filtriranja podataka je odstranjivanje irelevantnih i suvišnih informacija iz skupa podataka. To podrazumeva uklanjanje duplih i nepotpunih podataka, njihovu tranformaciju i jedinstven sistem podataka, izabiranje podgrupa podataka, određivanje broja promenjivih sa kojima je moguce raditi. Cilj transformacije podataka je promena izvornog podatka u drugačiji format tipa podataka. Postoje različiti tehnike koje se mogu primeniti za korak filtriranja i transformaciju podataka,

[3] http://en.wikipedia.org/wiki/Data_mining



a najčešće korišćene su; transformacija tipova podataka, neprekidna transformacija kolona, grupisanje, rad sa vrednošću koja nedostaje, brisanje abnormalnih slučajeva itd.

3) Kreiranje i izbor modela – je treći korak koji se primenjuje nakon filtriranja i transformacije podataka. Tek kada se podaci filtriraju i kada se promenljive transformišu u pogodne tipove podataka, može se započeti sa kreiranjem modela. Pre kreiranja modela treba da razumemo cilj data mining projekta i vrstu data mining zadatka koji će se koristiti. Za svaki data mining problem postoji nekoliko odgovarajućih algoritama. Preciznost algoritma zavisi od prirode podataka kao što su; broj stanja atributa koji se koriste za predviđanje, prenos vrednosti svakog atributa, veza između atributa itd. U ovom početnom delu projekta potrebno je sastaviti tim poslovnih analitičara koji su eksperti u određenoj oblasti.

4) Procena kvaliteta modela – U delu kreiranja modela mi kreiramo skup modela koristeći algoritme i tehnike data mining-a , ali nakon kreiranja moramo izvršiti i evaluaciju tog modela. Postoji nekoliko popularnih alata za evaluaciju kvaliteta modela. Najpoznatiji je lift dijagram. On koristi već istreniran model kako bi predvideo vrednosti koje će se dobiti iz skupa podataka koji se testira. Na osnovu vrednosti koje se dobiju i verovatnoće on grafički prikazuje model na dijagramu.

5) Kreiranje izveštaja – Nakon kreiranja modela i evaluacije kvaliteta tog modela vrši se kreiranje izveštaja koji se dostavljaju menadžerima na uvid. Većina data mining alata ima osobinu kreiranja izveštaja koji omogućuje korisnicima da generišu prethodno definisan izveštaj sa tekstualnim i grafičkim detaljima data mining modela. Postoje dva osnovna tipa izveštaja: izveštaji o pronađenim obrascima i izveštaji o predviđenim vrednostima modela.

6) Ocenjivanje modela – U mnogim data mining projektima, pronalaženje obrazaca i modela je samo pola posla; konačni cilj je upotreba tog modela za predviđanje. Predviđanje se još naziva i scoring u data mining terminologiji. Da bi dobili predviđene vrednosti moramo da imamo već istrenirani model i skup novih podataka.

7) Integracija data mining modela u aplikaciju – Integrisanje data mining modela u poslovne aplikacije predstavlja ponovnu primenu poslovne inteligencije na poslovni sistem tj. zatvaranje petlje za analizu. Sve više poslovnih aplikacija uključuje i data mining komponentu a prednosti data mining-a su velike. Na primer CRM (Customer Relationship Management) aplikacije mogu imati data mining osobine koje grupiše kupce u segmente, ERP (Enterprise Resource Planning) aplikacije mogu imati data mining osobine koje im koriste da predvide obim proizvodnje. On-line knjižara može dati potencijalnim kupcima preporuke knjiga. Integrisanje data mining osobina, pogotovo komponente za predviđanje u aplikacije jedan je od bitnijih koraka data mining projekta. Ovo je ključni korak za uvođenje data mining-a u masovnu upotrebu

8) Upravljanje modelom – Održavanje statusa data mining modela predstavlja pravi izazov. Svaki data mining model ima svoj životni ciklus. U nekim oblastima primene obrasci su relativno stabilni i modeli ne zahtevaju učestalo ponovno treniranje modela. Ali u mnogim oblastima obrasci se menjaju često. Trajanje jednog data mining modela je ograničeno. Nova verzija modela se mora praviti često. Određivanje preciznosti modela i kreiranje novih verzija ovog modela bi trebalo biti postignuto korišćenjem automatizovanih procesa.

4 Automatska personalizacija i data mining

Sposobnost personalizacije sistema da prilagodi sadržaj i preporuči stavke, podrazumeva da on mora biti u stanju da zaključi šta korisnik traži na osnovu prethodnih interakcija ili trenutne interakcije sa tim korisnikom, a verovatno i drugim korisnicima. Zadatak personalizacije može se posmatrati kao predviđanje problema: sistem mora pokušati da predvidi nivo interesovanja korisnika, ili korist od određene kategorije sadržaja, stranice ili stavki, i rangira ih u skladu sa predviđenim vrednostima.

Osim toga, zadatak pružanja personalizovanog sadržaja često je uramljena u smislu preporuka zadataka, u kojem sistem preporučuje stavke sa najvišim predviđenim interesnim vrednostima ili usluge za aktivnog korisnika. U principu, personalizacija sistema se može posmatrati kao mapiranje korisnika i stavki u skup “interesa od vrednosti”. Pogled na personalizaciju da funkcioniše kao predviđanje, proizilazi iz činjenice da mapiranje nije, uopšteno, definisano u celom domenu dostupnosti korisniku, a samim tim zahteva od sistema procenu vrednosti interesovanja za neke elemente u domenu.[4]

Automatski sistemi za personalizaciju, generalno, se razlikuju u vrsti podataka i metodima koji se koriste za kreiranje profila korisnika, dok se u domenu algoritamskog pristupa koristi predviđanje.

4.1 Pristupi personalizaciji

Sa arhitektonske i algoritamske tačke gledišta u personalizaciju sistema spadaju tri osnovne kategorije: *sistem baziran na pravilima*, *sadržaj za filtriranje sistema* i *kolabarativni sistem za filtriranje*. Primarni focus u ovom poglavlju je model baziran na kolabarativnom pristupu filtriranja u kojoj su modeli naučeni kroz različite data mining tehnike.

[4] Sandro Araya, Mariano Silva, Richard Weber, *A methodology for web usage mining and its application to target group identification*, Elsevier B.V. 2004



4.1.1 Sistem baziran na pravilima

Zasnovan na pravilima, sistem za filtriranje se oslanja na ručna ili automatski generisana pravila koja se koriste kao preporuka stavki korisnicima. Mnogi postojeći e-commerce Web sajtovi koji koriste personalizaciju ili preporuku tehnologije, koriste uputstvo zasnovano na pravilima sistema. Takvi sistemi omogućavaju Web sajt administratorima da odrede pravila, često zasnovana na demografskim, psihografskim ili drugim ličnim karakteristikama korisnika. U nekim slučajevima, pravila mogu biti zavisna od domena I da reflektuju određenje poslovne ciljeve sa Web sajta. Pravila se koriste da utiču na sadržaj kojim se sluzi korisnik čiji profil ispunjava jedan ili više pravila. Kao I većina sistema zasnovanim na pravilima, ovaj tip personalizacije oslanja se na znanje inženjeringa po sistemu dizajniranja baze pravila u skladu sa specifičnim karakteristikama domena ili istraživanja trzišta. Korisnički profili su uglavnom dobijeni kroz eksplicitne interakcije sa korisnicima. Neka istraživanja se fokusiraju na mašinskom učenju tehnike za klasifikaciju korisnika u jednoj od nekoliko kategorija na osnovu njihovih demografskih atributa, a samim tim automatski dobijaju pravila odlučivanja koja se mogu koristiti za personalizaciju.

Primarni nedostatak zasnovan na pravilima filtrirane tehnike, pored uobičajnog manjka znanja inženjerstva, potiče iz metode koje se koriste za generisanje korisničkih profila. Ulaz je obično subjektivan opis korisnika ili njihovih interesa od strane samih korisnika a samim tim je sklon pristrasnosti. Osim toga profili su često statični I na taj način se degradiraju perfomanse sistema tokom vremena, kako profil zastareva.

4.1.2 Sistem filtriranja na osnovu sadržaju

U ovom sistemu profil korisnika predstavlja opis sadržaja ili predmeta o kojima je korisnik ranije izrazio interes. Preporuka generacijskog zadatka u ovakvim sistemima obično uključuje poređenje izdvojenih karaktersitika od nevidljivih ili ne ocenjenih stvari u opisanom sadržaju korisničkog profila. Stavke koje se smatraju dovoljno sličnim profilima korisnika se preporučuju korisniku.

U vecini sistema filtriranja na osnovu sadržaja, posebno onih koji se koriste na Web-u i u e-commerce aplikacijama, opisi sadržaja su tekstualne karakteristike izdvojene sa Web stranice ili opisa proizvoda. Kao takvi, ovi sistemi se često oslanjaju na poznate dokumente tehnika modeliranja sa korenima u pronalaženju informacija filtriranja istraživanja. Predviđanje interesovanja korisnika za određenu stavku se može izvesti na osnovu izračunavanja vektora sličnosti (korišćenjem kosinusne sličnosti) ili korišćenjem pristupa verovatnoće kao sto su Bayesian-klasifikacija. Pored toga, u suprotnosti sa pristupima na osnovu kolaboracije, profili su individualni po prirodi, izgrađeni samo od karakteristika povezanih sa stavkama prethodno viđenim ili ocenjivanim od aktivnih korisnika.



Primeri ranih personalizovanih agenata koji koriste ovaj pristup su: Letezija, NewWeeder, Personal WebWatcher, InfoFinder, Syskill i Webert I Naïve Bayes, najblizi susedni pristup koristi Schwab.

Primarni nedostatak filtriranja na osnovu sadržaja je njihova tendencija da specializuju jedinice selekcije, jer profili se isključivo zasnivaju na prethodnom rejtingu korisnika. Korisničke studije su pokazale da korisnici pronalaze online preporuke najkorisnije kada preporučuju neočekivane stvari, što sugeriše da korišćenje slicnog sadržaja takođe može rezultirati nedostajanju važnog "pragmatičnog" odnosa između Web objekata kao što su zajedničke ili komplemetarne koristi u kontekstu konkretnog zadatka. Pored toga, sistem zasnovan na filtriranju na osnovu sadržaja zahteva da se stvari efikasno predstave uz pomoć ekstrahovanih tekstualnih karakteristika, koje ne daju uvek praktičnu hetoregenu prirodu Web podataka.

4.1.3 Filtriranje na osnovu kolaboracije

Kolaborativnim filtriranjem je pokušano da se nadomeste neki nedostaci ostalih pristupa. Posebno u kontekstu e-commerca, preporuke kod sistema zasnovanih na kolaborativnom filtriranju su postigle značajne uspehe. Generalno, ove tehnike obuhvataju podudaranja rejtinga trenutnih korisnika za objekte (filmove ili proizvode) sa sličnim korisnicima u cilju da proizvedu preporuke za objekte koje još nisu ocenili ili videli aktivni korisnici. Tradicionalno, primarna tehnika koja se koristi da bi se izvršio ovaj zadatak je standardna memorija na bazi k-Nearst-Neighbor (kNN) klasifikaciji pristupa, koji poredi profile ciljanih korisnika sa istoriskim profilima drugih korisnika u cilju pronalaženja top korisnika koji imaju slične ukuse I interesovanja.

Međutim, tehnike kolaborativnog filtriranja imaju svoja potencijalno ozbiljna ograničenja. Najvažnije od ovih ograničenja je zbog nedostatka skalabilnosti. U suštini, kNN zahteva da se susedno formiranje faza vrši kao online proces (tj. modeliranje faze se obavlja u realnom vremenu, za razliku od modela na bazi pristupa u kojem se model učenja obavlja offline). Kako se broj korisnika I predmeta povećava, ovaj pristup može dovesti do neprihvatljivog kašnjenja pružanja preporuke ili dinamičkog sadržaja u toku interakcije korisnika.

Drugo ograničenje kNN tehnike proizilazi iz oskudne prirode skupa podataka. Kako se broj stavki u bazi podataka povećava, gustina zapisa svakog korisnika u vezi sa ovim stavkama će se smanjivati. Ovo će značajno smanjiti mogućnost preklapanja posečenih ili ocenjenih predmeta među parovima korisnika, što dovodi do manje pouzdane izračunate korelacije. Osim toga, kolaborativno filtriranje se obično obavlja najbolje kad eksplicitni ne binarni korisnici daju ocene za slične objekte na raspolaganju. Na mnogim Web lokacijama međutim, može biti poželjno da se integriše prilagođena akcija širom sajta, uključujući različiti tip objekta, navigacione i stranice sa sadržajima, kao i

implicitno orjentisane proizvode korisnika kao što su korpa za kupovinu ili informacioni zahtevi proizvoda. [5]

Brojne strategije optimizacije predložene su i primenjene da bi se popravili nedostaci. Ove strategije obuhvataju sličnost indeksiranja i dimenzionalnosti smanjenja da se smanje troškovi u realnom vremenu za pretragu i otklone problemi, kao i offline grupisanje zapisnika korisnika, omogućavajući online komponentama sistema da traže u odgovarajućem klasteru. Model zasnovan na varijanti kolaborativnog filtriranja je poznat kao stavka u kojoj je, počev od istog korisničkog rejtinga profila baze podataka, stavka – tačka sličnosti matrica koja se gradi offline, i koristi se u fazi predviđanja za generisanje preporuka. Umesto zasnivanja stvari na sličnosti sadržaja opisa stavki, sličnost između stavki se zasniva na korisničkim ocenama. Svaka stavka predstavlja vektor, a sličnosti se obračunavaju na osnovu pokazatelja kao što su kosinus sličnosti i povezanost na bazi sličnosti. Proces preporuke predviđa rejting za stavke koje nisu ranije videte ili ocenjene od aktivnog korisnika korišćenjem ponderisanog zbira rejtinga stavki u tački u okolini stavke cilja. Evoluacija stavke pristupa kolaborativnog filtriranja je pokazala da stavka kolaborativnog filtriranja može dati preporuke koje su sličnog kvaliteta u odnosu na memoriju kolaborativnog pristupa.

Većina mining podataka personalizacije može da se posmatra kao proširenje kolaborativnog filtriranja. U ovim pristupima obrazac otkrića algoritama predstavlja ulaz istoriske ocene ili navigacionih profila prošlosti korisnika i generišu ukupni model korisnika. Korisnik modela, za uzvrat, može da se koristi zajedno sa profilom aktivnog korisnika, da bi predvideo buduće ponašanje korisnika ili generisao preporuke.

4.2 Pristup korisničkom profilisanju

Svi pristupi personalizaciji, kao i u većoj meri, personalizacija na osnovu data mininga, zahteva prikupljanje podataka koji tačno odražava interese korisnika i njihove interakcije sa aplikacijama i stavkama. Personalizovani sistemi razlikuju se ne samo u algoritmima koji se koriste za generisanje, preporuke ili predviđanja, vec i u načinu na koji se grade korisnički profili na osnovu pratećih podataka.

Personalizacija zasnovana na pravilima i sadržaju sistema generalno, grade individualni model interesa korisnika, i koriste ovaj profil da prilagode budućim interakcijama samo tom korisniku. Kao što je pomenuto ranije, sistem zasnovan na filtriranju sadržaja zahteva karakteristike stavki izdvojene iz opisa ili relacionih atributa saradnika sa stavkama u pozadinskoj bazi podataka. U takvim sistemima proces izgradnje profila korisnika zahteva dve faze. Prvo sistem mora da utvrdi nivo interesovanja korisnika u podskupu stavki. Ovaj zadatak se može implicitno

[5] Mobasher, B. Web *usage mining and personalization*. In Singh, M.P., ed.: Practical Handbook of Internet Computing. CRC Press 2005

postići pasivnim posmatranjem korisnika, i koristeći razne heuristike da klasifikuju stavke kao interesantne ili ne interesantne, ili se može zasnivati na eksplicitnim presudama korisnika dodeljivanjem ocene na stavke ili ručno indetifikovanje pozitivnih i negativnih primera. Glavni nedostatak pristupa na osnovu pojedinačnih profila je nedostatak slučajnog otkrivanja kao preporuke koji su veoma fokusirani na prethodne interese korisnika. Takođe sistem zavisi od dostupnosti opisnog sadržaja stavki koje se preporučuju.[6]

U slučaju zasnovanom na pravilima sistema, naročito onih zasnovanih na demografskom filtriranju, svaki profil korisnika se može predstaviti vektorom ličnih i demografskih karakteristika, često nazvanih otisak prsta (finger print). U e-commercu i analitičkoj Web aplikaciji otisak prsta posetilaca takođe može da sadrži izračunate atribute kao ukupan iznos potrošen i učestalost kupovine ili posete. Malo sistema koristi demografske podatke u okviru preporuke procesa. Ovo je zbog činjenice da se takvi podaci mnogo teže prikupljaju na Web-u, a kada se prikupe teže da budu lošeg kvaliteta. Takođe, preporuke zasnovane samo na demografskim podacima su se pokazale manje preciznim od onih zasnovanih na sadržaju stavki i ponšanju korisnika.

U kolaborativnom filtriranju sistem koristi ne samo profil aktivnog korisnika već takođe i vodi bazu podataka o profilima drugih korisnika. Za razliku od sadržaja filtriranja u kojoj stavka na stavku sličnosti čine osnovu za preporuku generisanja, kolaborativni sistemi se oslanjaju na sličnosti od korisnika do korisnika. Profili su uglavnom predstavljeni kao vektori ili skup ocena pružanja preferencija korisnika na podskup stavki. Aktivan korisnički profil se koristi da bi se pronašli drugi korisnici sa sličnim preferencijama, takozvane aktivne korisničke komsije. Imajte na umu da za razliku od sadržaja zasnovanog na filtriranju, stvarni opisi sadržaja artikla nisu deo profila.

Dok tradicionalno kolaboratovno filtriranje koristi samo ocenjene podatake, hibridni kolaborativni pristup koristi i sadržaj i rejting podataka korisnika. Osim toga u oba slučaja kolaborativnog i filtriranog sadržaja, istraživani su razni pristupi za integraciju ontološkog domena korisničkog profila. U prisustvu domena ontolgije, korisnik profila može zapravo reflektovati strukturu domena i na taj način zahtevati složenije zastupljenosti u odnosu na ravno predstavljanje korisnika u standarnim pristupima.

Bez obzira na algoritamski pristup personalizaciji, profilisani podaci korisnika mogu biti prikupljeni implicitno ili eksplicitno. Eksplicitno prikupljanje obično zahteva od korisnika aktivno učešće. U sistemima koji se oslanjaju na demografske i lične podatke, interakcija korisnika može poprimiti oblik učešća u online anketama u vreme registracije ili pružanje ličnih i finansijskih podataka prilikom kupovine (koji se onda mogu kombinovati sa offline demografskim podacima dostupnih kroz različite podatke agregacija usluga). Slično tome, sistem zasnovan na sadržaju filtriranja može takođe koristiti bilo direktno ili indirektno

[6] Mobasher, B. Web *usage mining and personalization*. In Singh, M.P., ed.: Practical Handbook of Internet Computing. CRC Press 2005



povratne informacije korisnika da odredi njihov nivo interesovanja u određenim tačkama. Tradicionalni kolaborativni sistem filtriranja koji se koristi u e-commercu, eksplicitne informacije korisnika koristi u vidu ocene o pojedinačnim stavkama. Međutim, mnogi kolaborativni sistemi, posebno Web personalizovani sistemi koji koriste *clickstream*[7] ili druge vrste ponašanja podataka, pokušavaju da mere interesovanja korisnika individualno ili u grupi stavki na osnovu heurističkog pokazatelja (kao što su vreme provedeno u gledanju stavke, bez obzira da li je stavka kupljena itd). Mnogi e-commerc sajtovi, kao što su amazon.com prati kupovinu svakog klijenta i istoriju aktivnosti i koristi informacije kao deo korisničkog profila.

Prednost korišćenja implicitno povratne informacije korisnika je da uklanja teret u vezi sa obezbeđivanjem ličnih podataka od korisnika. Sistem prikuplja relevantne podatke na osnovu posmatranog ponašanja korisnika i zaključuje specifične podatke o njemu. Implicitno profilisanje podrazumeva da sistem mora biti u stanju da prati ponašanje korisnika kako bi se utvrdili obrasci pretraživanja ili kupovine. Implicitni podaci mogu biti prikupljeni od klijenta ili od strane servera. Pristupi klasifikaciji mogu se klasifikovati na osnovu toga da li su ovi pristupi razvijeni da rade na strani klijenta ili na strani server. Ključna razlika između ovih personalizovanih pristupa je obim podataka koji je dostupan za personalizaciju sistema. Na strani klijenta, podaci su dostupni samo za individualne korisnike i samim tim pristup je moguće ostvariti samo individualno. Na strani servera, posao ima sposobnost da prikupi podatke o svim svojim posetiocima, i samim tim kako individualnih tako i kolaborativnih pristupa koji se primenjuju. S druge strane serveri obično imaju pristup interakcijama korisnika sa sadržajem na njihovom Web sajtu, a klijent može da pristupi podacima o interakciji pojedinaca sa više Web sajtova.

Većina klijentskih aplikacija zasnovana je na sadržaju sistema u cilju personalizovanih pretraga preko Web-a ili višestrukih spremišta. Nedostatak zajedničkog domena antologije preko Web sajta je nestruktuirana strana Web-a i proređena raspoloživost podataka koja smanjuje mogučnost za navigacionu personalizaciju nasuprot traženju na osnovu interakcije sa Web-om kao celinom.

Kolaborativna personalizacija sistema bazirana na korišćenju Web-a se oslanja na clickstream i navigacione podatke automatski prikupljene od strane Web-a i aplikacionih servera i skladišta u serverskom fajlu. Drugi izvor podataka o klijentima su transakcije baze podataka, pre prodaje i posle prodaje podržavanih podataka ili demografske informacije. Takvi podaci mogu biti dinamički prikupljeni sa Web lokacija ili kupljene od treće strane. U mnogim slučajevima podaci se čuvaju u različitim formatima u više različitih baza podataka.

4.3 Data mining pristup personalizaciji

[7] Pojam clickstream se odnosi na tok uzastopnih klikova (akcija) određenog korisnika na nekom Web sajtu.



Ono sto čini data mining pristup Web personalizaciji različit od drugog pristupa je da upotreba Web mining nije specifičan algoritam već sledi tipičan data mining ciklus. Kao takva, ona pruža veliku fleksibilnost za usklađivanje različitih podataka kanala na sveobuhvatan način i dozvoljava personalizaciji zadataka da budu bolje integrisani sa drugim postojećim aplikacijama. Osim toga, zbog fokusa data mining-a na efikasan model zasnovan na algoritmima, personalizovani sistem na osnovu data mining-a ima tendenciju da bude više skalabilniji od onih zasnovanih na tradicionalnom pristupu kao što je standardno kolaborativno filtriranje. Upotreba Web mining-a se može definisati kao automatsko otkrivanje i analiza šablona u clickstreamu i povezanih prikupljenih podataka nastale kao rezultat interakcije korisnika sa Web izvorom na jednom ili vise sajtova. Cilj upotrebe Web mining-a je da zarobi model i analizira modele ponašanja i profila korisnika u interakciji sa Web sajtom. Otkriveni obrasci se obično predstavljaju kao kolekcije stranica, objekata ili resursa koji su često priključeni grupama korisnika sa zajedničkim potrebama i interesima.[8]

Tradicionalno, cilj upotrebe Web mining-a je bio da podrži donošenje odluka procesa od strane operatora Web sajta u cilju sticanja boljeg razumevanja njihovih posetilaca, stvaranja efikasnije ili korisnije organizacije za Web sajtove i efikasnijeg marketinga. Međutim, ovi modeli se takođe mogu koristiti kao automatski adaptivni sistemi radi postizanja različitih funkcija za personalizaciju.

Ceo proces Web personalizacije baziran je na korišćenju Web mining-a koji se sastoji od tri faze: Priprema podataka i transformacija, patern otkrića i preporuke. Od ovih samo kasnija faza se obavlja u realnom vremenu.

Pripremna faza transformiše sirove Web log fajlove u profil korisnika ili Web transakcije podataka koji mogu biti procesuirani ili obrađeni od strane data mining zadataka. Ova faza takođe uključuje integraciju podataka iz više izvora, kao što su pozadina baze podataka, aplikacioni server i sadržaj sajta. Rezultati data mining faze se transformišu u agregatne korisničke modele, pogodne za upotrebu u preporučenoj fazi. Preporučeni motor smatra aktivni profil korisnika u saradnji sa otkrivenim paternom za pružanjem personalizovanog sadržaja.

[8] Sandro Araya, Mariano Silva, Richard Weber, *A methodology for web usage mining and its application to target group identification*, Elsevier B.V. 2004

5 Prikupljanje podataka i preprocesiranje

Gledanje na personalizaciju kao data mining aplikaciju ima za cilj da kreira skup korisnički orjentisanih modela podataka (korisničke profile), koji predstavljaju interese i aktivnosti svih korisnika koji se mogu koristiti kao ulaz za različita učenja mašinskog algoritma za obrazac (patern) otkrića. Izlaz iz ovih algoritama, odnosno šeme otkrića se mogu koristiti za predviđanje budućih interesa korisnika. Tačno predstavljanje ovih modela korisnika se razlikuje u zavisnosti od pristupa za postizanje personalizacije i preciznosti dostupnih informacija. Zadaci obrazaca otkrića se stoga razlikuju po složenosti zasnovane na ekspresivnosti predstavljanja odabranog profila korisnika i podataka na raspolaganju.

5.1 Izvori podataka za koriscenje Web mininga

Primarni izvori podataka koji se koriste u upotrebi Web mining-a su nekoliko log fajlova koji obuhvataju pristupe logovima Web servera i logove aplikacionih servera. Dodatni izvori podataka koji su takođe od suštinskog značaja za pripremu podataka i obrazca otkrića uključuje lokaciju fajlova i meta podataka (uključujući sadržaj funkcije i strukturne elemente strana), operativne baze podataka, aplikacioni templejti i domen znanja. U nekim slučajevima i za neke korisnike, dodatni podaci mogu biti dostupni zahvaljujući klijent strani ili proxy nivou prikupljanja podataka (internet provajder), kao i od spoljnih clickstream ili demografskih izvora podataka (npr. ComScore, NetRatings, MediaMetrix i Acxiom).[9]

Najvažniji od ovih izvora za upotrebu Web mininga su clickstream podaci zabeleženi automatski od strane Web-a i aplikacionih servera u log fajlove. Ovi podaci predstavljaju fine strukture navigacionih ponašanja posetioca. Svaki pristup serveru koji odgovara HTTP zahtevu generiše jedan unos u evidencijama pristupa servera. Svaki unos loga (u zavisnosti od formata) može da sadrži polja indetifikacije vremena i datuma zahteva, ip adresu klijenta, izvor zahteva, moguće parametre korišćene u pozivanju Web aplikacije, status zahteva, HTTP metod koji se koristi, agent korisnika (tip i verzija browsera i operativnog sistema), odgovarajući resurs Weba i, ako su dostupni, na strani klijenta cookies koji identifikuju jedinstvenog ponovnog posetioca.

U zavistnosti od ciljeva analize ovi podaci treba da se transformišu i prikupe na različitim nivoima apstrakcije. U upotrebi Web mining-a najosnovniji nivo

[9] Sandro Araya, Mariano Silva, Richard Weber, *A methodology for web usage mining and its application to target group identification*, Elsevier B.V. 2004



podataka apstrakcije je prikaz stranice. Prikaz stanice je ukupna zastupljenost kolekcije Web objekata koji doprinosi prikazu na Web pretrazivaču zahvaljujući jednoj akciji korisnika (kao što je clicktrue). Konceptualno, svaki prikaz stranice može da se posmatra kao skup Web objekata ili sredstava koji predstavljaju specifičan „događaj korisnika“, na primer čitanje članaka, pregled stranice proizvoda ili dodavanje proizvoda u kupovnoj korpi. Na nivou korisnika najosnovniji nivo ponašanja apstrakcije jeste sesija. Sesija je niz prikaza stranica od strane jednog korisnika tokom jedne posete. Pojam sesija može se dalje apstrahovati i izborom podskupa prikaza stranica u sesije koje su značajne ili relevantne za analizu zadataka. Sesija se moze koristiti direktno kao profil korisnika, međutim ako je cilj analize da snimi ponašanje korisnika tokom vremena (na primer preko više sesija) sve sesije koje pripadaju korisniku se mogu kombinovati i prikupiti da bi se kreirao profil za tog korisnika.

Sadržaj podataka na sajtu je kolekcija objekata i odnosa koji se prenose korisniku. U najvećem delu, ovi podaci se sastoje od kombinacija tekstualnih materijala ili slika. Izvori podataka korišćeni za isporuku ili generisanje ovih podataka uključuje statičku HTML/XML stranu, multimedijalne fajlove, dinamički generisane strane segmenata i skripte i zbirke podataka iz operativne baze podataka. Sadržaj podataka takođe uključuje semantičke ili strukturalne meta podatke ugrađene u sajt ili individualne stranice, kao što su opisne ključne reči, dokument atributa, semantički tagovi ili HTTP varijable. [10]

Strukturu podataka predstavlja prikaz dizajnera sadržaja organizacije u okviru sajta. Ova organizacija je zarobljena preko inter stranica povezane strukturom između strane, što se ogleda kroz hiper veze. Struktura podataka takođe sadrži unutar stranica strukturu sadržaja u okviru stranice. Na primer, HTML i XML dokumenta mogu biti predstavljena kao stablo struktura kroz prostor tagova na stranici. Hiper veza strukture za sajt je obično zarobljena od strane automatski generisane „mape sajta“, obično predstavljanih kao usmereni grafikon. Sajt mapiranih alatka mora da ima mogućnost da snimi i predstavi unutrašnji i spoljašnji prikaz stranica. Za dinamički generisane stranice sajt mapiranja alatki mora ili uključiti suštinska znanja osnovnih aplikacija i skripti ili mora da ima sposobnost da stvori segmente sadržaja korišćenjem uzorka parametara prosleđene takvoj aplikaciji ili skripti.

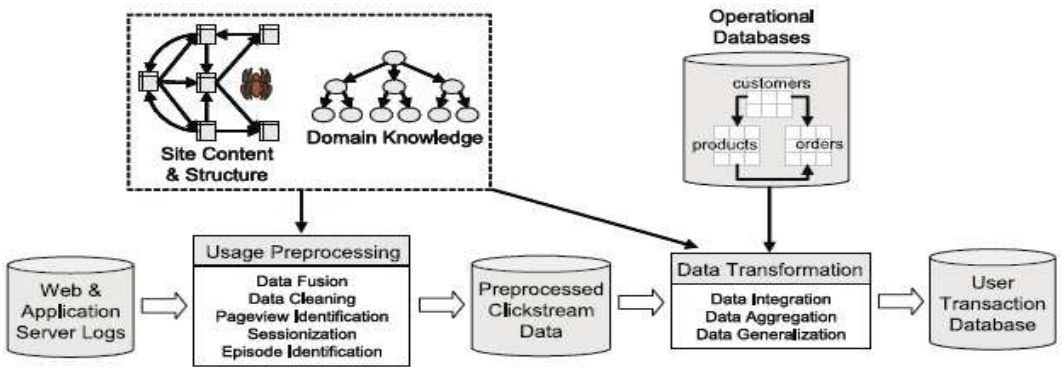
Na kraju, operativne baze podataka za sajt mogu sadržati dodatne informacije o korisniku i stavkama. Takvi podaci mogu podrazumevati demografske informacije o registrovanim korisnicama, ocene korisnika za razne predmete kao što su proizvodi ili filmovi, poslednja kupovina ili istorija posete korisnika, kao i drugi eksplicitno ili implicitno predstavljeni interesi korisnika. Proizvod baze podataka ili sistema za upravljanje sadržajem može da sadrži i dodatne sadržaje deskriptora i relacionih atributa koji se mogu koristiti kao deo reprezantacije sadržaja informacija za stavke. Neki od ovih podataka se može

[10] Baoyao Zhou, Siu Cheung Hui, and Alvis C. M. Fong, *Web Usage Mining for Semantic Web Personalization*, School of Computer Engineering, Nanyang Technological University, Singapore, 2004

snimiti anonimno dokle god postoji mogućnost da se napravi razlika između različitih korisnika.

5.2 Podaci preprocesiranja za Web mining

Cilj faze preprocesuiranja u upotrebi Web mining-a je da transformiše sirove podatke clickstream-a u skup korisničkih profila. Sa navigacione tačke gledišta svaki takav profil obuhvata razgraničen niz ili skup pregleda stranica predstavljajući korisničke sesije. Ovi podaci se mogu koristiti kao ulaz za različite data mining algoritme ili dalju transformisanost i apstrahovanost. Korišćenje preprocesuiranja Web podataka predstavlja jedinstven broj izazova koji su dovele do različitih algoritama i heuristickih tehnika za preprocesuiranje podataka kao sto su fuzije i čišćenje, identifikacija korisnika i sesija, identifikacija prikaza stranice. Uspešna primena data mining tehnike u korišćenju Web podataka veoma zavisi od pravilne primene zadatka preprocesuiranja.



Slika 2[11] - Koraci u pripremi podataka za korišćenje Web mininga

Podaci fuzije se odnose na spajanje logova fajla sa nekoliko Web i aplikacionih servera. Ovo može zahtevati globalnu sinhronizaciju preko ovih servera. U odsustvu zajedničkih ugrađenih IDS sesija, heurističkih metoda zasnovanih na „upućivanju“ polja unosa loga servera, zajedno sa različitim sesijama i metodama indetifikacije korisnika može se koristiti za obavljanje objedinjavanja. Ovaj korak je od suštinskog značaja za „inter-site“ korišćenje Web mining-a, gde se analiza ponašanja korisnika vrši preko log fajlova za bliske Web sajtove.

Podaci za čišćenje podrazumevaju zadatke kao što su uklanjanje suvišnih referenci na ugrađene objekte, stil fajlovi, grafike ili audio datoteke kao i uklanjanje referenci zbog spajder navigacije. Sledeći zadatak može da vrši održavanje spiska poznatih spajdera koristeći heuristiku ili korišćenjem klasifikacije algoritama za

[11] izvor - http://www.igi-global.com/viewtitlesample.aspx?id=10785



izgradnju modela spajder i Web robota navigacije. Takođe, nisu svi zahtevi klijent strana beleženi u pristupu log servera. Na strani klijenta ili proxy strani keširanje često može dovesti do nedostatka referenci te stranice ili objekata koji su keširani. Većina referenci koje nedostaju može se heuristički zaključiti kroz proces koji se zove put završetka koji se oslanja na poznavanju strukture sajta i posrednu informaciju sa loga servera. U slučaju dinamički generisane stranice, oblik aplikacije koristeci HTTP post metod rezultira u svim ili u delu parametra korisničkog unosa, nije dodata url adresa pristupa od strane korisnika, a samim tim ne pojavljuje se ni u logovima servera.

U korišćenju Web mining-a neophodno je napraviti razliku između različitih aktivnosti korisnika. U odsustvu mehanizma autentifikacije, čest pristup razlikovanja između jedinstvenih posetilaca jeste upotreba kolačića na strani klijenta. Međutim, ne koriste svi sajtovi kolačiće, a zbog privatnosti, na strani klijenata kolačići su ponekad onemogućeni od strane korisnika. IP adresa sama po sebi generalno je nedovoljna za mapiranje logova ulaza za skup jedinstvenih posetilaca. Ovo je uglavnom zbog korišćenja ISP proxy servera koji daju rotirajuće IP adrese za klijente dok oni pregledaju Web. U takvim slučajevima, moguće je da se preciznije identifikuje jedinstveni korisnik putem kombinacija IP adrese i drugih informacija kao sto su korisnički agenti i referisanja.

Pod predpostavkom da jedinstvene evidencije korisnika mogu biti identifikovane, podrazumevamo da niz prijavljenih aktivnosti pripadaju istom korisniku kao korisniku log aktivnosti. Sesionizacija je proces segmentacije korisničkih aktivnosti logovanja svakog korisnika u sesijama, svaki predstavlja jednu posetu na sajtu. Web sajtovi bez korišćenja dodatnih informacija autentifikacije korisnika i bez mehanizama kao što su ugrađene IDS sesije, se moraju oslanjati na heurističke metode za sesioniranje. Cilj heurističkog sesioniranja je rekonstrukcija, od clickstream podataka, stvarnog niza radnji jednog korisnika u toku jedne posete sajta. Generalno heuristička sesionizacija se deli na dve osnovne kategorije: vremenski orjentisane ili strukturno orjentisane. Vremenski orjentisani primenjuju ili globalno ili lokalno vreme – procena da se napravi razlika između uzastopnih sesija, dok strukturalna orjentisanost koristi statistički strukture sajta ili implicitno povezivanje struktura iz logova servera. Razne heuristike za sesionizaciju su indetifikovane i proučavane.

Posle osnovnog koraka clickstream procesuiranja, podaci iz razlicitih drugih izvora moraju biti integrisani. Integracija sadržaja, strukture i korisničkih podataka u raznim fazama procesa korišćenja Web mining-a mogu biti od suštinskog značaja u pružanju mogućnosti za dalju analizu i razlog otkrivanja obrazca. Na primer, integracija semantičkih znanja iz sadržaja sajta ili semantičke osobine proizvoda mogu se koristiti od strane personalizacije sistema da bi pružio više korisčnikih preporuka. U e-commerce aplikacijama, integracija oba kupca i podataka proizvoda (npr. Demografije, ocene, istorije kupovine) iz operativne baze podataka sa upotrebom podataka mogu dozvoljavati otkrivanje važnih pokazatelja poslovne inteligencije kao što su ocene konverzacije klijenata i vrednosti života. Upotreba strukture podataka je neophodna tokom preprocesuiranja (Na primer, u identifikaciji prikaza stranice, sesonizaciji i u putanji završetka). Ali se takođe može koristiti za poboljšanje rezultata modela na bazi personalizacije tehnike.

Direktni izvor semantičkih znanja koja se mogu integrisati u mining proces je kolekcija sadržaja koja je u vezi sa stavkama ili prikazima stranica na Web sajtu. Ove karakteristike obuhvataju ključne reci, fraze, kategorije imena ili specifične atribute u vezi sa stvarima ili proizvodima kao sto je cena, marka itd. Sadržaj preprocesuiranja uključuje ekstrakciju relevantnih karakteristika iz teksta i meta podataka.

6 Tehnike i alati otkrivanja paterna u Web mining-u

Tipovi i nivoi analize korišćeni na integrisanim podacima zavise od primarnih ciljeva analitičara i željenih izlaza. Ova sekcija opisuje najčešće tipove otkrivanja paterna i analize koji se koriste kod Web mining-a i koji su predstavljeni na slici 11. Osnovni tipovi analize koji se javljaju kod Web usage mining-a su[12]:

- Analiza sesija i korisnika (Session and Visitor Analysis)
- Asocijativna i korelativna analiza (Association and Corelation Analysis)
- Klasterovanje i segmentacija korisnika (Cluster Analysis and Visitor Segmentation)
- Analiza sekvencijalnih i navigacionih paterna (Analysis of Sequential and Navigational Patterns)
- Klasifikacija i modeliranje korisnika (Web user modeling classification)



Slika 3 - Tehnike Web usage mining-a i zadaci koje rešavaju

[12] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: *Web usage mining: Discovery and applications of usage patterns from web data*, Department of Computer Science and Engineering, University of Minnesota 1999



6.1 Analiza sesija i korisnika

Statistička analiza preprocesiranih sesija konstituiše uobičajenu formu analize. U ovom slučaju, podaci su agregirani po predefinisanim jedinicama, kao što su dani, sesije, posetioci, ili domeni. Standardne statističke tehnike mogu biti korišćene na ovim podacima da se dobije znanje o ponašanju korisnika. Ovo je pristup preuzet od najkomercijalnijih alata koji su dostupni za analizu Web logova. Izveštaji bazirani na ovom tipu analize mogu uključivati informacije o najčešće pristupanim stranama, prosečnom vremenu zadržavanja na stranici, prosečnoj putanji kroz sajt, najčešćim ulaznim i izlaznim tačkama, i o ostalim agregatnim merama. Uprkos nedostatku dubine u ovakvoj vrsti analize, rezultujuće znanje potencijalno može bti korisno za poboljšanje performansi sistema i radi pružanja podrške marketinškom odlučivanju. Nadalje, komercijalni Web alati analitike u velikoj meri inkorporiraju razne data mining algoritme i time rezultuju sofisticiranijim sajtovima i korisničkom metrikom.

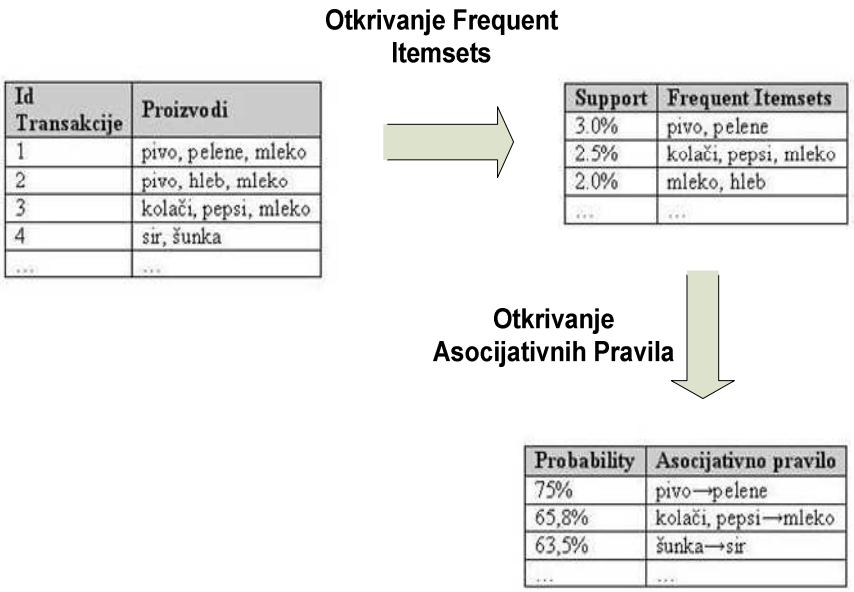
Druga forma analize integrisanih podataka je OLAP. OLAP pruža bolje integrisan okvir za analizu sa većim nivom fleksibilnosti. Izvori podataka za OLAP analize su najčešće multidimenzionalni Data Warehouse ili u slučaju Web mining-a modifikovani Data Warehouse (Data Webhouse), koji integrišu korišćenje, sadržaj i podatke e-commerc sajtova na različitim nivoima agregacije za svaku dimenziju. OLAP alati dozvoljavaju promene u agregacionim nivoima po svakoj dimenziji tokom analize. Dimenzije analize u takvoj strukturi mogu biti bazirane na različitim poljima koja su dostupna u log fajlovima i mogu uključivati vremensko trajanje, domen, zahtevani resurs, browser korisnka, reference, itd. Ovo omogućava da analiza bude izvršena na delovima loga koji je povezan sa određenim vremenskim intervalom ili na višem nivou abstrakcije sa odgovarajućom strukturom URL putanje. Izlazi OLAP upita mogu takođe biti korišćeni kao ulaz za razne data mining tehnike ili alate za vizualizaciju.

6.2 Asocijativna i korelativna analiza

Otkriće asocijativnih pravila i statistička korelativna analiza na podacima o korišćenju Web-a rezultuje otkrivanjem grupa proizvoda koji se kupuju zajedno ili stranica kojima se najčešće pristupa. Ovo, omogućava kompanijama koje vrše online prodaju da organizuju sadržaj sajta efikasnije ili da predlože neke efektnije kombinovane rasprodaje (prodaje).

Algoritam za otkrivanje asocijativnih pravila pronalazi grupe proizvoda koje se često pojavljuju zajedno u mnogim transakcijama (uvek se postavlja neki prag koji mora da se pređe pri zaključivanju). Postoje dva osnovna koraka algoritma za otkrivanje asocijativnih pravila. Prvi korak je računski intenzivniji i osnovni cilj je da se otkriju frequent itemsets. Drugi korak algoritma ima za cilj generisanje

asocijativnih pravila na osnovu frequent itemsets iz prethodnog koraka. Ilustracija algoritma je prikazana na slici 4.



Slika 4 - Prikaz dvo-koračnog algoritma za generisanje asocijativnih pravila

Radi boljeg pojašnjenja principa algoritma prvo će biti objašnjeni određeni pojmovi koji se koriste i koji su od izuzetnog značaja za sam algoritam.

Itemset – predstavlja kolekciju predmeta (stavki). Svaka stavka je vrednost atributa. Na primer u kupovnoj korpi, itemset sadrži kolekciju proizvoda kao što su kolači, Pepsi i mleko. Kod istraživanja demografskih karakteristika kupca, itemset sadrži kolekciju atributa kao što su {Pol = 'Muški', Obrazovanje= 'Bachelor'}. Svaki itemset ima takođe i veličinu, koja predstavlja broj stavki koje sadrži. Na primer veličina itemset-a {Kolači, Pepsi, Mleko} je 3.

Frequent itemsets - su oni itemset koji su relativno zastupljeni u dataset-u. Prag zastupljenosti za itemset se definiše korišćenjem parametra support koji je dalje objašnjen.

Support – se koristi da meri zastupljenost nekog itemset-a. Support nekog itemseta {A, B} je dobijen preko ukupnog broja transakcija koje sadrže i A i B.

Support ({A, B}) = Broj transakcija (A, B)

Minimum Support - je prag parametar koji treba odrediti pre izvršavanja asocijativnog modeliranja. On nam govori da smo zainteresovani samo za one itemsets i pravila koja zadovoljavaju bar minimalan support u postojecem datasetu. Parametar Minimum_Support se koristi da ograniči itemsets, ali ne i pravila.



Probability (Confidence) – je svojstvo asocijativnog pravila. Probability pravila A→B je izračunato korišćenjem supporta itemseta {A, B}podeljenog sa support-om {A}. Probability se takođe naziva confidence u data mining istraživačkom društvu. Definiše se na sledeći način:

Probability (A => B) = Probability (B|A) = Support (A, B)/ Support (A)

Minimum_Probability je prag parametar koji se mora odrediti pre pokretanja algoritma. Označava da je korisnik zainteresovan samo za ona pravila koja imaju visoku probability pre nego minimalnu probability. Minimum_Probability nema uticaja na itemsets, ali ima uticaja na pravila.

Importance - koja se u literaturi naziva interesting score ili lift se koristi za procenu itemsets i pravila. Importance nekog itemseta je definisana korišćenjem sledeće formule:

Importance ({A,B}) = Probability (A, B)/ (Probability (A)* Probability (B))

Ako je importance = 1, A i B su nezavisni items, što u prevodu znači da kupovina proizvoda A i kupovina proizvoda B su dva nezavisna događaja. Ako je importance < 1, A i B su negativno korelisani (zavisni). To znači ako kupac kupuje proizvod A, vrlo retko u istoj transakciji kupuje i proizvod B. Za pravila, importance se izračunava primenom sledeće formule

Importance (A => B) = log (p (B|A)/p (B|not A))

Kada je vrednost importance 0 to znači da nema povezanosti između A i B. Pozitivna vrednost importance parametra znači da probability B raste kada je A istinito. Negativna importance znači da se probability B smanjuje kada je A istinito.

Sa slike 4. se jasno vidi da se pre prvog koraka algoritma identifikuju transakcije koje se čuvaju u operativnoj ili nekoj drugoj bazi nad kojom se vrši analiza. Postavljanjem parametra minimum_support na određenu vrednost generišu se skupovi frequent itemsets sa odgovarajućim vrednostima support parametra što predstavlja prvi korak algoritma generisanja asocijativnih pravila. Na primer iz druge tabele se vidi da 3% od svih transakcija sadrže pivo i pelene. Drugi korak algoritma je otkrivanje asocijativnih pravila. Broj kvalifikovanih asocijativnih pravila je baziran na parametru Minimum_Probability. Na primer ukoliko se postavi vrednost parametra Minimum_Probability na 30%, to znači da 30% kupaca koji kupi proizvod A takođe kupuje i proizvod B, kod pravila A→B i to je kvalifikovano pravilo. Sa slike iz tabele asocijativnih pravila imamo pravilo pivo→pelene sa vrednošću parametra probability od 75%, što znači da 75% onih kupaca koji kupe pivo kupe i pelene. .

Asocijativna pravila se izražavaju na sledeći način. Asocijativno pravilo r je izraz forme X → Y (σr, αr), gde su X i Y itemsets, σ je support itemsets X * Y koja reprezentuje verovatnoću da se X i Y pojave zajedno u transakciji, i α je confidence za pravilo r, reprezentujući uslovnu verovatnoću da se Y pojavljuje u transakciji, pod pretpostavkom da se X pojavljuje u toj transakciji. Iz prethodnog



objašnjenja asocijativno pravilo između piva i pelena se može iskaziti kao pivo→pelene[3%,75%] što znači da 75% kupaca koji kupuju pivo ujedno kupuju i pelene, a to čine 3% svih kupaca.

Asocijativna pravila mogu se takođe iskoristiti za optimizaciju strukture sajta. Na primer ukoliko sajt ne pruža direktne linkove između dve strane A i B, otkrivanje pravila A→B može da indikuje da uspostavljanje direktnog hiperlinka od A do B može pomoći korisnicima da brže i lakše nađu potrebne informacije. Ako sesiju posmatramo kao transakciju, algoritmi mininig-a za generisanje asocijativnih pravila se mogu koristiti da otkriju asocijativne relacije između stranica koje se zahtevaju. Na primer asocijativno pravilo strana A → strana B [5% 80%] govori da 80% korisnika koji pogledaju stranicu A će takođe pogledati i stranicu B., a 5% svih korisnika pogleda obe stranice.

Obe asocijativne analize (i sa proizvodima i sa stranicama) i statistička korelaciona analiza (uglavnom među mušterijama ili posetiocima) su korišćene uspešno u Web personalizaciji i sistemima preporuke.

6.3 Klasterovanje i segmentacija korisnika

Klasterovanje je data mining tehnika koja grupiše kolekciju objekata koji imaju slične karakteristike. U domenu korišćenja Web-a, postoje dve vrste interesantnih klastera koji mogu biti istraženi (otkriveni):

- Korisnički klaster
- Klaster stranica.

Klasterovanje korisničkih zapisa (sesija i transakcija) je jedan od najčešćih zadataka analize u Web mining-u i Web analitici. Klasterovanje korisnika teži da oformi grupe korisnika koje ispoljavaju slične paterne "surfovanja". Takvo znanje je specijalno korisno za izvođenje demografije korisnika u cilju ostvarivanja segmentacije tržišta u aplikacijama e-commerc ili pri pružanju personalizovanog Web sadržaja korisnicima sa sličnim interesima. Dalje analize korisničkih grupa, bazirane na demografskim atributima (kao što su godine, pol, primanja, itd...) mogu voditi otkrivanju važnih pravila poslovne inteligencije. Klasterovanje zasnovano na ponašanju se takođe može koristiti za kreiranje Web udruženja korisnika (Web-based user communities), koja reflektuju slične interese grupe korisnika, i za otkrivanje modela (profila) korisnika koji se mogu koristiti kao dinamičke preporuke pri Web personalizaciji aplikacija.

Klasterovanje stranica (objekata) se može izvoditi, zasnovano na podacima o ponašanju (počinjući od korisničkih sesija ili transakcija) ili osobina sadržaja povezanih sa stranicama ili objektima (ključne reči ili atributi). U slučaju klasterovanja zasnovanom na sadržaju, rezultat mogu biti kolekcije stranica ili proizvoda povezanih sa samom temom ili kategorijom. U klasterovanju



zasnovanom na ponašanju, objekti kojima se najčešće pristupa ili koji se najčešće kupuju zajedno može biti automatski organizovano u grupe. Takođe, može biti iskorišćeno za prikaz statičkih ili dinamičkih HTML stranica koje predlažu određene hiperlinkove korisniku u zavisnosti od njegove istorije kretanja kroz sajt ili istorije kupovne aktivnosti.

6.4 Analiza sekvencijalnih i navigacionih paterna

Tehnika otkrivanja sekvencijalnih paterna pokušava da pronađe paterne između sesija koji govore na primer da prisustvo seta objekata (proizvoda) je praćeno nekim proizvodom u određenom intervalu sesija ili epizoda. Koristeći ovaj pristup, Web marketing eksperti mogu da predvide buduće paterne koji će biti od pomoći u kreiranju reklamne kampanje za određenu grupu korisnika. Drugi tipovi privremene analize koji se mogu izvoditi kao sekvencijalni paterni uključuju analizu trenda i analizu sličnosti. U kontekstu podataka o ponašanju na Web-u, otkrivanje sekvencijalnih paterna se može koristiti za sakupljanje čestih navigacionih putanja među korisničkim tragovima.

Shvatanje Web transakcije kao sekvence pageview-ova omogućava upotrebu brojnih, korisnih i dobro istraženih modela u otkrivanju i analizi navigacionih paterna korisnika. Jedan takav pristup je modelovanje navigacione aktivnosti na Web sajtu kao Markov model; svaka stranica (ili kategorija) može biti reprezentovana kao stanje, a tranzicione verovatnoće između ovih stanja mogu prezentovati verovatnoću da će korisnik preći iz jednog stanja u drugo (najjednostavnije rečeno: preći sa jedne stranice na drugu!). Ova reprezentacija dozvoljava izračunavanje brojnih korisnih korisničkih metrika ili metrika sajta. Na primer jedna može da izračunava verovatnoću da će korisnik napraviti kupovinu (porudžbinu), na osnovu toga što je korisnik izvršio pretragu online kataloga. Markov model je prikazan kao fundamentalna mašinerija za predviđanje linkova dovoljno dobra kao i za predviđanje Web-a u minimiziranju nesigurnosti sistema. Cilj ovakvog pristupa je predviđanje sledeće korisnikove akcije zasnovane na prethodnom "surfing" ponašanju korisnika. Takođe se koriste da otkriju verovatnoće korisničkih navigacionih putanja na Web sajtu. [13]

Sofisticiranija statistička tehnika kao što je mešavina Markovog modela, je takođe bila korišćena za klasterovanje navigacionih sekvenci i za izvođenje eksploratorne analize navigacionog ponašanja korisnika na Web sajtu.

Drugi način efikasnog reprezentovanja navigacionih putanja je ubacivanje svake putanje u strukturu stabla (tree structure). Dobro poznat primer ovakvog pristupa je ideja agregatnog stabla predstavljenog kao deo WUM (Web Utilization Miner-a) sistema. Svaki čvor u stablu predstavlja navigacionu podsekvencu u jednoj sesiji. Ovaj pristup i njegovi dodaci imaju potvrđeno korišćenje u evaluaciji navigacionog dizajna sajta.

[13] R. Cooley, B. Mobasher, J. Srivastava: *Web Mining: Information and Pattern Discovery on the World Wide Web*, Department of computer science and engineering University of Minnesota, Minneapolis, MN 55455, USA 1997

6.5 Klasifikacija i modeliranje korisnika

Klasifikacija je zadatak mapiranja (svrstavanja) podataka (objekata) u jednu od nekoliko predefinisanih klasa. Što se tiče Web domena, zainteresovani smo za kreiranje ili razvijanje profila korisnika koji pripada određenoj klasi odnosno kategoriji. Ovo zahteva ekstrakciju i selekciju osobina koje najbolje opisuju karakteristike određene klase odnosno kategorije. Klasifikacija se može izvršiti primenom algoritama učenja koji se nadgledaju, kao što su tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, neural networks, i suport vector machines. Takođe je moguće koristiti prethodno otkrivene klastere i asocijativna pravila za klasifikaciju novih korisnika.

Klasifikacione tehnike igraju izuzetno važnu ulogu u analitičkim Web aplikacijama za modelovanje korisnika prema raznim predefinisanim metrikama. Na primer za datu kolekciju korisničkih transakcija, suma kupovina za svakog korisnika u određenom periodu vremena može biti izračunata. Klasifikacioni model tada može biti izgrađen na osnovu ovih obogaćenih podataka u cilju klasifikacije korisnika u one koji imaju veliku sklonost kupovini i na one koji nemaju sklonost kupovini, uzimajući u obzir osobine kao što su korisnički demografski atributi kao i njihove navigacione karakteristike.

Još jedna bitna tehnika klasifikacije i modeliranja korisnika u Web domenu jeste Web personalizacija i sistemi za preporučivanje. Na primer aplikacije za filtriranje koje sarađuju u postojećim sistemima za preporuku koriste k-nearest neighbor classifiers za predviđanje rangiranja korisnika ili kupovnu sklonost mereći korelacije između posmatranog korisnika i prethodnih korisničkih transakcija. Mnogi od pristupa Web mining-a o kojima je bilo reči mogu biti korišćeni za automatsko otkrivanje modela i primenu ovih modela radi obezbeđivanja personalizovanog sadržaja aktivnim korisnicima.

Paterni ponašanja otkriveni pomoću Web mining-a su efektivni u otkrivanju (pronalaženju) objekat-objekat (proizvod-proizvod) i korisnik-korisnik relacija i sličnosti na nivou korisničkih sesija. Ipak, bez doprinosa nekoj dubljoj analizi, takvi paterni bi malo doprinosili fundamentalnim razlozima zbog čega su ti objekti ili korisnici grupisani zajedno. Nadalje, nasledna i povećana heterogenost Web-a zahteva da Web aplikacije efektivnije integrišu različite vrste podataka kroz višestruke kanale i iz različitih izvora.

Ipak, fokus na tehnike i arhitekturu za što efektivniju integraciju i mining sadržaja, ponašanja, i strukture podataka iz različitih izvora izgleda vodi ka sledećoj generaciji više korisnih i inteligentnijih aplikacija i ka sofisticiranijim alatkama za Web mining koje će izvući inteligenciju iz korisničkih transakcija na Web-u.



Moguće je shvatiti neke od semantika sajtova integrišući pristup filtriranja sadržaja baziranog na ključnim rečima sa mining tehnikama. Ipak, u cilju da se shvate i otkriju kompleksnije relacije na višem semantičkom nivou zasnovanom na atributima povezanim sa strukturiranim objektima, bilo bi neophodno zakoračiti ispred reprezentacija zasnovanih na ključnim rečima i automatski integrisati relacione strukture sa ontologijom domena u preprocesiranju i mining procesu. Napori u ovom pravcu imaju izgleda da budu najplodonosniji u kreiranju efektnijeg Web mining-a, modelovanja korisnika i personalizaciji sistema koji su konzistentni sa pojavljivanjem i širenjem semantičkog Web-a.

7 Analiza i alati za analizu otkrivenih paterna

Paterni koji su otkriveni na Web-u pomoću ranije opisanih tehnika, neće biti od velike koristi analitičarima ukoliko ne postoje mehanizmi i alati koji će im pomoći u njihovom boljem razumevanju. Pored razvoja tehnika za otkrivanje korisnih paterna unutar Web logova, postoji i potreba za razvojem tehnika i alata koji omogućavaju analizu otkrivenih paterna. Od ovih tehnika se očekuje statističko predstavljanje brojnih podataka, korišćenje grafike, vizualizacija, primena analize ponašanja, kao i korišćenje upita nad bazama podataka. Analiza ponašanja korisnika prilikom pristupa Web-u je nova oblast istraživanja.[14]

Vizualizacija se veoma uspešno koristi u cilju pomaganja korisnicima da razumeju različite vrste fenomena kako prirodnih, tako i apstraktnih. Dakle, to je prirodan način za razumevanje ponašanja korisnika Web-a. Pitkow je razvio WebViz sistem za vizualizaciju paterna u pristupu WWW-u. Predložena je i paradigma Web putanja ("Web path paradigm") u kojoj je za ekstrakciju podsekvenci u paternima kretanja po Web-u korišćen skup logova za pristup serveru. WebViz sistem omogućava analitičaru selektivnu analizu dela web-a za koji je zainteresovan, bez obrade delova koji nisu od interesa za analizu. Web je predstavljen kao usmeren graf, pri čemu čvorovi predstavljaju Web strane, a grane hiperlinkove između strana.

OLAP (Online Analytical Processing) se pojavio kao snažna paradigma za strategijske analize baza podataka u poslovnim okruženjima. Nedavno je pokazano da funkcije i zahtevi koji se stavljaju pred OLAP procesiranje zahtevaju dizajniranje nove informacione strukture. To je prednost razvoja "data cube" (kocke podataka) informacionog modela i tehnika za njegovu efikasniju implementaciju. Međutim, pokazalo se da analize koje se zahtevaju u postupku Web mining-a imaju mnoštvo zajedničkih karakteristika sa skladištima podataka, pa su i OLAP tehnike sasvim primenljive. Informacije o korisničkim pristupima u logovima servera su modelirane kao jedna "append-only" kolekcija, koja raste tokom vremena. Pošto veličina logova servera veoma brzo raste, nemoguće je obezbediti njihovu online analizu. Stoga, postoji potreba za sumiranjem podataka unutar logova, na različite načine, kako bi njihova online analiza postala izvodljiva.

[14] Cooley, R., Mobasher, B., Srivastava, J.: *Data preparation for mining world wide web browsing patterns*, Journal of Knowledge and Information Systems, 1999



Kako veličina servera raste prilično brzo, možda neće biti moguće omogućiti online analize svih podataka. Kako bilo, postoji potreba da se sumiraju podaci loga, veoma često na različite načine, kako bi se omogućile online analize. Može se zahtevati za potrebe zaštite da se delovi loga učine (ne)vidljivim za različite analize. Jedan od razloga koji su doprineli velikom uspehu relacionih baza podataka je i postojanje deklarativnog jezika upita koji omogućava aplikaciji da izrazi koji uslovi moraju biti zadovoljeni pre nego što se navede kako se dobijaju zahtevani podaci. Ovo se može omogućiti na najmanje dva načina, Prvo, konstante se mogu smestiti u baze podataka koje treba pretraživati. Drugo, upiti se mogu izvesti nad znanjem koje je izdvojeno iz procesa mining-a, i u tom slučaju upiti su potrebniji od podataka.

8 Yahoo! Personalizacija

Yahoo je bila jedna od prvih lokacija na Webu za korišćenje personalizacije velikih razmera. U ovom case study preneću iskustva dizajnera koji su projektovali funkciju personalizacije, dati uvid u probleme Web personalizacije i predložiti ideje za budućnost.

U ovom delu koncetracija je na tri primera personalizacije: My Yahoo!, Yahoo! Companion i Inside Yahoo! Search.

My Yahoo! (my.yahoo.com) je prilagođena lična kopija Yahoo! Korisnici mogu izabrati iz stotine modula kao što su vesti, berza, vreme i sportski rezultati i staviti ih na jednu ili više Web strana. Sadržaj svakog modula se tada ažuira automatski, tako da korisnici mogu da vide ono što žele onako kako oni žele. To pruža korisnicima najnovije informacije o svakoj temi ali samo sa određenim stvarima o kojima oni žele da budu informisani. Primer My Yahoo stranice prikazana je na sledecoj slici. Mi ćemo ovde pomenuti nekoliko opštih tema o toj stranici:

- Personalizacija se često dešava unutar modula. Na primer korisnici mogu da izaberu koje TV kanale žele da uključe u svoj TV vodič u odnosu na to koji kablovski sistem koriste. Ostali moduli su opširniji, na primer zdravstvene vesti.

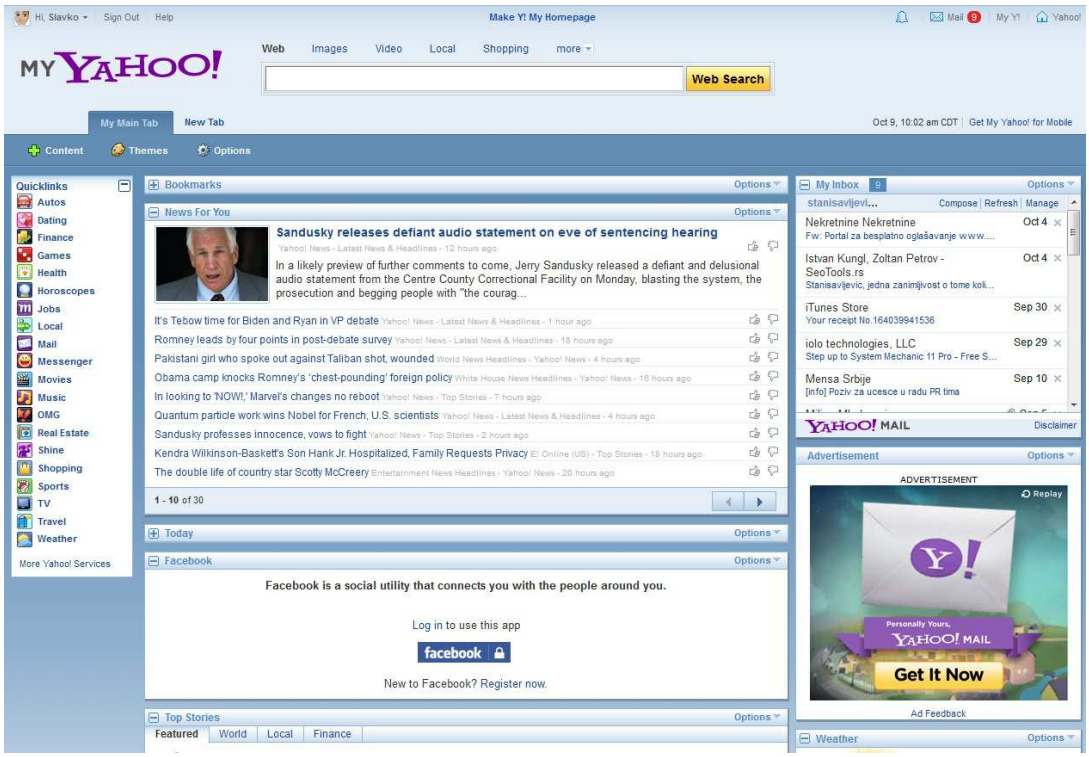
- Ne samo da je sadržaj prilagođen već se i layout (raspored) moze prilagoditi takođe.

- Neki sadržaji se automatski personalizuju. Primer takvog sadržaja je sportski modul koji navodi timove iz korisničkog podneblja nakon dobijanja te informacije od korisnika.

- My Yahoo opcija omogućava My Yahoo stranici da automatski ažurira u bilo kom korisnički određenom intervalu od 15 minuta do nekoliko sati. Stranica se uvek gradi na on-the-fly podudaranju korisničkih preference sa

dostupnog sadržaja. Arhitektura je dovoljno efikasna da bude u stanju da obezbedi servis za milione ljudi sa hiljadu izvora pri hiljadu promena dnevno, koristeći relativno mali broj off-the-shelf kompjutera. Arhitektura je potpuno skalabilna. Kako korisnička baza raste jednostavno se dodaje više hardwara eleminišući potrebu za skupim hardware-skim rešenjima.

- Moduli mogu biti izabrani (sa duge) liste, ali takođe mogu biti dodati klikom na dugme sa originalne strane sadržaja. Na primer svaka stranica o vremenu (weather.yahoo.com) sadrži “add to my yahoo!” dugme, koje dodaje stranicu direktno u korisnikov My Yahoo nalog. Takođe svaki modul na My Yahoo stranici ima edit i remove dugme, omogućavajući korisnicima da direktno manipulisu svojom stranicom, bez ikakve potrebe da posete edit-layout stranu.



Slika 5 - Primer My Yahoo! strane

Yahoo! Companion (companion.yahoo.com) je browser toolbar sa koje korisnik može direktno pristupiti većini Yahoo funckija sa bilo koje stranice na Webu. U izvesnom smislu, to je kao Mini My Yahoo! koji zauzima mali prostor na vrhu stranice i uvek je sa Vama. Izgled toolbara se može promeniti u bilo kom trenutku, a promene ostaju sa korisnicima čak iako se prebace na drugi kompjuter.

Održavanje bookmarka je dobar primer upotrebe Yahoo companiona. Za korisnika, interfejs je sličan bilo kojoj drugoj funkciji bookmarka, ali razlika je u tome što se ovi bookmarkovi čuvaju na serveru, tj. oni su dostupni i dosledni bez obzira na računar koji se koristi. Drugi primer je mogućnost da se izabere nekoliko alatki i da se zamene u bilo kom trenutku.



Inside Yahoo! Search results. Desetine miliona različitih upita se šalje na yahoo search svakog dana. Nemoguće je, naravno, prilagoditi svaku od njih međutim, nekoliko hiljada fraza su dovoljno jasne i Yahoo povezuje sadržaj dovoljno dobro da možemo dopuniti uobičajna Web traženja sa direktnim, fokusiranim sadržajem koji ponekad može biti personalizovan. Na primer, ako neko traži ime filma, ukazaće mu se yahoo! Movies, prikazaće sliku filma, glumca kao i vreme aktuelne projekcije. Ukoliko korisnik pogleda vreme projekcije i unese poštanski broj, ta stranica će se automatski prilagoditi bioskopima u njegovoj okolini. Jednim klikom nakon traženja filma, može se videti projekcija u nečijem okruženju. Na sličan nacin, ako je potraga za kineskom hranom biće upućen na yahoo yellom pages i biće prikazana lista kineskih restorana u blizini. Naravno lokacija se može promeniti u svakom trenutku.

8.1 Oblasti za personalizaciju širih razmera

Potencijal skalabilnosti mora biti ugrađen u svaku Web personalizaciju proizvoda direktno od početka. Ljudi očekuju da njihovi kompjuteri komuniciraju sa njima brzo, odugovlačenje udaljava korisnika. Yahoo je uvek bio opsednut brzinom i efikasnošću u velikoj meri jer imaju jako veliku bazu korisnika ali takođe jer veruju da je jednostavnost i praktičnost najvažnija u svakoj personalizovanoj alatki korisnika. Lični podaci o Yahoo korisnicima se održava u specijalno dizajniranoj user database. Zbog izuzetno visoke stope transakcija Yahoo se nije odlučio za korišćenje komercijalno dostupnih baza podataka već je izgradio sopstveni prilagođeni software. Dodate su mnogobrojne osobine ovom jezgru UDB kako je njihova baza korisnika rasla, uključujući optimizaciju, keširanje, puna suvišna mehanizacija komunikacije između UDB i My Yahoo! strane. Takođe su dodali replikaciju podataka i sposobnost distribucije, omogućavajući replikaciju i distribuciju UDB preko sigurnih veza u udaljenim mestima u Aziji i Evropi. Ova ista tehnologija je primenjena lokalno kako bi se omogućio backup u svakom trenutku. Baza korisnika Yahoo-a je toliko masivna, i menja se tako velikom brzinom, da postojeći sekundarni backup mehanizmi nisu mogući. Sa naglaskom na pouzdanost, Yahoo je bio primoran da razvije nove tehnologije samo za tu svrhu, i ovo se lepo isplatilo na duge staze.

8.2 Privatnost I bezbednost

Pitanje privatnosti je suviše veliko da bi se diskutovalo ovde. Ali ipak nijedna diskusija o personalizaciji se ne može tretirati bez nje. Svaka kompanija koja prikuplja lične informacije mora čuvati tu informaciju sa svojim (poslovnim) životom. To je jako važno. Neograničeno deljenje ovih informacija sa drugim kompanijama ili čak drugim nezavisnim odeljenjima unutar iste kompanije mogu

imati katastrofalne rezultate. Treba biti čuvan onoliko koliko i tajne trgovine. Lozinke korisnika se uvek čuvaju u šifrovanom obliku, šifriraju se svi osetljivi podaci, čuvaju se na mašinama sa ograničenim pristupom itd. Takođe se obavljaju bezbednosno revizorske procene i predlažu se potrebne promene kao i zapošljavanje nekoliko internih ljudi posvećenih isključivo pitanju bezbednosti I privatnosti.

Uvek će postojati tenzija između upotrebe ličnih podataka radi poboljšanja usluga korisnika i upotrebe istih podataka radi profita kompanije. Neophodno je imati ljude sa punim radnim vremenom koji sluze kao šampioni potrošača, koji su potpognuti posmatračima I revizorima.

8.3 Korisnički interfejs

Upotrebljivost je idalje najveće tehničko pitanje za personalizaciju velikih razmera. U radu će se razmatrati samo jedan primer: Pitanje predvidljivosti. Bilo je mnogo govora o funkciji personalizacije koja uči šta korisnici žele i pokušava da izgovori. Prilagođjenje novine koje ističu samo vesti koje korisnik želi da čuje je dobar primer. Glavna slabost ovakvog sistema je njihova nepredvidivost. Većina korisnika očekuje da ima makar intiutivno shvatanje onoga što im je dato i oni očekuju da vide isto dosledno ponašanje. Biti iznenađen je divno ukoliko je potpuno pozitivno iznenađenje, ali sve ukupno, biti nepredvidiv je negativno. Konkretno, ako ljudi nisu sigurni kako nešto radi, manje su skloni eksperimentisanju, jer se plaše da će pokvariti nešto ili doći u stanje u kome neće moći opozvati operaciju. Svaka funkcija personalizacije treba da podstiče eksperimetnisanje.

U slučaju vesti, na primer nije jasno da li ljude žele lične vesti jer oni često žele iste vesti koje svako drugi dobija. Ovo ne znači da lične vesti nemaju vrednost. Vesti o nekoj kompaniji ili nekom gradu, školu, rodbini može biti izuzetno vredna, ali ovo su direktni primeri nasuprot apstraktnim primerima. Recimo, moćne crne kutije su generalno opasne ukoliko su im rezultati intuitivni, dosledni i predvidivi. Dobijanje vremenske prognoze i vesti o lokalnom sportskom timu su očigledni. Dobijanje vesti o raku iz nekih medicinskih casopisa iz prošlosti ili traženje nekih medicinskih termina moze da zbuni korisnika u najboljem slučaju, a u najgorem slučaju moze da ugrozi poverenje korisnika i stvori ozbiljnu zabrinutost u privatnost u umu korisnika.

8.4 Neka opšta zapažanja i naučne lekcije

Ovde su navedena nekoliko zapažanja i saznanja o Web personalizaciji velikih razmera. Koncetracija je na manje očigledna pitanja više nego na sveobuhvatna.



Većina korisnika uzme ono što im je dato i nikada ga ne prilagodi svojim potrebama. Jedna od atrakcija Weba je jednostavnost u kojoj se informacije mogu dobiti. U nekim slučajevima, ovo je šteta, jer učimo ljude da ulazu premalo napora, imaju suviše kratak raspon pažnje i idu najlakšim rutama. Veoma iznenađujuća statistika je da većina aktivnih My Yahoo korisnika ne prilagode svoju stranu. Oni rade sa podrazumevanim (default) vrednostima. Moglo bi biti tri razloga za to:

1. Standardna strana (default) je tako dobra da nema razloga da se uloži više napora.
2. Alatke za prilagođavanje je toliko teško koristiti da ljudi ne žele time da se zamaraju.
3. Mnogim ljudima ne treba složena personalizacija.

Odgovor je verovatno kombinacija sva tri odgovora a isto verovatno važi i za većinu drugih personalizacija, ne samo za yahoo.

Veliki napor se treba uložiti u podrazumevanu stranicu(default page). Očigledan cilj je poboljšanje alata za prilagođavanje. Ali poboljšanje standardne strane za ljude koji je ne prilagođavaju obično dobija manje pažnje. Ovo neće biti uvek presudno, Yahoo ulaže značajan napor čineći standardnu stranu boljom koliko je moguće. Najbolji primer je upotreba zip kodova (ili drugim kodovima koji pružaju lokaciju). Znajući poštanski broj može se automatski izabrati strana koja prikazuje vremensku prognozu, sportske ekipe, lokalne vesti, lokalne događaje, lokalne gužve u saobraćaju itd.

Napredni korisnici mogu učiniti neverovatne stvari. Suprotno od korisnika koji ne prilagođavaju svoje strane su iskusni korisnici. Česta greška je praviti predpostavke da niko nikada neće želeti da uradi to. My Yahoo! prilagođena strana je veća od 500KB, može da se skladišti portfolio veći od 200 prilagođavanja. Pre tri godine bi ovo izgledalo besmisleno. Povećavajući napore u pojednostavljivanje interfejsa za prilagođavanje, trudili su se da sačuvaju sve sposobnosti My Yahoo! za napredne korisnike. Glavna lekcija ovde (kao i za većinu oblasti Yahoo-a) je da se dizajnira sve za beskonačan rast koliko god je moguće. Nekoristiti veštačke granice osim ako je apsolutno neophodno.

Prilagođavanje bi trebalo da Vas prati koliko god je to moguće. Ukoliko se prijavite za My Yahoo! i kreirate portfolijo deonica, i kasnije idete na područje tržista akcija, taj portfolio je idalje sa Vama. Ukoliko rasporedite vaše ikone ili bookmarkove na izvestan način i prebacite se na drugi računar ili odete negde iz grada, on ostaje uvek sa vama. Čuvanje informacija o korisnicima u centralnoj bazi podataka pre nego na računaru korisnika pomaže osobi da vidi iste informacije od kuće i sa posla (naravno podrazumeva se da je baza podataka poverljiva i sigurna).

Ljudi obično ne shvataju koncept prilagođavanja. Ono što zvuči očigledno istraživacima nije očigledno i za većinu ljudi. Ljudi nisu navikli na računare koji stvaraju iznenađujuće, naizgled inteligentne rezultate. Oni su navikli na statične stvari. Većina ljudi koja vidi My Yahoo! stranicu prvi put misli da je to samo još jedan način da se predstavi sadržaj i propusti činjenicu da korisnik može da ga



prilagodi. Zato je neophodno da se predstavi svaka alatka za prilagođavanje na najjednostavniji način korisniku.

Uvideli smo kako znajući poštanski broj može omogućiti mnoge automatske funkcije prilagođavanja.

Ali ukoliko je korisnički interfejs tako da poštanski broj mora biti unešen, onda će svi ne američki korisnici biti odbijeni. Neki Web sajtovi dodaju provere određenih pravila oko adresa ili broja telefona koja važe samo u SADu, ili predpostavljaju da neko koristi određeni browser na izvesnom operativnom sistemu i neproverava da li njihov sajt može biti koristan i za druge koji koriste različite platforme.

8.5 Učenje od korisnika

Bez obzira na to koliko dobro je alat dizajniran za krajnje korisnike, oni će ga koristiti na neočekivan način. Ovo posebno važi za potpuno nove aplikacije kao i za većinu alata za personalizaciju. Srećom, Web aplikacije su lakše za učenje. Yahoo kontstntno gleda logove i gradi specijalne alate kako bi video ne samo brojeve već i način korišćenja promene i neobične događaje.

8.6 Zaključak

Povezivanje ljudi i kompjutera u privatnom smislu je veoma teško. Previše pokušaja je učinjeno bez pravog osvrtaja na to šta ljudi zaista žele, šta mogu da koriste i kako najbolje da odgovara njihovim potrebama. Količina i najvažnije dubina dostupnih ličnih podataka je neverovatna. To će biti veliki izazov za sve učesnike koji kombinuju poslovanje, tehnologiju i društvena pitanja da rade na način koji koriste potrošačima a da pritom ne ugrožava njihovu privatnost.

Personalizovane karakteristike su trenutno velika korist naprednim korisnicima – onima koji imaju dovoljno samopouzdanja da eksperimentišu sa ovim opcijama i koriste vreme da stvore nešto što zaista odražava njihova lična interesovanja. Glavni izazov personalizacije velikih razmera je da se smanji unos podataka, time stvarajući jednostavnije za manje iskusne korisnike da prilagode njihove stranice i napraviti je jasnijom za početnike kako bi lakše uvideli opciju prilagođavanja i personalizacije. Učenje direktno od korisnika ima veliki potencijal ali i velike barijere. Skalabilnost je od suštinskog značaja. Biti u mogućnosti da brzo služi milione korisnika, pouzdano i jeftino je veliki deo uspeha.

9 Zaključak

U ovom radu je predstavljena sveobuhvatna rasprava procesa Web personalizacije posmatrana kao primenom Web mining-a koji mora biti podržan tokom različitih faza tipičnog data mining ciklusa. Razgovarano o mnostvu aktivnosti i tehnika koje se koriste u razlicitim fazama ovog ciklusa, uključujući preprocesuiranje i integraciju podataka iz više izvora, i tehnike otkrivanja obrazaca koje se primenjuju na ove podatke. Takođe su predstavljene brojne specifične preporuke algoritama za kombinovanje otkrivenog znanja trenutnih statusa korisničkih aktivnosti na Web sajtu kako bi obezbedio lični sadržaj korisnika.

Ključni deo procesa personalizacije je generisanje korisnickog modela. Najčešci korišćeni korisnički modeli su jos uvek prilično pojednostavljeni, predstavljajući korisnika kao vektor ocene ili korišćenim setom ključnih reči. Čak i tamo gde su multidimenzionalne ili ontoloske informacije dostupne, podaci su generalno mapirani u jednoj korisničkoj stavki koja je pogodnija za većinu podataka i tehnika masinskog učenja. Kako bi se obezbedila najkorisnija i efikasnija preporuka, personalizovani sistemi treba da ugrade izražajnije modele. Neke od rasprava o integraciji semantickog znanja i ontologije u mining procesu ukazuju na to su neki koraci napravljeni u tom pravcu.

Jos jedan važan i težak izazov je modelovanje u korisničkom kontekstu. Posebno profili koji se najčešće koriste danas manjkaju u sposobnosti da modeliraju korisnički kontekst i dinamiku. Korisnici mogu pristupiti različitim stavkama iz različitih razloga i pod različitim kontekstima. Modeliranje konteksta i upotrebe u okviru preporuke generacija treba se dodatno istraziti. Takođe, korisnički interesi i potrebe se menjaju sa vremenom. Indentifikovanje ovih promena i prilagođavanje je ključni cilj personalizacije. Međutim veoma malo istraživanja je potrošeno u evoluciju korisničkih ponašanja tokom vremena i njihovog uticaja na preporuke. Ovo je delom zbog kompromisa između ekspresivnosti profila i skalabilnosti u pogledu broja aktivnih korisnika.

Rešenja za ove vazne izazove će verovatno dovesti do stvaranja nove generacije efikasnije i korisnije Web personalizacije i preporučenih sistema koji mogu biti raspoređeni u vise složenom Web okruženju.



10 Literatura

[1] Sandro Araya, Mariano Silva, Richard Weber, *A methodology for web usage mining and its application to target group identification*, Elsevier B.V. 2004

[2] R. Cooley, B. Mobasher, J. Srivastava: *Web Mining: Information and Pattern Discovery on the World Wide Web*, Department of computer science and engineering University of Minnesota, Minneapolis, MN 55455, USA 1997

[3] Baoyao Zhou, Siu Cheung Hui, and Alvis C. M. Fong, *Web Usage Mining for Semantic Web Personalization*, School of Computer Engineering, Nanyang Technological University, Singapore, 2004

[4] Mobasher, B. Web *usage mining and personalization*. In Singh, M.P., ed.: Practical Handbook of Internet Computing. CRC Press 2005

[5] Cooley, R., Mobasher, B., Srivastava, J.: *Data preparation for mining world wide web browsing patterns*, Journal of Knowledge and Information Systems, 1999

[6] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: *Web usage mining: Discovery and applications of usage patterns from web data*, Department of Computer Science and Engineering, University of Minnesota 1999

[7] Kohavi, R., Provost, F. Applications *of data mining to electronic commerce. Data Mining and Knowledge Discovery, 2001*

[8] Pazzani, M., Billsus, D. *Learning and revising user profiles: The identification of interesting web sites, 1997*

Online izvori:

[9] http://en.wikipedia.org/wiki/Web_personalization

[10] http://en.wikipedia.org/wiki/Web_usage_mining

[11] http://en.wikipedia.org/wiki/Data_mining

[12] http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[13] http://www.igi-global.com/viewtitlesample.aspx?id=10785