Ekonomski horizonti, Maj - Avgust 2019, Volumen 21, Sveska 2, 177 - 193
UDC: 33 ISSN: 1450-863 X

www. ekfak.kg.ac.rs

Pregledni članak
UDK: 004:378.4]:005
doi:10.5937/ekonhor1902177B

TRANSFORMACIJA *WEB* PODATAKA U ZNANJE - IMPLIKACIJE ZA MENADŽMENT

Zita Bošnjak*, Olivera Grljević i Saša Bošnjak

Ekonomski fakultet u Subotici, Univerzitet u Novom Sadu

Preduzeća svakodnevno prikupljaju različite podatke o *online* ponašanju pojedinaca na svojim računarima, koje uključuje njihovo pretraživanje sajtova, *online* kupovine, postavljanje komentara i sl. Ove podatke jednim imenom nazivamo *web* podacima. U njima se kriju indikatori interesa, navika, preferencija ili uobičajenih modaliteta ponašanja pojedinaca, tzv. paterni. Međutim, mnogo je korisnije kada preduzeće prikuplja podatke o svim svojim korisnicima, te analizom ovih podataka stekne uvid u njihove navike i tendencije. Uočavanjem i interpretiranjem ovih paterna, menadžment može donositi odluke koje su adekvatnije, budući da su zasnovane na informacijama i saznanjima deriviranim iz podataka, te bolje zadovoljavaju potrebe komitenata. Postupak otkrivanja paterna, tzv. rudarenje *web* podataka, transformiše iste u znanje. U radu je istraženo na koji način metode i tehnike rudarenja podataka mogu da se primene na *web*-bazirane forme podataka i na koji način otkrivanje paterna u *web* sadržajima, strukturi i načinu upotrebe *web-a* utiče na menadžment.

Ključne reči: nestrukturirani podaci, rudarenje *web-a*, segmentacija komitenata, modelovanje *online* pristupa, kolaborativno filtriranje

JEL Classification: O33, L190, D85, Z13

UVOD

World Wide Web (*web*) je nastao i doživeo ekspanziju devedesetih godina prošlog veka, zahvaljujući širokoj upotrebi mikrokompjutera, razvoju *hardware*-a (pre svega, mikroprocesora, memorijskih elemenata, i tehnologija za skladištenje podataka), razvoju korisnički orijentisanih i jednostavnih za upotrebu *software*-skih alata i neverovatnim mogućnostima koje je *web* kao globalni informacioni sistem stvorio u domenu poslovanja. Tehnologija druge generacije *web*-a je omogućila da korisnici sami kreiraju sadržaj na *web*-u, te se *web* kakav danas poznajemo sastoji od *web* stranica, slika, video i drugih *online* materijala kojima se može pristupiti putem čitača *web*-a, ali i interaktivnih medija (društveni mediji) i korisnički generisanih sadržaja. U želji da saznaju informacije o svojim komitentima, preduzeća više nisu prinuđena da informacije prikupljaju kroz intervjue, upitnike ili kroz drugačiji vid interakcije sa korisnicima, jer su one dobrovoljno već postavljene na *web*u. Međutim,

* *Korenspondencija*: Z. Bošnjak, Univerzitet u Novom Sadu, Ekonomski fakultet u Subotici, Segedinski put 9-11, 24000 Subotica, Republika Srbija; e-mail: bzita@ef.uns.ac.rs

neprekidni porast obima generisanih podataka i diversifikovanje njihovih izvora su značajno usložnili postupke za izbor adekvatnih metrika i sticanje uvida u podatke na osnovu kojih preduzeće može preduzimati određene korake ka poboljšanju poslovanja. Ogromne količine podataka i informacija su otežale pretraživanje i nalaženje određene, specifične informacije, te je sve više vremena potrebno za kolekciju i transformaciju podataka, a sve manje vremena ostaje za strategijsko planiranje (Markov & Larose, 2007, 4). Ovo je dovelo do nastanka i razvoja inteligentnih tehnika i alata koji pružaju podršku u obavljanju sve kompleksnijeg zadatka analize *web* podataka i ekstrakcije informacija i znanja iz njih.

Otkrivanje znanja sadržanog u podacima (*Knowledge Discovery in Data*), je oblast koja se oslanja na dostignuća mašinskog učenja kao dela veštačke inteligencije, sa ciljem ekstrakcije potencijalno korisnih i interesantnih saznanja iz podataka velikog obima. Tradicionalno, tehnike rudarenja podataka (*Data Mining*), bile su primenjene na podatke koji su se kroz vreme prikupljali u relacionim ili transakcionim bazama ili skladištima podataka preduzeća, ali se one, u principu, mogu primeniti na bilo koju vrstu repozitorijuma podataka, pa i na *web* podatke.

U radu su razmatrani koncepti, tehnike i mogućnosti primene rudarenja *web*-a, discipline koja se rapidno razvija zahvaljujući velikim potencijalima za unapređenje poslovanja, naročito e-biznisa. Kako rudarenje *web*-a može da se sprovede kroz različite pristupe i primenom različitih tehnologija, njegova implementacija nije uniformna i jasno određena već varira u zavisnosti od zadataka, domena primene i cilja analize.

Shodno tome, predmet istraživanja su ključni aspekti o kojima je potrebno voditi računa prilikom opredeljivanja za primenu novih tehnologija i tokom sprovođenja analiza nad *web* podacima. Istraživanje se bavi najčešćim vrstama zadataka na koje se ova relativno nova disciplina može uspešno primeniti, te opisuje specifičnosti različitih pristupa u njihovom rešavanju.

U radu smo pošli od hipoteze da rudarenje *web* podataka ima veoma heterogenu primenu koja donosi benefite u mnogim sferama poslovanja, zahvaljujući otkrivenim (sa)znanjima koja bi inače ostala skrivena u mnoštvu *web* podataka.

Da bismo testirali polaznu hipotezu, istražili smo relevantnu literaturu i sagledali o kojim sve primenama i kakvim rezultatima rudarenja *web*-a je izveštavano i u kojim zadacima, koje su prednosti koje rudarenje *web*-a pruža u različitim sferama poslovanja u odnosu na druge analitičke pristupe i koje sve mogućnosti inkorporiranja nove tehnologije u poslovanje stoje menadžmentu na raspolaganju. Shodno tome, najpre je dat pregled različitih načina rudarenja *web* podataka, u zavisnosti od izvora raspoloživih podataka (podaci o posetama *web* stranama, podaci o međupovezanosti *web* strana, podaci o ključnim temama relevantnim za dokumente na *web*u i sl.) i vrste saznanja koja se žele derivirati (da li se želi saznati profil posetilaca *web* stranice koji su tokom posete izvršili *online* kupovinu proizvoda, ili derivirati model za automatizovanu klasifikaciju velikih količina dokumenata u zadate kategorije, ili pak informacije o tome koje proizvode preporučiti posetiocu sajta, kako bi se maksimizirala šansa za potencijalnom *online* kupovinom). Nadalje, ukazano je na empirijski potvrđena iskustva u rudarenju podataka, saopštena u relevantnim izvorima. Fokus rada je na primenljivosti i koristima rudarenja sadržaja, strukture i načina upotrebe *web*-a u poslovnom okruženju, a ne na samim tehnologijama pretraživanja informacija i njihovom efikasnošću.

Struktura rada je sledeća: u drugom delu je dat pregled literature o teorijsko-metodološkim aspektima transformacije *web* podataka u znanje. U trećem delu rada, opisana su tri ključna aspekta rudarenja *web* podataka, i to:

- postupci rudarenja sadržaja *web*-a: algoritam klasterovanja, sistemi za pronalaženje informacija, pristup kolaborativnog filtriranja, i specifični pristup analizi netekstualnih sadržaja;
- rudarenje strukture *web*-a i mogućnosti analize strukture rangiranjem relevantnih stranica; i
- rudarenje upotrebe *web*-a.

Mogućnosti primene pristupa opisanih u prethodnim delovima rada, u različitim sferama poslovanja su opisane u četvrtom delu, uz nekoliko primera dobre prakse. Zaključna razmatranja i budući pravci istraživanja rudarenja *web*-a dati su u petom odeljku.

PREGLED LITERATURE O TEORIJSKO-METODOLOŠKIM ASPEKTIMA TRANSFORMACIJE *WEB* PODATAKA U ZNANJE

Otkrivanje znanja sadržanog u podacima (KDD) je proces kojim se ekstrahuju značajne informacija i (sa) znanja skrivena u mnoštvu podataka i dokumenata. Generička metodologija, koja se najčešće koristi u upravljanju procesom otkrivanja znanja, te poprima obeležja standarda, je CRISP-DM metodologija (*Cross Industry Standard Process for Data Mining*), (Shearer, 2000; Jiawei & Kamber, 2001; Guandong, Yanchun & Lin, 2011). Ona opisuje kontinualan, iterativni proces koji se sastoji iz niza koraka koji su domen-nezavisni, odnosno, primenljivi u svim industrijama i oblastima poslovanja. Prema CRISP-DM metodologiji, neophodno je, najpre, poznavati oblast poslovanja da bi poslovna pitanja mogla da se prevedu u ciljeve rudarenja podataka. U narednom koraku, potrebno je prikupiti i razumeti relevantne podatke, te ih pripremiti za dalju obradu. Postupci pripreme podataka, tzv. preprocesiranje, umnogome zavise od njihove izvorne forme i cilja analize, a sastoje se od integracije, pročišćavanja, filtriranja i transformacije podataka u adekvatnu formu. Pripremu podataka u CRISP-DM metodologiji prati modelovanje, tj. ekstrahovanje skrivenih relacija, zakonomernosti ili shema ponašanja, a primenjuju se naučne metode i tehnike rudarenja podataka. Evaluacija i provera izgrađenog modela treba da pokažu da li on ispunjava poslovne ciljeve. U slučaju pozitivnog ishoda, potrebno je interpretirati uočene paterne ponašanja i razviti plan akcije za dalju implementaciju otkrivenih znanja. Sekvenca navedenih koraka nije striktna, već je uobičajeno vraćanje sa narednih na prethodne korake radi dobijanja boljih međurezultata. Otkrivena saznanja se koriste za definisanje još užih poslovnih zadataka na koje se, kontinualnim rudarenjem podataka, uz stečena iskustva kroz prethodne iteracije, traže odgovori.

IBM korporacija je 2015. osmislila novu metodologiju, poznatu kao ASUM-DM (*Analytics Solutions Unified Method for Data Mining/Predictive Analytics*). Ova metodologija je ekstenzija i poboljšanje postojeće CRISP-DM metodologije (IBM Analythics, 2016), budući da se u većoj meri bavi infrastrukturnim i operativnim aspektima rudarenja podataka i prediktivne analitike, kao i menadžerskim aktivnostima u fazi implementacije.

E. Yoneki, J. M. Tirado, Q. Guo i O. Serban (2016), opisali su ekspert-centričnu metodologiju za ekstrakciju znanja iz *web* podataka. Prema njihovoj zamisli, celokupan zadatak otkrivanja znanja se vezuje za eksperta i njegovu ekspertizu. Naime, od samog početka procesa, odnosno, od koraka prikupljanja podataka, pa sve do ekstrahovanja paterna i implementacije otkrivenih saznanja u poslovanju, ekspertu stoji na raspolaganju skup *software*-skih alata koji mu pomažu da samostalno, uz minimalne intervencije razvojnog tima, sprovede ceo iterativni proces.

Kada rudarenje podataka koristimo za klasterovanje, klasifikaciju, predikciju ili regresiju velikih količina *web* podataka u nameri da izvučemo korist za poslovanje, govorimo o disciplini rudarenja *web*-a (Markov & Larose, 2007; Liu, 2007; Palau, Montaner, Lopez & de la Rosa, 2016). Ona ima tri različita aspekta:

- Rudarenje sadržaja *web*-a (*Web Content Mining*) - zadatak se sastoji od pronalaženja sadržine *web* stranica i rezultata *web* pretrage. Svaka *web* stranica je osmišljena sa namerom da posluži posetiocima sajta, te sadrži relevantne podatke. Međutim, oni nisu ograničeni na tekstualnu formu, već mogu biti u formi slike, grafičkih elemenata ili tabela, što značajno usložnjava analizu sadržaja.
- Rudarenje strukture *web*-a (*Web Structure Mining*) - zadatak je da se analizira hiperlink struktura, odnosno, podaci o načinu na koji su povezane i organizovane *web* stranice. Informacije o strukturi

unutar stranice objašnjavaju kako su uređeni različiti HTML (*Hypertext Markup Language*) ili XML (*Extensible Markup Language*) tagovi unutar posmatrane strane. Informacije o spoljašnjoj povezanosti stanice sa drugim *web* stranama su date u formi mreže *hyperlink*-ova koji povezuju međusobno više stranica.

- Rudarenje upotrebe *web*-a (*Web Usage Mining*) - zadatak je da se analizira navigacioni put posetilaca *web*-a. Podaci koji daju informacije o shemi pristupa *web* stranicama se najčešće čuvaju u proširenom *log* formatu (*Extended Common Log Format*) u *log* datotekama na serverima, i sadrže informacije kao što su: IP (Internet Protokol) adresa, referenca posećene stranice, vreme i mesto pristupa.

Postupci pripreme i modelovanja *web* podataka su specifični za svaki od navedenih aspekata. Za rudarenje sadržine *web*-a karakteristično je predstavljanje dokumenata pomoću preprocesiranih reči koje dokument sadrži, a koje se nazivaju karakteristikama. Karakteristike su dimenzije po kojima se u daljim postupcima analize dokumenti međusobno upoređuju po sličnosti (Liu, 2007). Svi dokumenti se formalno predstavljaju u vektorskom modelu prostora (*Vector Space Model*). To je matrična prezentacija svih dokumenata polaznog skupa, koja se kreira u postupku opisanom u (Cheng, Healey, McHugh & Wang, 2001):

- Za svaku karakteristiku se izračunava stepen učestalosti pojavljivanja ili frekventnost u posmatranom ulaznom skupu dokumenata (*Degree of Frequency*);
- Reči koje se retko pojavljuju se eliminišu iz dalje analize. Vrednost donjeg praga frekventnosti se proizvoljno postavlja na početku procesa rudarenja podataka, čime se redukuje dimenzionalnost analize;
- Svakoj od preostalih karakteristika se pridružuje *n* stepeni značajnosti, po jedan za svaki od *n* dokumenata iz polaznog skupa dokumenata;
- Stepeni značajnosti se zapisuju u formi matrice *n* x *j*, gde *j* označava broj karakteristika ulaznog skupa dokumenata. Ova matrica se naziva vektorski model prostora (*Vector Space Model*), a vrednosti u ćelijama matrice se izračunavaju kao TF-IDF (*Term Frequency-Inverse Document Frequency*) mera, odnosno, odnos frekventnosti termina u posmatranom dokumentu prema broju pojavljivanja istog termina u svim dokumentima ulaznog skupa.

Za pronalaženje informacija od interesa za korisnika, u postupku rudarenja sadržine *web*-a se koriste tehnike filtriranja informacija. Ove tehnike posmatraju temu/predmet dokumenata pri selektovanju onih koji su relevantni za korisnika. Najčešće korišćene metode filtriranja su upiti u vektorskom prostoru (*Vector-Space Queries*), inteligentni agenti (*Intelligent Agents*) i vizuelizacija informacija (*Information Visualization*) (Cheng, Healey, McHugh & Wang, 2001; Kumar & Mohamed, 2018) Kod upita u vektorskom prostoru se pretražuju i rangiraju dokumenti po (kosinusnoj) sličnosti sa vektorskom prezentacijom zadatog upita, što je analogno ranije opisanom postupku izračunavanja sličnosti dokumenta sa centrima klastera.

Performanse sistema za pronalaženje informacija se mogu unaprediti pristupom kolaborativnog filtriranja (*Collaborative Filtering*), koji svojim kvalitetom nadilazi puku analizu sadržine dokumenata, a tiče se atributa kao što su preferencije i ukus korisnika i njihovo poimanje kvaliteta proizvoda, usluga ili drugih entiteta od interesa. J. L. Herlocker, J. A. Konstan, A. Borchers i J. Riedl (1999, 232), izneli su pretpostavku da korisnici *web*-a koji iskazuju slično ponašanje imaju slične interese, tj. da postoji visok stepen korelacije u njihovim preferencijama. Uzimajući ovu pretpostavku za polazište, razvijen je postupak kolaborativnog filtriranja (Resnick & Varian, 1997, 57), koji je osnova sistema za davanje preporuka (*Recommender Systems*).

Po definiciji, rudarenje slika (*Image Mining*) se bavi ekstrakcijom slikovnih paterna iz velikih kolekcija slika (Fayyad, Djorgovski, & Weir, 1996), a ne razumevanjem i/ili ekstrakcijom specifičnih karakteristika jedne slike, niti nalaženjem relevantnih slika. Postupak rudarenja slika se sastoji od skladištenja i procesuiranja, kako bi se poboljšao kvalitet slike, generisanja bitnih karakteristika na osnovu slike, indeksiranja i pronalaženja slika i otkrivanja paterna i znanja (Madhumathi & Selvadoss Thanamani, 2014, 1818). Tehnike rudarenja podataka

se primenjuju na generisane karakteristike i otkrivaju se značajni paterni, od kojih se, nakon evaluacije i interpretacije, dobija finalno znanje, koje se primenjuje u aplikacijama.

Video je primer multimedijalnih podataka, jer sadrži tekst, sliku, meta-podatke, visuelni i audio zapis: video zapis sadrži sekvencu slika sa temporalnim informacijama, dok se audio zapis sastoji od govora, muzike i specijalnih zvukova, dok je tekstualni zapis lingvistička forma unutar videa (Vijayakumar & Nedunchezhian, 2012). Kako bi se video podaci mogli analizirati, moraju se prevesti u strukturiranu formu (Rui & Huang, 2000). Model video podataka je reprezentacija videa koja se bazira na njegovim karakteristikama, sadržini i nameni analize (Kokkoras, Jiang, Vlahavas, Elmagarmid, Houstis & Aref, 2002). Model se pravi segmentacijom ili anotacijom videa. M. Petkovic i T. D. Jonker (2001), predložili su model sa četri nivoa podataka:

- nivo sirovih podataka, sa sekvencom "uramljenih slika" (*Frames*) i video atributima,
- nivo karakteristika, koji sadrži domen-nezavisne karakteristike koje se mogu automatizovano generisati iz sirovih podataka, opisujući boje, tekture, oblike i pokrete,
- nivo objekata, koji sadrži entitete sa istaknutim prostornim dimenzijama, pridružene "uramljenim slikama", i
- nivo događaja, sa entitetima koji imaju istaknute vremenske dodatke koji opisuju kretanje i međusobnu interakciju objekata u prostoru i vremenu.

Postoje dve vrste rudarenja audio podataka (*Audio Data Mining*) (Leavitt, 2002):

- indeksiranje, bazirano na tekstu, koje prevodi izgovoreni tekst u tekstualni zapis te uz pomoć posebnih rečnika identifikuje šta je izgovoreno,
- fonemski-bazirano indeksiranje, koje analizira i identifikuje zvuke u delu audio zapisa kako bi se kreirao indeks.

Rečnik fonema se koristi za prevođenje upita u odgovarajući fonetski niz, a sistem pretražuje indekse kako bi našao onaj koji najviše odgovara zadatom upitu.

Tekst unutar videa se može izdvojiti na više načina (Ma, Lu, Zhang & Li, 2002): iz scene (natpisi na bilbordima, ciradi kamiona ili na majicama), iz mehanički pridruženog tekstualnog sadržaja (dopunske informacije za bolje razumevanje pojedinih sekvenci), ili automatskim prepoznavanjem izgovorenog teksta.

Za rudarenje strukture *web*-a je presudno na koji način će se izračunavati značaj, odnosno, rang *web* strana. Izračunavanje ranga *web* stranice vrši se na osnovu razlike sume svih ulaznih i sume svih izlaznih *link*-ova sa te stranice. Osnovu nešto drugačijeg pristupa izračunavanju ranga, čini ideja da su *hyperlink*-ovi indikatori ljudskog prosuđivanja o međusobnoj relevantnosti, te da pojava *link*-a na stranici *p* ka stranici *q* nosi latentnu informaciju o tome da je autor koji je kreirao stranicu *p*, i u nju uključio *link* na stranicu *q*, „preneo" određeni stepen značaja stranice *p* na stranicu *q*. Posledično, što veći broj linkova sa drugih *web* stranica vodi ka nekoj stranici (ulazni *link*-ovi), to je veći njen značaj, odnosno, autoritet. Kompleksniji pristupi rangiranju uzimaju u obzir širu organizaciju *web*-a, u kojoj, pored autoritativnih stranica sa mnogo ulaznih *link*-ova, otkrivaju i stranice koje su povezane sa mnogim autoritetima, odnosno imaju mnogo izlaznih *link*-ova, te predstavljaju centar razgranavanja ili vrstu „jezgra", tzv. habovi (*Hub Pages*). *Hub*-ovi koji povezuju autoritete za zajedničku temu omogućavaju da se uoče i odbace „lažni" autoriteti, tj. nepovezane stranice sa velikim brojem ulaznih *link*-ova. Svako poboljšanje performansi postojećih metoda pretrage je usko povezano sa pitanjima efikasnosti algoritama pretraživanja i raspoloživim kapacitetima za skladištenje podataka.

Osnovni koncepti rudarenja upotrebe *web*-a objasnio je B. Liu (2007). Rudarenje upotrebe *web*-a koristi podatke o svim aktivnostima posetilaca *web*-a, koji se automatski generišu od momenta prijave na *web*, do momenta odjave: odakle je poseta izvršena, kojom putanjom se posetilac kretao tokom pretraživanja sajta, vreme provedeno na svakoj stranici, kuda je nakon posete stranici korisnik otišao, itd. Ovi podaci se čuvaju u *log* datotekama na serverima. Svaki klik mišem nakon što se korisnik ulogovao odgovara jednom zahtevu za *web* stranicom, a sekvenca klikova odgovara sekvenci linkova na stranice koje je

korisnik posetio. Korisnička poseta *web*-u se naziva sesijom, a podaci o sekvencama stranica posećenih tokom jedne sesije se nazivaju *clickstream* podacima ili *web* klikovima (*Web clicks*). Metoda generisanja asocijativnih pravila iz transakcionih podataka, poznata kao analiza potrošačke korpe (*Market-Basket Analysis*) se može iskoristiti i za analizu *web log* podataka (Jiawei & Kamber, 2001; Li & Feng, 2010). Asocijativno pravilo je pravilo oblika $X \Rightarrow Y$, koje tumačimo tako da kupovina artikla X povlači sa sobom kupovinu artikla Y u istoj transakciji. U *web* okruženju, isto pravilo ukazuje na relaciju između HTML stranica X i Y, koje se učestalo pojavljuju jedna pored druge u korisničkim sesijama (Markov & Larose, 2007). Asocijativno pravilo ne nosi informaciju o hronologiji obavljenih poseta sajtu. Za ovakve vrste analiza se koriste tehnike za generisanje sekvencijalnih asocijativnih pravila, koje uključuju i temporalnu komponentu. U sekvencijalnim pravilima, poseta *web* sajtu navedenom u antecedensu pravila usledila je pre posete *web* sajtovima navedenim u konsekvensu pravila, odnosno, analizira se sekvenca klikova u vremenu (Markov & Larose, 2007.) Postojeći algoritmi za rudarenje upotrebe *web*-a generišu preveliki broj pravila, u kojima je teško razlučiti koja pravila su bitna, a koja nemaju potencijalnu vrednost za korisnika (Cheng, Healey, McHugh & Wang, 2001). Stoga je potrebno odrediti kriterijume kojima će se osigurati implementacija samo korisnih asocijativnih pravila, dok će se odbaciti sva pravila koja se u nedovoljnoj meri mogu iskoristiti za postizanje željenih efekata analize. Ove kriterijume nazivamo merama interesantnosti pravila (Tan, Kumar & Srivastava, 2004; Hilderman & Hamilton, 2013).

Pitanje tačnosti analize *log* datoteka, usled keširanja *web* stranica i potreba da se zadatak delegira preduzećima specijalizovanim za *web* analitiku dovelo je do nastanka drugačijeg pristupa prikupljanju podataka za analizu upotrebe *web*-a, tzv. tagovanja stranica (*Page Tagging*). Metod koristi *JavaScript* ugnežđen na *web* stranici, kako bi svaki put kada korisnik putem čitača *web*-a zatraži stranicu, ili klikne mišem na link, generisao zahtev analitičkom serveru kod trećih lica. Oba metoda prikupljanja podataka o upotrebi *web*-a se mogu koristiti za izveštavanje o saobraćaju na *web* stranicama.

ASPEKTI RUDARENJA *WEB* PODATAKA

Neverovatne mogućnosti *web*-a kao globalnog informacionog sistema u domenu poslovanja možemo posmatrati sa tri aspekta: mogućnosti koje pruža analiza sadržaja *web*-a, koje proizilaze iz analize njegove strukture i koje stvaraju (sa)znanja o njegovoj upotrebi.

Rudarenje sadržaja *web*-a

Rudarenje sadržaja *web*-a se definiše kao „istraživački metod za izradu ponovljivih i validnih zaključaka iz podataka o njihovom kontekstu" (Krippendorff, 1980, 36). Sadržaj *web*-a, kao repozitorijum podataka koje su korisnici postavili na uvid javnosti, nosi informacije o stavovima, preferencijama, mišljenjima i ponašanju korisnika *web*-a, što može biti veoma značajno u mnogim sferama poslovanja.

Prvi aspekt *web*-a, kao globalnog informacionog sistema, je automatizovano pohranjivanje, pristupanje, pronalaženje, organizovanje i predstavljanje podataka u *web* dokumentima. Pojam dokumenta je ranije bio vezan za tekstualne datoteke generisane u nekom programu za obradu teksta, ali se danas koristi za opisivanje bilo kojeg tipa datoteke koja se nekom aplikacijom može generisati. *Web* stranica je, takođe, dokument, koji je obično pisan u HTML jeziku i kojem se može pristupiti putem čitača *web*-a, navođenjem URL adrese. Osim tekstualne sadržine, pisane nekim od prirodnih jezika u slobodnoj formi (bez unapred određene ili propisane strukture, ili je ona samo delimično uređena), *web* stranica može da sadrži slike, audio i video zapise i *hyperlink*-ove na druge dokumente. Najšire posmatrano, e-servisi, arhivirani mejlovi, i drugi sadržaji se, takođe, mogu podvesti pod *web* dokumente.

Suština rudarenja sadržaja *web*-a jeste u sposobnosti pronalaženja dokumenata relevantnih za korisnika. Analiza sadržaja je široko rasprostranjen metod za objektivno i sistematično kvantitativno ispitivanje sadržaja koji se prenosi, a može biti koristan za otkrivanje ili uvid u preferencije i ponašanje korisnika, u kompleksne društvene i komunikacione trendove i paterne koje korisnici generišu (Kim &

Kuljis, 2010, 373). Zbog polu- ili nestruktuirane forme dokumenata, interaktivnosti, decentralizovanosti i mrežne strukture hiperlinkova, klasični sistemi za upravljanje bazama podataka ne mogu poslužiti za pronalaženje potrebnih dokumenata u mnoštvu *web* dokumenata, već se koriste noviji pristupi.

Klasterovanje

Zadatak klasterovanja može se opisati kao segmentacija heterogenog polaznog skupa u podskupove elemenata sa visokim stepenom međusobne kohezije. U kontekstu rudarenja *web*-a, polaznu populaciju mogu činiti kako dokumenti, tako i *web* stranice, koje grupišemo u podgrupe prema njihovom značenju i smislu termina navedenim u njima (klasterovanje po sličnosti), bilo korisnici *web*-a, na osnovu aktivnosti koje obavljaju kada dođu na *web*. Kada je u pitanju rudarenje sadržaja *web*-a, klasteruju se dokumenti, dok se prilikom rudarenja upotrebe *web*-a klasteruju *web* posetioci na osnovu načina korišćenja *web*-a. Iako se termini korisnik i posetilac *web*-a često koriste kao sinonimi, u kontekstu rudarenja *web*-a se oni razlikuju. Razlika proizilazi iz vrste podataka koje se automatizovano generišu i pamte o korisnicima/posetiocima *web*-a. Termin korisnik je opštiji i označava pojedinca koji pristupa *web*u. Korisniku se pridružuje jedinstveni identifikacioni kod, koji ga razlikuje od drugih korisnika *web*-a, a ostaje nepromenjen čak i ako korisnik pristupa *web*u sa različitih uređaja ili promeni svoj *web* pretraživač. Pojam posetioc označava pojedinca koji pristupa *web*u sa određenog uređaja (laptop, mobilni telefon, i dr.), koristeći određeni pretraživač (Google, 2019). Na taj način se za jednog korisnika vezuje više posetilaca *web*-u. Za potrebe analize *web* podataka je značajno da pristup korisnika sa drugog uređaja ili iz drugog pretraživača generiše drugi identifikacioni kod posetioca, jer se na taj način mogu uočiti razlike u *online* ponašanju istog korisnika, npr. drugačije se ponaša u slučaju kada pristupa *web*-u sa laptop računara nego ako mu pristupa iz aplikacije na mobilnom telefonu. Shodno rečenom, u nastavku će se termin posetioc *web*-a koristiti jedino kada je bitno naglasiti da se prilikom rudarenja podataka uzima u obzir uređaj/pretraživač kojim se korisnik služi. U suprotnom će se koristiti termin korisnik.

Algoritmi klasterovanja koriste matricu vektorskog prostora za izračunavanje optimalnog broja klastera i njihovih centara. Udaljenosti svakog dokumenta ulaznog skupa do izračunatih centara klastera se zapisuju u vektorski model dokumenta. Na osnovu nje, vrši se diskretno klasterovanje dokumenata u onu particiju polaznog skupa do čijeg centra je stepen sličnosti dokumenta najveći. Algoritmi klasterovanja mogu pronaći dokumente koji su relevantni za korisnika veoma brzo i tačno, pa se mogu smatrati efikasnom tehnologijom za pronalaženje informacija na *web*-u (Fan, Liu, Tong, Zhao, Nie, 2016, 42).

Sistemi za pronalaženje informacija

Uobičajeno je da korisnici *web*-a iskazuju potrebu za određenom informacijom u formi korisnički definisanog upita. Nakon što korisnik specificira svoj upit, pronalaze se relevantni dokumenti, u skladu sa potrebama korisnika. Sistemi za pronalaženje informacija (*Information Retrieval Systems*) koriste tehnike filtriranja i pretražuju dokumente po predmetu ili temama za koji se vezuje sadržaj, te na taj način prevazilaze nedostatke sistema za upravljanje strukturiranim repozitorijumima podataka. Istovremeno, zadatak sistema za pronalaženje informacija je da isključe iz rezultata pretrage što više dokumenata irelevantnih za korisnički upit. Alati koji filtriraju sadržaj dokumenata upoređuju formalnu reprezentaciju sadržine koja se nalazi u *web* dokumentima sa formalnom reprezentacijom sadržine koja interesuje korisnika, a koja je navedena kroz korisnički upit, te na taj način selektuju pravu informaciju za svakog korisnika.

Kolaborativno filtriranje

U realnom svetu se često dešava da nam prijatelji daju savet, ili preporuku, o interesantnim proizvodima koje bismo trebali kupiti, knjizi koju vredi pročitati, filmu koji bi nam se mogao dopasti i sl. Preporučeni entiteti mogu biti i *online* resursi (*web* stranice) ili *online* aktivnosti (registracija na forum ili grupu,

online kupovina i dr.). Formalno, za ovakav scenario kažemo da oni sarađuju - vrše kolaboraciju sa nama u procesu selekcije. Koristeći postupak kolaborativnog filtriranja, sistemi za davanje preporuka savetuju korisnike na osnovu informacija o ponašanjima i preferencijama drugih korisnika.

Ovi sistemi neretko podstiču korisnike da eksplicitno ocene entitet, ili iznesu svoje lične preferencije, koje se memorišu u sistemu. Korisnici sistema za kolaborativno filtriranje dele svoje analitičko prosuđivanje i rangiranje entiteta (ocenjivanje kupljenih proizvoda ili usluga), tako da drugi korisnici mogu lakše odlučivati koje proizvode da kupe ili koju drugu akciju da preduzmu. Na osnovu ocenjenog prethodnog iskustva i stepena zadovoljstva ranijih korisnika, sistem za preporučivanje kreira personalizovanu preporuku entiteta koji bi mogli biti interesantni novom korisniku. Na osnovu ocena korisnika, moguće je vršiti njihovo klasterovanje u segmente onih koji imaju isti ukus ili informacione potrebe.

Performanse sistema za davanje preporuka su upravo proporcionalne stepenu kolaboracije korisnika (Palau, Montaner, Lopez & de la Rosa, 2016, 145). Zato je metoda kolaborativnog filtriranja naročito korisna u svetu koji je sve više umrežen putem interneta i u kojem se mreža dokumenata na *web*-u gradi zajedničkim naporima samih korisnika.

Analiza netekstualnog sadržaja web dokumenata

Iako sadržaj *web* dokumenata, najvećim delom, čine tekstualni podaci, u dokumentima se mogu naći slike, video i audio zapisi, arhivirani mejlovi i drugi sadržaji. Slike se mogu automatizovano klasifikovati ili klasterovati po vrednostima bazičnih boja (RGB komponenta), ili vrednostima teksture. Kao mera sličnosti slike sa zadatim graničnim uslovima se koristi entropija. Rudarenje slika je veliku primenu našlo u medicini, gde se medicinske slike klasifikuju radi potvrđivanja postavljenih dijagnoza (Babu & Mehtre, 1995).

Cilj rudarenja video podataka (*Video Data Mining*) je ne samo da se automatizovano otkrije sadržina i struktura videa, karakteristike pokretnih objekata i njihova prostorna i vremenska korelacija, već se otkrivaju paterni u aktivnostima objekata i snimljeni događaji, uz veoma malo znanja o samoj sadržini videa. Audio zapis igra značajnu ulogu u detekciji i prepoznavanju događaja u videu. Rudarenje audio podataka se može koristiti za automatizovano razlikovanje više govornika, analizu izgovorenog teksta, detekciju emocija, i sl. Potencijalne aplikacije rudarenja videa uključuju anotaciju, pretraživanje, rudarenje saobraćajnih informacija, detekciju događaja ili anomalija iz videa snimljenih nadzornim kamerama, analizu i uočavanje paterna i trendova. U poređenju sa rudarenjem drugih tipova podataka, rudarenje video podataka je tek u povoju.

Analiza strukture *web*-a

Milioni *online* korisnika *web*-a, sa različitim namerama i pobudama, kontinualno kreiraju hiperlinkovima povezanu sadržinu *web*-a. Zbog toga je struktura *web*-a kompleksna i nemoguće ju je planirati ili uticati na njenu evoluciju. Rudarenje strukture *web*-a je proces saznavanja informacija iz organizacione strukture *web* stranica, koja se sastoji od *hyperlink*-ova koji povezuju određene stranice formatirane u HTML jeziku. Grubo posmatrano, zadatak je otkriti one stranice koje su relevantne za postavljeni upit, pri čemu je ocena kvaliteta pretrage predmet subjektivnog rasuđivanja ljudi, usled inherentnog ličnosnog aspekta kriterijuma relevantnosti.

Korisnici *web*-a mogu postaviti specifične upite, za koje postoji mali broj dokumenata koji sadrže odgovor na navedeni upit. U ovakvim slučajevima, teško je identifikovati malobrojne relevantne dokumente koji sadrže zahtevanu informaciju u mnoštvu *web* dokumenata. Ovaj fenomen je poznat kao problem retkog pojavljivanja (*Scarcity Problem*). Nasuprot ovog scenarija, korisnici *web*-a mogu postaviti upit čija tema je opšta ili široka, te se može naći na hiljade relevantnih *web* stranica koje sadrže traženu informaciju i odgovaraju na zadati upit. Ovaj problem je poznat kao problem preobimnosti (*Abundance Problem*). U ovom scenariju je potrebno pronaći manji podskup najznačajnijih, najvažnijih dokumenata, odnosno, filtrirati iz velike rezultujuće kolekcije

tzv. autoritativne dokumente. Budući da nisu uvek najautoritativnije stranice na kojima se neki pojam učestalo pojavljuje, već to mogu biti stranice na kojima se pojam ne mora ni pojaviti (Honda, kao vodeći proizvođač automobila na svom sajtu ne pominje termin „proizvođač automobila"), jasno je da tekst-bazirani pristupi nisu adekvatni za rangiranje stranica po autoritetu, već se koriste pristupi koji se oslanjaju na hiperlinkove koji povezuju *web* stranice. Rezultati pokazuju da tekst, sadržan u dokumentu koji navodi neki drugi dokument kao referencu, često ima veću diskriminativnu i deskriptivnu vrednost nego tekst u originalnom dokumentu (Glover, Tsioutsiouliklis, Lawrence, Pennock & Flake, 2002, 566).

Rangiranje relevantnih stranica

Postoji više pristupa pronalaženju relevantnih stranica u kontekstu *hyperlink*-ovske strukture *web*-a. Oni se oslanjaju na postupak rangiranja stranica, pri čemu se pod rangom podrazumeva pridružen numerički pokazatelj. Ovakvo posmatranje je previše pojednostavljeno, s obzirom na to da je broj ulaznih *link*-ova veliki kod popularnih stranica (www.yahoo.com), pa bi se oni mogli smatrati relevantnim za sve upite.

Iako su *link*-ovi i sadržina i dalje među najznačajnijim parametrima za rangiranje *web* stranica, savremeni sistemi za rangiranje uključuju mnogo više pokazatelja, od kojih nijedan zasebno ne daje pravu sliku ranga, ali nastoji da najznačajnije stranice istakne i njihov sadržaj učini vidljivijim. Google u svom pretraživaču koristi preko 200 faktora za određivanje ranga stranice, ali njihova specifikacija potpada pod poslovnu tajnu (Search Engine Land, 2017).

Rudarenje upotrebe *web*-a

Podaci koji su prikupljeni u log datotekama na *web* serverima imaju određeni stepen analogije sa podacima iz transakcionih baza podataka o kupovinama. Svaka zahtevana *web* stranica može se smatrati analognom jednom artiklu u transakciji, dok se skup svih stranica koje je određeni korisnik tražio tokom jedne posete *web* sajtu može smatrati analognim jednoj transakciji, odnosno, skupu artikala koji su se našli u potrošačkoj korpi prilikom kupovine. Primer asocijativnog pravila o određenom *web* sajtu mogao bi da glasi: „Ako je korisnik zahtevao *web* stranice *A*, *B* i *C*, postoji indikacija sa merom sigurnosti od 23% da će takođe zahtevati stranice *D* i *E*."

Rudarenje (načina) upotrebe *web*-a je fokusirano na analizu podataka o posećenim *web* stranicama tokom korisničkih sesija i omogućava ekstrahovanje saznanja o ponašanju posetilaca *web*-a. Interesantne zakonomernosti, ili sheme ponašanja, kojih ranije nismo bili svesni, donose potencijalnu korist kako vlasnicima sajta, tako i posetiocima. Poteškoća sa kojom se susreću analitičari podataka prikupljenih u *log* datotekama na *web* serverima, je tendencija generisanja prevelikog broja rezultujućih pravila, među kojima je, čak i upotrebom mera interesantnosti, teško odabrati ona pravila koja daju informacije korisne u datom domenu primene. Nadalje, obe metode prikupljanja podataka o upotrebi *web*-a (iz *log* fajlova i tagovanjem), mogu se koristiti za izveštavanje o saobraćaju na *web* stranicama. Za koji pristup će se preduzeće opredeliti, zavisi i od razlike u ceni analize ako se ona obavlja "u kući" ili angažovanjem spoljnih saradnika.

MOGUĆNOSTI PRIMENE RUDARENJA *WEB*-A

Najširu primenu analize sadržaja *web*-a nalazimo u društvenim i humanističkim domenima, jer su u sadržaje utkani svi socijalni procesi značajni za korisnike *web*-a - kroz upotrebu simbola, preko poruka koje se prenose, kroz uključene slike ili fenomen organizovanja. I. Kim i J. Kuljis (2010), opisali su specifičnu upotrebu analize sadržaja *web* dokumenata u ispitivanju kulturološkog uticaja na dizajn i upotrebu *blog*-ova. S obzirom na to da korisnici samostalno kreiraju i održavaju *blog*-ove, pretpostavka koja je ispitana je da blogovi oslikavaju sistem vrednosti i preferencije autora, koji su rezultat kulturološkog nasleđa. Rezultati istraživanja u Južnoj Koreji i Velikoj Britaniji su opovrgli pretpostavku i pokazali da *blog*-eri iz zemalja koje tradicionalno

imaju nizak stepen tolerantnosti na rizik i nesigurnost (kao što je Južna Koreja), manje verovatno otkrivaju personalne informacije o zaposlenju, nego *blog*-eri iz zemalja koje su tradicionalno tolerantne na rizik i neizvesnost (kao što je Velika Britanija), ali da zato češće daju informacije o godinama i *link* za kontakte.

Struktura *web link*-ova može ukazivati na sličnost među povezanim *web* stranicama. Stoga, ukoliko smo pronašli stranicu *p*, koju smatramo relevantnom za neku određenu temu, odnosno, dovoljno autoritativnom za datu temu, uvidom u *link*-ove koji okružuju stranicu *p* možemo dobiti odgovor na pitanje koje druge teme su po mišljenju korisnika koji su kreirali stranice i *link*-ove na *web*u, povezane sa polaznom temom. Ako je stranica *p* visoko referentna, postojaće enormno velik broj nezavisnih mišljenja o povezanosti stranice *p* sa drugim stranicama, ali ako iskoristimo notacije autoriteta i *hub*-ova, dovoljno je da pronađemo najbliže (lokalne) autoritete oko *p*. Lokalni autoriteti su neka vrsta sažetka opštih tema koje su u vezi sa temom stranice *p*. Na osnovu uvida u *hyperlink* strukturu *web*-a, možemo kategorisati *web* stranice ili steći uvid u hijerarhijsku ili mrežnu strukturu postojećih veza na *web* sajtu, u okviru određenog domena.

Po analogiji sa merom autoritativnosti *web* stranica, analiza strukture *web* linkova se može iskoristiti za merenje statusa, uticaja (*impact*-a) na društvenim mrežama. U tom smislu su razvijeni mehanizmi za merenje „značaja" pojedinaca kroz citiranost njihovih radova, ili pak za merenje *impact* faktora časopisa u implicitno kreiranoj mreži naučnih radova koji su povezani citiranim referencama, a koje se koristi u bibliometrijskim evaluacijama. Najpoznatija mera autoritativnosti u ovom kontekstu je *impact* faktor časopisa (Garfield, 1972), koji se za posmatranu godinu izračunava kao prosečna vrednost broja citata naučnih radova objavljenih u tom časopisu u prethodne dve godine (Egghe & Rousseau, 1990). *Impact* faktor je, u osnovi, mehanizam rangiranja koji prilikom izračunavanja autoriteta uzima u obzir samo ulazne *link*-ove u mreži.

Podaci o upotrebi *web*-a ukazuju na interakciju korisnika i sajta, na osnovu koje se kreira korisnikov model kao slika njegovog ponašanja, interesovanja i ličnih preferencija. Rudarenje upotrebe *web*-a, u osnovi, ima dve primene: analizu saobraćaja na *web* stranicama, i primenu u e-komercu. Analiza saobraćaja uključuje analizu podataka o navigacionoj putanji korisnika koju je sledio tokom svoje sesije. Beleže se i podaci o tome koliko puta je tokom pretraživanja učitao određene stranice. Analiza obično otkriva detalje o posetiocima, kao što je stil pretraživanja, preferencije, prosleđivanje proizvoda/usluga prijateljima, broj klikova na *link* i pogodaka na određenu stranicu, itd. Zahvaljujući otkrivanju asocijativnih pravila, ponašanje posetilaca se može predvideti upoređujući sheme pretraživanja sajta sa shemama pretrage koje su ekstrahovane rudarenjem *log* podataka drugih korisnika. Na osnovu prepoznatih sličnosti i uočenih zajedničkih interesovanja, sistemi za preporučivanje mogu sugerisati posetiocima najadekvatniju putanju do preferiranih sajtova, ili pak uobičajenu putanju do određene elektronske kupovine i personalizovati *web* sadržaj (Siddiqui & Aljahdali, 2013, 42).

Velike kompanije širom sveta su odavno shvatile da e-komerc nije puka prodaja i kupovina putem interneta, već mogućnost povećanja efikasnosti u konkurentskoj borbi na tržištu, iskorištavanjem znanja skrivenih u velikim količinama podataka raspoloživih *online*. Integracija tehnika rudarenja *web*-a sa e-komerc aplikacijama omogućava vlasnicima e-prodavnica da poboljšaju svoje performanse i usluge, te prikupe informacije o korisnicima i ponašanju korisnika koji pristupaju proizvodima ili uslugama putem *web* sajta. E-komerc sajtovi generišu podatke koji nose informaciju o razlozima, dinamici i načinima navigacije kroz *web* sajtove, te njihova analiza može ukazati na bolji pristup željenim *web* sadržajima, poboljšati tok kupoprodajnih procesa i generisati novu vrednost za konzumente. U nameri da se udovolji potrebama posetilaca u što većoj meri, pogotovo komitentima i lojalnim konzumentima, tokom njihove posete se na *web*-u nude kastomizovane usluge. Pojedini sajtovi, na osnovu uočenih sličnosti u ponašanju posetilaca i predviđenih interesa posmatranog posetioca, pripremaju kastomizovane kataloge proizvoda (Shukla, Silakari & Chande, 2013, 8). Većina vodećih e-komerc preduzeća u svetu

(Yahoo!, Amazon, eBay, IBM, itd.), prilagodila je svoje *web* prezentacije i sisteme za preporučivanje za personalizovani pristup.

Rudarenje *web* podataka je značajno u e-komercu, za upravljanje odnosima sa komitentima (Li & Feng, 2010, 279). Naglasak je na privlačenju novih komitenata, zadržavanju postojećih, unakrsnoj prodaji proizvoda i odlivu potrošača. Profilisanje posetilaca omogućava preduzećima da predvide ko su njihovi potencijalni komitenti ili konzumenti proizvoda/usluga, i kakve sheme ponašanja mogu očekivati od njih. Rudarenje upotrebe *web*-a rezultuje analitičkim pregledom ponašanja posetilaca kada posete *web* stranicu preduzeća; uvidom u to koji deo populacije oni čine (u odnosu na njihove godine, pol, lokaciju, i druge karakteristike); kako su dospeli na *web* stranicu preduzeća; šta su najčešće posećivali (koji sadržaj je najinteresantniji za njih); ukupan učinak poseta; i dr. Rudarenje upotrebe *web*-a može biti bitno za marketinške aktivnosti, jer otkriva učestale navigacione puteve kroz *web* sajt preduzeća. „Svaki put kada posetilac sajta klikne na link, sliku, ili neki drugi objekat na stranici, ta infomacija biva zabeležena i zapamćena. Možete otkriti navike svakog pojedinca, ali je korisnije kada memorišete na hiljade navigacionih putanja i saznate globalne navike i tendencije vaših korisnika“ (Jokar, Honarvar, Hamirzadeh & Esfandiari, 2016, 321). Agregirane informacije u formi sveukupne statistike poseta su korisne donosiocima odluka, jer su jasan indikator učestalosti posećivanja svake stranice, vremena provedenog na svakoj stranici, aktivnosti posetilaca na određenoj stranici, broja poseta spram broja komercijalnih aktivnosti naručivanja ili kupovine, popularnosti svih proizvoda ili usluga, raspoloživih izbora za konzumenta, itd. Informacija o posećenim stranicama i njihovom redosledu, takođe, govori i o najčešćem mestu napuštanja *web* sajta preduzeća, kao i o tome gde je najviše proizvoda dodato u potrošačku korpu ili uklonjeno iz nje. Analizom sekvence klikova se može utvrditi efikasnost sajta. Za to je potrebno kvantifikovati ponašanje posetilaca tokom sesije, odnosno evidentirati sve prodajne transakcije tokom posete. Asocijativna pravila o upotrebi *web*-a mogu potpomoći bolju organizaciju sadržine, ili dati preporuke za efektivnu unakrsnu prodaju (Liu, 2007). Sekvenca klikova može ukazati na najbolje mesto za postavljanje oglasa, ukoliko znamo sa kojih stranica su posetioci došli na posmatrani sajt, a može pokazati da li je *online* marketinška kampanja uspešna ili ne (postoji li veza sa kupovinama preko *web*-a). Podaci o navigaciji kroz *web* stranice mogu ukazati na (ne) adekvatnost *online* obrazaca, načina selekcije robe za virtuelnu korpu i načina plaćanja.

Osim e-komerc sajtova, postoje i sajtovi koji nisu namenjeni prodaji, već služe kao katalozi proizvoda, dok se transakcije obavljaju *offline*. Praćenje *clickstream* podataka na ovim sajtovima pruža napredne informacije o potražnji, značajne za snabdevanje, planiranje zaliha i proizvodnju (Huang & Mieghem, 2014, 334). Autori ističu da preduzeća koja koriste podatke o posetama i klikovima mogu smanjiti troškove skladištenja i naručivanja za 3-5%.

Prednosti rudarenja upotrebe *web*-a se koriste i za poboljšanje performansi *web* servera i njihovih aplikacija, kroz različite strategije keširanja sadržaja i pred-pristupa (*Pre-Fetching*), kako bi se skratilo vreme odziva na postavljeni upit korisnika. Istovremeno vlasnici *web* stranica mogu poboljšati upotrebljivost *web* strana kroz bolji dizajn i implementaciju, npr. optimizovati sajt za pretragu po često korištenim ključnim rečima ili eliminisati posrednike preko kojih su posetioci došli na sajt.

Primeri dobre prakse

AMSOIL Inc, prvi proizvođač sintetičkog maziva za vozila i mašine, sa preko četrdeset godina tradicije, opredelio se za *software*-sko rešenje *Mozenda,* za pretraživanje *web*-a kao podršku svom poslovanju. Mozenda Cloud-Hosted Web Harvesting Solution omogućava AMSOIL-u da ekstrahuje i organizuje nestrukturirane *web* podatke kako bi razvio i održavao resurse za planiranje (Mozenda, 2018). Tim za strateško planiranje u AMSOIL-u koristi podatke koje Mozenda prikuplja sa *web*-a za planiranje distributivne mreže u maloprodaji. Ovo, pre svega, podrazumeva mapiranje maloprodaja i servisnih lokacija, na kojima se obavlja zamena ulja za vozila i mehanizaciju, kako bi se identifikovale potencijalne lokacije za nove maloprodaje. Istraživanje konkurentskih cena

je jedna od najčešćih primena opcije prikupljanja *web* podataka za svako preduzeće koje se bavi *online* prodajom. Alat Mozenda neprekidno prikuplja javno dostupne informacije sa e-komerc sajtova, kako bi se bolje razumelo tržište i pratile kategorije proizvoda i konkurentske cene naspram cena AMSOIL-a. Od e-komerc operacija, preko maloprodaje, do planiranja maloprodajne mreže, brzi pristup nestrukturiranim *web* podacima pomaže AMSOIL-u da neometano rešava različite strateške i taktičke zahteve i da se uspešno nosi sa većim brendovima, kao što su: Mobil, Pennzoil, Shell, Castrol i Valvoline.

Tesco.com je e-komerc ogranak Tesco PLC-a, britanskog trgovca prehrambenim proizvodima i robom široke potrošnje, koji posluje u Velikoj Britaniji, Evropi, Aziji i Severnoj Americi. Predstavljen 2000, Tesco.com trguje prehrambenim proizvodima, robom široke potrošnje, odećom, bankarskim i osiguravajućim uslugama. Tesco programeri i analitičari podataka, radi boljeg razumevanja koje proizvode i koje *web* stranice korisnici koriste i koji su navigacioni putevi rezultovali najvećim brojem konverzija, uveli su *software*-ski paket Splunk Enterprise, koji je preduzeću omogućio da na osnovu novih uvida u *online* podatke poboljša zadovoljstvo komitenata; smanji potencijalne gubitke prihoda; ubrza razvojni ciklus i poboljša saradnju među timovima. Splunk digital intelligence solutions (Splunk, 2013), prevazilazi klasičnu marketing analitiku i pruža celokupan uvid u korisnikovu interakciju kroz različite digitalne kanale, uključujući *web*, mobilnu komunikaciju, društvene medije i *offline* interakciju. Splunk Enterprise rešenja se koriste zajedno sa drugim alatima za *web* analitiku, da bi se ekstrahovale informacije iz istorijskih i podataka u realnom vremenu, generisanih kako na klijentskoj, tako i serverskoj strani aplikacije.

Proizvođač nameštaja i artikala za opremanje enterijera, IKEA, sa preko 60 godina iskustva u maloprodajama, u preko 40 zemalja sveta, se, takođe, opredelila za inteligentnu analizu *web* podataka kako bi unapredila svoje poslovanje (Harapiak, 2013). Prevashodno, demografski i psihografski podaci o kupcima iskorišteni su za otkrivanje njihovih navika življenja i, konsekventno, za unapređenje vizuelnog doživljaja doma kroz izložbene salone uređene kao realni životni prostor. Analizirajući podatke o ponašanju potrošača, IKEA je otkrila šablon kupovine koji su nazvali domino efektom: kupovina samo jednog komada nameštaja povlači za sobom određene druge kupovine, kako bi se on što bolje uklopio u postojeći prostor. Takođe, otkrili su da demografski podaci komitenata, pored podataka o izvršenim transakcijama, nisu dovoljni za analizu njihovih potrošačkih navika, jer ne nose informaciju o njihovom sistemu vrednosti. Uključivanjem psihografskih podataka, otkrili su zanimljive šablone ponašanja. Jedan od njih se ticao usko specifičnih navika stanovnika Pitsburga, za koje su posebno kreirali katalog u kojem su prevashodno isticali cenovnu prednost na koju su u ovoj regiji stanovnici bili posebno osetljivi, za razliku od drugih karakteristika nameštaja za koji nisu iskazivali senzibilitet (Dudovskiy, 2017). U IKEA-i ističu da oslanjanjem na klasične analize ovakvu specifičnost nikada ne bi uočili.

AstraZeneca Co. je britansko-švedsko multinacionalno farmaceutsko i biotehnološko preduzeće, sa sedištem u Londonu. Nakon implementacije *software*-a za rudarenje podataka, AstraZeneca je ostvarila značajno poboljšanje standarda za svoja medicinska istraživanja. Platforma za prikupljanje podataka iz najrazličitijih biomedicinskih izvora, Linked Life Data, je omogućila interaktivno otkrivanje uzročno-posledičnih veza biomemedicinskih entiteta i objašnjenje uočenih kauzaliteta, što je doprinelo da se uključe/isključe određeni kriterijumi iz kliničkih ispitivanja (AstraZeneca, 2019). Alat Lexiquest Mine application, čiji vendor je SPSS, je potrebne informacije ekstrahovao iz nestrukturiranog tekstualnog sadržaja iz repozitorijuma koji svakodnevno primi na hiljade novih tekstualnih dokumenata (Thomas, 2002).

LIAT (*Leeward Islands Air Transport*) je avionska kompanija sa Kariba, koja opslužuje međuostrvski saobraćaj na 15 destinacija. Preduzeće je rudarenjem tekstualnih poruka putnika unapredilo korisnički servis. Izgrađeni su prediktivni klasifikacioni modeli u alatu RapidMiner, koji su preusmeravali poruke korisnika relevantnim odeljenjima (RapidMiner, 2019a). Na taj način je Odeljenje za odnose sa

korisnicima oslobođeno manuelnog prosleđivanja poruka, te su se mogli fokusirati na davanje kvalitetnih odgovora. LIAT tvrdi da je negativni sentiment korisnika na društvenim mrežama opao sa 90% na 40%. Ovaj rani uspeh je otvorio put primeni rudarenja podataka u drugim poslovnim operacijama.

Vodeći provajder mobilne telefonije u Austriji, Mobilkom Austria, prima više od 800.000 e-mail poruka svakog meseca, od kojih čak i nakon filtriranja *spam*-a (neželjena pošta) ostaje oko 80.000 korisničkih zahteva. Budući da u komunikaciji sa preduzećem korisnici očekuju brz odgovor, Mobilkom je započeo rudarenje sadržine ovih *e-mail*-ova u alatu RapidMiner - Data Science Platform. *E-mail*-ovi su automatizovano klasifikovani po temama i prosleđeni osoblju za podršku, koje je bilo kompetentno za datu temu (RapidMiner, 2019b). Na ovaj način je zagarantovan kvalitetan odgovor u najkraćem mogućem roku.

PayPal je najbrži i najbezbedniji *online* način plaćanja i primanja novca iz celog sveta. Sa 143 miliona aktivnih računa na 193 tržišta i 26 valuta, PayPal omogućava globalnu trgovinu, u kojoj dnevno obradi preko 8 miliona transakcija. Jedan od neprekidnih zadataka kompanije je upravljanje zadovoljstvom korisnika i smanjivanje njihovog odliva. U pozadini je saznanje o tome šta korisnike čini zadovoljnima i na koji način se njihov doživljaj proizvoda može poboljšati. Primenjujući analizu tekstualnih povratnih informacija od korisnika iz 60 zemalja sveta, uspeli su da identifikuju „top promotere" i „top ogovarače" preduzeća (RapidMiner, 2019c).

Sve više provajdera *software*-skih alata (Featured Customers, 2019), bilo da su ti alati u celosti posvećeni rudarenju podataka ili pak imaju zasebne platforme ili module za inteligentne analize, izveštavaju na svojim *web* stranicama o pozitivnim iskustvima i najrazličitijim primerima dobre prakse u rudarenju *web*-a.

ZAKLJUČAK

Inkorporacija dostignuća veštačke inteligencije u svim domenima poslovanja, smatra se danas jednim od ključnih strateških opredeljenja, koje zahteva ne samo veliki stepen inovativnosti, već i transformaciju poslovanja. Zbog kompleksnosti novih tehnologija i promena koje one zahtevaju, svi koji žele ostati konkurentni u budućnosti, moraju započeti sa zaokretom ka novim tehnologijama i eksploatacijom njihovih mogućnosti već sada. Imajući ovo na umu, u radu su prikazani mnogostruki aspekti, zadaci i mogućnosti rudarenja *web* podataka, jer njihova adekvatna primena omogućava preduzećima ekstrakciju vrednih informacija i znanja, koji inače ostaju skriveni u velikim kolekcijama podataka na *web*u, a njihovi potencijali neiskorišćeni. Na osnovu izrudarenih paterna, preduzeća mogu preduzimati određene korake ka unapređenju poslovanja, te je značajno da postoji celovit i sistematičan prikaz široke palete diversifikovanih metoda koji se u pojedinim pristupima rudarenju *web* podataka koriste. U radu su prikazani različiti aspekti rudarenja *web*-a i mapirani uz različite metodološke pristupe: uz rudarenje sadržaja *web*-a se vezuju metode klasterovanja, sistemi za pronalaženje informacija i kolaborativno filtriranje; rudarenje strukture *web*-a je povezano sa metodom rangiranja relevantnih stranica u kontekstu hiperlinkovske strukture *web*-a; uz rudarenje upotrebe *web*-a je vezan metod otkrivanja asocijativnih pravila.

Nadalje, dat je pregled specifičnih zadataka koje pojedini aspekti rudarenja *web*-a rešavaju i opisani su tipovi (sa)znanja koji se mogu derivirati. Na osnovu ovakvih informacija i saznanja, preduzeće može bolje planirati poslovne strategije i ostvariti bržu ekspanziju. Istaknute su mnogobrojne mogućnosti i prednosti koje rudarenje *web* podataka pruža u različitim sferama poslovanja u odnosu na druge analitičke pristupe: dublja analiza socijalnih i društvenih procesa, merenje statusa ili uticaja (dokumenata, pojedinaca i dr. entiteta), kategorisanje *web* stranica, uvid u hijerarhiju *web* strana, analiza *web* saobraćaja, otkrivanje učestalih navigacionih puteva kroz *web* sajt preduzeća, procena efikasnosti sajta i poboljšanje upotrebljivosti *web* strana, uvid u interakciju korisnika i sajta radi personalizacije *web* sajtova/kataloga proizvoda, preporučivanje proizvoda/usluga, profilisanje korisnika za bolje upravljanje odnosima sa komitentima, otkrivanje naprednih informacija o potražnji, značajnih za

snabdevanje, planiranje zaliha i proizvodnju, poboljšanje performansi *web* servera i njihovih aplikacija. Dosadašnja praksa je pokazala, a o tome svedoče i odabrani primeri dobre prakse navedeni u radu, da svaki aspekt upotrebe inteligentnih tehnika doprinosi poboljšanom poslovanju, a da se kvalitet postignutih rezultata značajno poboljšava kombinacijom više metoda (Zaiane, Li & Hayward, 2004). Sve ovo govori u prilog polaznoj pretpostavci da rudarenje podataka ima heterogenu primenu, koja donosi benefite u mnogim sferama poslovanja. Pa ipak, pojedine inteligentne tehnologije, uključujući i rudarenje *web*-a, još uvek su u ranoj fazi prihvatanja, čiju širu zastupljenost možemo očekivati tek kada više *software*-skih provajdera bude uključilo dostignuća mašinskog učenja u svoja rešenja, odnosno, kada se proširi ponuda specijalizovanih alata za ovu namenu. Do tada, rudarenje *web* podataka će predstavljati konkurentsku prednost i preduzeća koja su preduzela aktivnosti oko njegove primene, ili već postižu određene rezultate, nerado dele svoja iskustva sa širom javnošću. Zbog toga se u referentnim izvorima može naći samo manji broj primera dobre prakse, i to pretežno iz domena e-komerca.

U oblasti veštačke inteligencije se neprekidno razvijaju različiti pristupi analizi podataka i proširuje paleta mogućih aplikacija. Na primer, Tableau i VoiceBase (Tableau-VoiceBase, 2019), su zajednički razvili alat za analizu izgovorenog teksta, koji omogućava da se zvučni zapisi kontakt centara zapišu bez transkribovanja, vizualizuju i pruže uvid u nestrukturirane telefonske razgovore i bazu korsnika (Sevilla, 2019). Na ovaj način će službe marketinga, prodaje, proizvodnje i dr. moći poboljšati odlučivanje i preduzeti adekvatne akcije. Obzirom da rad daje presek trenutnog stanja u oblasti rudarenja *web* podataka, ovakve i slične inovativne aplikacije nisu njime obuhvaćene. Isto tako, nove klase podataka, generisane kroz mobilne aplikacije, društvene medije, senzore, mobilne uređaje, su segment na koji će se preduzeća sve više oslanjati u operativnom radu. Razumevanje upotrebe ovih mnogostrukih kanala i kreiranje analitičkih mogućnosti njihove obrade u cilju boljeg *online* i *offline* poslovanja su predmet budućih istraživanja.

REFERENCE

AstraZeneca, Co. (2019). *AstraZeneca: Early Hypotheses Testing Through Linked Data*. Retrieved Jun 6, 2019, from https://www.ontotext.com/knowledgehub/case-studies/causality-mining-pharma/

Babu, G. P., & Mehtre, B. M. (1995). Color indexing for efficient image retrieval. *Multimedia Tools and applications, 1*(4), 327-348.

Cheng, G., Healey, M. J., McHugh, J. A. M., & Wang, J. T. L. (2001). *Mining the World Wide Web - An Information Search Approach*. Dordrecht, The Netherlands: Kluwer Acaddemic Publishers.

Dudovskiy, J. (2017). *IKEA Segmentation, Targeting and Positioning: Targeting Cost-Conscious Customers.* Retrieved Jun 6, 2019, from https://research-methodology.net/ikea-segmentation-targeting-positioning-targeting-cost-conscious-customers/

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics: quantitative methods in library, documentation and information science*. Amsterdam, The Netherlands: Elsevier Science Publishers.

Fayyad, U. M., Djorgovski, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.). *Advances in Knowledge Discovery and Data Mining*, (pp. 471-493). Menlo Park, California: AAAI Press.

Fan, G. L., Liu, Y. W., Tong, J. Q., Zhao, S. H., & Nie, Z. Q. (2016). Application of K-means algorithm to web text mining based on average density optimization. *Journal of Digital Information Management*, *14*(1), 41-46.

FeaturedCustomers, Co. (2019). *Vendor Directory*. Retrieved Jun 12, 2019, from https://www.featuredcustomers.com/vendors

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471-479. doi:10.1126/science.178.4060.471

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, 562-569, ACM. doi:10.1145/511446.511520

Google. (2019). *Google analythics - Users (new, returning, unique) explained in great detail*. Retrieved Jun 12, 2019, from https://www.optimizesmart.com/understanding-users-in-google-analytics

Guandong, X., Yanchun, Z., & Lin, L. (2011). *Web mining and social networking: Techniques and applications*. Berlin, Germany: Springer

Harapiak, C. (2013). IKEA's International Expansion. *International Journal of Business Knowledge and Innovation in Practice*, *1*(1).

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). *An algorithmic framework for performing collaborative filtering.* In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 230-237), Berkeley, California. doi:10.1145/312624.312682

Hilderman, R., & Hamilton, H. J. (2013). *Knowledge Discovery and Measures of Interest*. Berlin, Germany: Springer Science & Business Media.

Huang, T., & Mieghem, J. A. V. (2014). Clickstream data and inventory management: Model and empirical analysis. *Production and Operations Management*, *23*(3), 333-347. doi.org/10.1111/poms.12046

IBM Analythics. (2016). *Analytics solutions unified method - Implementations with Agile principles*. Retrieved Jun 12, 2019, from ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf

Jiawei, H., & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco, SF: Morgan Kaufman Publishers.

Jokar, N., Honarvar, A. R., Hamirzadeh, S. A., & Esfandiari, K. (2016). Web mining and web usage mining techniques. *Bulletin de la Société des Sciences de Liège*, *85*, 321-328.

Kim, I., & Kuljis, J. (2010). Applying content analysis to web-based content. *Journal of Computing and Information Technology*, *18*(4), 369-375. doi:10.2498/cit.1001924

Kokkoras, F., Jiang, H., Vlahavas, I., Elmagarmid, A. K., Houstis, E. N., & Aref, W. G. (2002). Smart videotext: A video data model based on conceptual graphs. *ACM Multimedia Systems*, *8*(4), 328-338. doi.org/10.1007/s005300200

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. London, UK: Sage Publications.

Kumar, P. B. C., & Mohamed, D. (2018). Two-stage information filters for single and multiple sensors, and their square-root versions. *Automatica*, 98, 20-27. doi:10.1016/j.automatica.2018.09.001

Li, M., & Feng, C. (2010). *Overview of Web mining technology and its application in e-commerce*. In 2nd International Conference on Computer Engineering and Technology (pp. 277-280), IEEE Explore Digital Library. doi:10.1109/ICCET.2010.5485404

Leavitt, N. (2002). Let's Hear It for Audio Mining. *IEEE Computer Magazine on Technology News, 35*(10), 23-25. doi:10.1109/mc.2002.1039511

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

Ma, Y. F., Lu, L., Zhang, H. J., & Li, M. (2002). *A user attention model for video summarization*. In: Proceedings of the tenth ACM international conference on multimedia (pp. 533-542). doi:10.1145/641007.641116

Madhumathi, K., & Selvadoss Thanamani, A. (2014, March). Image mining: Frameworks and techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 2,* (Special Issue 1), Proceedings of International Conference on Global Innovations in Computing Technology (ICGICT'14), Tirupur, Tamilnadu, India.

Markov, Z., & Larose, T. D. (2007). *Data Mining the Web - Uncovering Patterns in Web Content, Structure, and Usage*. Chichester, UK: John Wiley and Sons.

Mozenda, Co. (2018). *How To Scale Your Web Content Harvesting Operation*. Retrieved Jun 6, 2019, from https://www.mozenda.com/scale-web-content-harvesting-operation/

Palau, J., Montaner, M., Lopez, B., & de la Rosa, J. L. (2016). Collaboration analysis in the recommender system using social networks. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(5), 137-151.

Petkovic, M., & Jonker, W. (2001). *Content-based retrieval of spatio-temporal video events*. In: Proceedings of multimedia computing and information management track of IRMA international conference. Retrieved Jun 6, 2019, from https://pdfs.semanticscholar.org/f2e0/f7557452c19e737a873b7444ca3e61e2b7fa.pdf

RapidMiner. (2019a). *Improving Customer Service with Text Mining and Auto-classification.* Retrieved Jun 6, 2019, from https://rapidminer.com/resource/customer-service-auto-classification/

RapidMiner. (2019b). *Optimization of Customer Support for Mobilkom Austria*. Retrieved Jun 6, 2019, from https://rapidminer.com/resource/optimization-customer-support-mobilkom-austria

RapidMiner. (2019c). *Sentiment Analysis at PayPal Using RapidMiner.* Retrieved Jun 6, 2019, from https://rapidminer.com/resource/sentiment-analysis-paypal/

Resnick, P., & Varian, H. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Rui, Y., & Huang, T. D. (2000). A unified framework for video summarization. *Browsing and retrieval, image and video processing handbook*, 705-715.

Search Engine Land. (2017). *How Google measures the authority of web pages*. Retrieved January 23, 2019, from https://searchengineland.com/google-authority-metric-274231

Sevilla, C. G. (2019). *VoiceBase and Tableau Deliver New Insights through Speech Analytics*. Retrieved January 23, 2019, from https://www.pcmag.com/article/367331/voicebase-and-tableau-deliver-new-insights-through-speech-an

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13-22.

Shukla, R., Silakari, S., & Chande, P. K. (2013). Web Personalization Systems and Web Usage Mining: A Review. *International Journal of Computer Applications, 72*(21), 6-13. doi:10.5120/12664-9264

Siddiqui, A. T., & Aljahdali, S. (2013). Web Mining Techniques in e-commerce Applications. *International Journal of Computer Applications, 69*(8), 39-43. doi:10.5120/11864-7648

Splunk. (2013). *Splunk for Digital Intelligence.* Retrieved February 4, 2019, from https://www.splunk.com/web_assets/pdfs/secure/Splunk_for_Digital_Intelligence.pdf

Tableau-VoiceBase, Inc. (2019). *Leverage VoiceBase's Open Data Architecture to Visualize AI Powered Speech Analythics in Tableau*. Retrieved Jun 6, 2019, from https://www.voicebase.com/tableau-speech-analytics/

Tan, P-N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems, 29*(4), 293-313. doi.org/10.1016/S0306-4379(03)00072-3

Thomas, D. (2002). *Data mining boosts drug research at AstraZeneca.* Retrieved Jun 6, 2019, from https://www.computerweekly.com/news/2240046811/Data-mining-boosts-drug-research-at-astraZeneca.

Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International Journal of Multimedia Information Retrieval*, *1*(3), 153-172. doi:10.1007/s13735-012-0016-2

Zaiane, O. R., Li, J., & Hayward, R. (2004). *Mission-Based Navigational Behaviour Modeling for Web Recommender Systems*. In International Workshop on Knowledge Discovery on the Web (WebKDD 2004), Advances in Web Mining and Web Usage Analysis (pp. 37-55).

Yoneki, E., Tirado, J. M., Guo, Q., & Serban, O. (2016). MAKI: Tools for web data knowledge extraction. *Technical report UCAM-CL-TR-881*, Cambridge, United Kingdom: University of Cambridge Computer Laboratory.

Primljeno 11. april 2019,
nakon revizije,
prihvaćeno za publikovanje 20. avgusta 2019.
Elektronska verzija objavljena 23. avgusta 2019.

Zita Bošnjak je redovni profesor na Ekonomskom fakultetu u Subotici, Univerzitet u Novom Sadu. Zvanje doktora informatičkih nauka stekla je na Ekonomskom fakultetu u Subotici. Oblasti njenog naučno-istraživačkog interesovanja su: inteligentni sistemi, upravljanje znanjem, i inteligentne analize podataka.

Olivera Grljević je docent na Ekonomskom fakultetu u Subotici, Univerzitet u Novom Sadu. Oblasti njenih istraživanja su: sentiment analiza, *text mining* i *data mining.*

Saša Bošnjak je redovni profesor na Ekonomskom fakultetu u Subotici, Univerzitet u Novom Sadu. Zvanje doktora informatičkih nauka stekao je na Ekonomskom fakultetu u Subotici. Oblasti njegovog istraživačkog interesovanja su: baze podataka, metode razvoja *software*-a i internet tehnologije.