UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

Uporaba rudarjenja po spletu za učinkovito poosebljanje in načrtovanje spletišč

Ljubljana, junij 2007 BOŠTJAN KOŽUH

IZJAVA

Študent Boštjan Kožuh izjavljam, da sem avtor tega magistrskega dela, ki sem ga napisal pod mentorstvom prof. dr. Jurija Jakliča, in skladno s 1. odstavkom 21. člena Zakona o avtorskih in sorodnih pravicah dovolim objavo magistrskega dela na fakultetnih spletnih straneh.

V Ljubljani, dne 14.6.2007

Podpis:

KAZALO

Kazalo slik

Kazalo tabel

1 UVOD

1.1 OPREDELITEV PROBLEMATIKE

Organizacije po svetu bržkone niso še nikoli doživljale tako korenitih sprememb kot jih zdaj. Internet in svetovni splet sta povečala pritisk po globalni navzočnosti in narekujeta prenos moči z lastnikov kapitala in nepremičnin na lastnike idej, informacij in znanja; pretok informacij je postal hiter in nemoten, organizacije se decentralizirajo, hierarhija se plošči, učinkovitost in odzivnost pa se povečujeta. Svetovni splet je močno spremenil naš koncept komunikacije in medsebojnega vplivanja, saj je postal eden izmed najpomembnejših kanalov za komunikacijo organizacij s strankami; te namreč svetovni splet vedno pogosteje uporabljajo za iskanje informacij o novih izdelkih in storitvah, zato morajo tudi organizacije najti ustrezne načine, da uporabnike privabijo na spletišče, jim predstavijo svoje sporočilo in, kar je najpomembnejše, da jih prepričajo, da izvedejo želena transakcijska dejanja in se spet vrnejo.

Svetovni splet še vedno raste s presenetljivo hitrostjo, tako z vidika števila obiskov na spletnih straneh kot tudi z vidika velikosti in zapletenosti spletišč. Ker do svetovnega spleta dostopajo vse boj heterogeni in globalno razporejeni uporabniki z različnimi interesi in cilji in ker ti za dostop uporabljajo različne naprave in vmesnike, postaja oblikovanje spletnega mesta, upravljanje spletnega strežnika in zagotavljanje enostavnosti pri navigaciji skozi spletno mesto vse bolj zapletena naloga. Ker te težave vplivajo na strateško uspešnost številnih organizacij, je bilo v izboljševanje spletne uporabnosti in razvoj tehnik za načrtovanje spletnih strani vloženih veliko raziskovalnih naporov, vendar pa so številna spletišča še vedno preveč kompleksna, tako da številnim uporabnikom ne uspe uspešno dokončati zastavljenih nalog, mnogi pa se prav zaradi težav na spletišče tudi več ne vrnejo.

Vse bolj se kaže, da je zaradi velikega števila spletnih strani in (pre)obilja informacij eden najučinkovitejših pristopov, ki jih lahko podjetja uporabijo za poglabljanje odnosa s svojimi strankami, spletno pooosebljanje. Spletno pooosebljanje je proces spreminjanja funkcionalnosti, videza, strukture in vsebine spletnih strani z namenom, da bi posameznemu uporabniku povečali relevantnost predstavljene vsebine (Markellou, Rigou, Sirmakessis, 2005, str. 30) in mu olajšali doseganje zastavljenih ciljev. Organizacije lahko spletno pooosebljanje omogočajo na različne načine; po začetnih poskusih spletnega pooosebljanja z različnimi tehnologijami se je pokazalo, da je najprimernejše t. i. opazovalno pooosebljanje, ki temelji na predpostavki, da lahko namige za uspešno pooosebljanje najdemo kar v preteklem navigacijskem vedenju obiskovalcev.

Količina podatkov, ki jih je treba analizirati pri opazovalnem poosebljanju, tako velika, da je ni mogoče obvladati brez močnih orodij in metod, hkrati pa tudi klasične metode preiskovanja in analiziranja podatkov niso več uporabne. Zato Srivastava et. al. (2000) za ta namen predlagajo uporabo rudarjenja po spletu, natančneje rudarjenja po podatkih o uporabi spleta. Rudarjenje po spletu je posebna aplikacija tehnik podatkovnega rudarjenja in ga lahko opredelimo kot (pol)samodejni proces raziskovanja in analiziranja velikih količin podatkov iz spletnih virov, z namenom, da bi odkrili uporabne in zanimive informacije, vzorce in povezave (Woon, Ng, Lim, 2005, str. 374), rudarjenje po podatkih o uporabi spleta pa je posebna vrsta rudarjenja po spletu, ki je usmerjena k odkrivanju znanja o ljudeh, ki splet uporabljajo, njihovih interesih in pričakovanjih, težavah, s katerimi se spopadajo, in njihovih zahtevah (Galeas, 2005). Grobo rečeno lahko celoten proces spletnega poosebljanja na osnovi rudarjenja po spletu razdelimo na štiri osnovne faze: zbiranje podatkov, priprava podatkov, odkrivanje vzorcev uporabe ter uporaba odkritih vzorcev za izvedbo poosebljanja. Proces je večdisciplinaren in kompleksen ter od sodelujočih zahteva temeljito poznavanje spletnih tehnologij in metodologije podatkovnega rudarjenja, pa tudi problematike izbranega spletnega mesta in njegove vrednosti za obiskovalce. Glede na vse navedene lastnosti se tega procesa ne da podpreti z enim samim orodjem, prav tako pa tudi ni enostavnega recepta in zagotovila za njegovo uspešno izvedbo.

Spletno poosebljanje pomeni priložnost tako za lastnike spletišč kot za uporabnike. Uporabniki po eni strani cenijo boljšo odzivnost, ki je rezultat vnaprejšnjih kalkulacij in nalaganja vsebin v lokalni predpomnilnik že pred izvedbo zahteve, hkrati pa se s prilagodljivimi in poosebljenimi spletišči močno poveča njihova uporabnost za uporabnika, kar poveča verjetnost uspešnega izvajanja zastavljenih nalog in vodi v izboljšanje uporabniške izkušnje. Podobno lahko veliko pridobijo tudi ponudniki vsebin: po eni strani lahko koristno uporabijo spoznanja, ki jih prinese uporaba modeliranja, po drugi strani pa so pomembni kratkoročni in dolgoročni finančni učinki, ki jih prinesejo zadovoljni uporabniki (Davison, 2004, str. 435). Poleg tega je organizacijam pri doseganju njihove poslovne strategije lahko v veliko pomoč tudi možnost usmerjanja obiskovalca po spletišču, saj se s tem lahko močno poveča verjetnost, da bodo uporabniki na posamezni strani izvedli tista dejanja, ki so želena z vidika posameznega poslovnega modela. Pri tem je zelo pomembno, da se zavedamo, da je treba pred vpeljavo rešitev za poosebljanje natančno definirati, katere oblike poosebljanja so v določenem primeru smiselne in potrebne, da bi uporabnikom omogočali lažje doseganje njihovih ciljev in izpolnjevanje njihovih nalog; pri tem ne smemo prezreti niti njihovih morebitnih težav in pomislekov, še posebej na področju zasebnosti in varovanja podatkov.

1.2 NAMEN IN CILJ MAGISTRSKEGA DELA

Namen dela je analizirati koristi uporabe rudarjenja po spletu za potrebe spletnega poosebljanja, odgovoriti na vprašanje, kdaj in zakaj naj ga podjetja in organizacije

uporabljajo, prikazati učinkovite pristope k spletnemu poosebljanju z uporabo omenjenega tipa rudarjenja ter identificirati težave, s katerimi se organizacije srečujejo pri procesu. Cilj dela je zato prepoznati ključne naloge v procesu odkrivanja znanja o uporabnikih spleta in pokazati na učinkovito arhitekturno rešitev vključevanja rudarjenja po spletu v poosebljanje in prilagajanje spletišč, s tem pa pripraviti osnovo, ki bi organizacijam pomagala odgovoriti na naslednja vprašanja:

- Kdaj in čemu se lotiti uvajanja spletnega poosebljanja?

- Kako poosebljati in prilagajati spletišče z vidika vsebine, strukture in iskanja z namenom povečevanja poslovne vrednosti spletišča?

- Kako izsledke rudarjenja po spletu uporabiti za boljše in učinkovitejše delovanje spletišča?

1.3 METODE PREUČEVANJA IN ZASNOVA DELA

Pri izdelavi magistrskega dela se bom oprl na preučevanje in analizo teoretične podlage s področja podatkovnega rudarjenja, rudarjenja po spletu in spletnega poosebljanja ter analizo raziskovalnih projektov in praktičnih primerov uvajanja in izvajanja posameznih nalog v procesu spletnega poosebljanja z rudarjenjem po spletu. Dognanja na tem področju so razmeroma nova in v mnogih pogledih še v nastajanju, zato bom snov pretežno črpal iz člankov, objavljenih v strokovnih revijah, v zbornikih konferenc in simpozijev, pa tudi iz člankov, zbranih v preglednih monografijah, in člankov, objavljenih na spletnih straneh najvidnejših strokovnjakov s tega področja.

Sistemi za spletno poosebljanje z uporabo rudarjenja po spletu se razvijajo predvsem v okviru raziskovalnih projektov, katerih izsledke pa je pogosto težko prenesti v prakso. Kjer bo mogoče in smiselno, bom poskušal združiti različne teoretične poglede na problematiko in najti pristop, ki je za podjetja ob tehnologiji in znanju, ki sta jim na voljo, najbolj uresničljiv in koristen. Pri tem bom uporabil tudi lastno znanje in izkušnje z načrtovanjem spletnih rešitev in podatkovnim rudarjenjem, pridobljeno med magistrskim študijem, na različnih tečajih in z delom pri različnih projektih. Izkušnje in pridobljena znanja bom nadgradil z informacijami, ki sem jih pridobil v razgovorih z osebami, ki so odgovorne za rudarjenje po spletu in razvoj poosebljenih rešitev v nekaterih vodilnih podjetjih na tem področju. Zbrana spoznanja bom na koncu uporabil pri izdelavi in kritičnem ovrednotenju projekta rudarjenja po spletu za poosebljanje spletišča Statističnega urada Republike Slovenije, hkrati pa bom tudi nakazal, katere so največje prepreke in priložnosti za uporabo rudarjenja po spletu za namene spletnega poosebljanja.

Magistrsko delo je razdeljeno na šest poglavij. Po uvodnem poglavju, v katerem sta opredeljena problematika in namen dela, nadaljujem z opredelitvijo rudarjenja po podatkih kot procesa odkrivanja znanja v podatkih, opišem kratko zgodovino razvoja procesa, njegovo metodologijo in večdisciplinarnost ter prikažem taksonomijo rudarjenja po podatkih s poudarkom na rudarjenju po spletu. Zaradi trenutne neusklajenosti ali neobstoja strokovnih izrazov za poimenovanje različnih vrst rudarjenja po podatkih, pozornost namenim tudi iskanju ustreznih slovenskih prevodov.

Tretje poglavje je namenjeno natančni predstavitvi procesa rudarjenja po podatkih o uporabi spleta, njegovega namena in težav, s katerimi se je treba spopasti, da proces uspešno izvedemo. Opišem posamezne faze v procesu in skušam v množici različnih pristopov najti tiste, ki so se izkazali za najučinkovitejše in najprimernejše za uporabo tudi zunaj raziskovalnih okolij. Poglavje zaključim s prikazom analize znanja, pridobljenega v procesu rudarjenja, in nakažem njegove možne aplikacije.

V četrtem poglavju se posvetim analizi spletnega poosebljanja, ki je po mnenju mnogih najperspektivnejša aplikacija rudarjenja po podatkih o uporabi spleta. Ugotavljam, kaj spletno poosebljanje sploh je, kakšen je potencial spletnega poosebljanja v organizacijah ter kdaj in kako ga je smiselno uporabiti. Nato nadaljujem z umeščanjem rudarjenja po podatkih o uporabi spleta v proces spletnega poosebljanja in opis nadgradim z analizo primernosti in postopkov vključevanja semantičnega znanja in drugih oblik rudarjenja po spletu v spletno poosebljanje. Poglavje zaokrožim s predstavitvijo zahtev in pogojev za razvoj integrirane arhitekturne rešitve poosebljanja spletišča z uporabo rudarjenja po spletu.

Peto poglavje je namenjeno uporabi teoretičnih spoznanj za izvedbo spletnega poosebljanje na spletišču Statističnega urada RS ter analizi potenciala rudarjenja po spletu in spletnega poosebljanja. Prikažem, katere oblike poosebljanja bi bilo smiselno uvesti in na kratko opišem način njihove izvedbe. Poleg tega se v skladu s formalno definicijo procesa lotim izdelave preizkusnega projekta spletnega poosebljanja z uporabo priporočanja vsebine in podrobno opišem izvedbo posameznih nalog ter utemeljim način njihove izvedbe. Poglavje končam z analizo dobljenih rezultatov in primernosti predstavljene oblike poosebljanja na obravnavanem spletišču ter analizo koristnosti spletnega poosebljanja na splošno.

V zadnjem, šestem poglavju, povzamem sklepne misli in nakažem nadaljnje smeri razvoja spletnega poosebljanja in rudarjenja po spletu.

2 ODKRIVANJE ZNANJA V PODATKIH

Tradicionalna metoda preoblikovanja podatkov v znanje se zanaša na ročno analizo in interpretacijo enega ali več analitikov, ki se temeljito seznanijo s podatki, nato pa nastopajo kot posredniki med podatki ter njihovimi uporabniki in izdelki. Tak način raziskovanja podatkov je počasen, drag in močno subjektiven, z dramatično rastjo količine podatkov pa na številnih področjih tudi povsem nepraktičen (Fayyad, Piatetsky-Shapiro, Smyth, 1996, str. 38). Ker so nam prav računalniki omogočili zbirati več podatkov, kot jih lahko pregledamo, je povsem logično, da tudi pri odkrivanju pomembnih vzorcev in struktur iz velikih količin podatkov uporabimo njihove računske tehnike. Tak način se uporablja tudi na področju odkrivanja znanja v podatkih, ki je v zadnjih letih postal eden izmed najbolj razširjenih pristopov k reševanju problema preobilja podatkov.

2.1 OPREDELITEV ODKRIVANJA ZNANJA V PODATKIH

Na abstraktni ravni se področje odkrivanja znanja v podatkih ukvarja z razvojem metod in tehnik za razumevanje podatkov. Osnovna težava, ki jo skuša rešiti proces odkrivanja znanja, je, kako podrobne podatke, ki jih je običajno preveč, da bi jih lahko pregledali ali razumeli, preoblikovati v drugačno obliko, ki je bodisi kompaktnejša (npr. kratko poročilo), abstraktnejša (npr. model procesa, v katerem so podatki nastali) ali pa uporabnejša (npr. napovedovalni model za ocenjevanje vrednosti bodočih primerov) (Fayyad, Piatetsky-Shapiro, Smyth, 1996, str. 37).

Natančna opredelitev odkrivanja znanja v podatkih ni enostavna naloga, zato se ne gre čuditi, da splošno sprejeta definicija ne obstaja. V literaturi se najpogosteje navaja, da je odkrivanje znanja v podatkih *»netrivialni proces odkrivanja veljavnih, novih, uporabnih in v končni fazi razumljivih vzorcev v podatkih«* (Fayyad, Piatetsky-Shapiro, Smyth, 1996, str. 40). Klösgen in Żytkow (Klösgen, Żytkow, 2002) taki definiciji nasprotujeta in pravita, da odkrivanje znanja v podatkih ne potrebuje posebne definicije, da proces ni vedno netrivialen in da v podatkih pravzaprav ne iščemo vzorcev, ampak znanje o področju, ki ga podatki predstavljajo. Trdita tudi, da je v končni fazi vsako pridobljeno znanje novo in uporabno, s čimer se sicer lahko strinjamo, vseeno pa je bistvo procesa odkrivanje informacij, vzorcev in povezav, ki niso očitni ali so celo neintuitivni. Vsekakor pa se proces odkrivanja znanja v podatkih v smislu, kot ga bom obravnaval v tem delu, razlikuje od splošnega procesa odkrivanja znanja vsaj glede uporabe informacijske tehnologije in njegovega namena. V tej zvezi se zdi najprimernejša definicija, ki pravi, da je odkrivanje znanja v podatkih *»netrivialen samodejni ali polsamodejni proces raziskovanja in analiziranja velikih količin podatkov z namenom odkrivanja veljavnih vzorcev in povezav, ki organizaciji pomagajo pri boljšem odločanju«* (Berry, Linoff, 1997, str. 5).

Zgodovinsko gledano poznamo za označevanje procesa odkrivanja uporabnih vzorcev v podatkih več izrazov – npr. podatkovno rudarjenje (ang. data mining – DM), izkopavanje znanja (ang. knowledge extraction), odkrivanje informacij (ang. information discovery), žetev informacij (ang. information harvesting), obdelava podatkovnih vzorcev (ang. data pattern processing) ipd. Najbolj se je uveljavil izraz podatkovno rudarjenje, ki izhaja iz področja statistike, vendar pa je prispodoba rudarjenja nekoliko zavajajoča in ni najboljša; pri »tradicionalnem« rudarjenju namreč izkopavamo rudo, ne pa zemlje, iz katere se le-ta izloča. Boljša analogija je *izkopavanje znanja* (ang. knowledge mining), saj je novo znanje pričakovani rezultat procesa, podobno kot sta to pri »tradicionalnem« rudarjenju ruda ali zlato. V tem smislu je bil leta 1989 na prvi delavnici o odkrivanju znanja (Workshop on Knowledge Discovery in Databases, Detroit, ZDA) kot najprimernejši predlagan izraz *odkrivanje znanja v zbirkah podatkov* (ang. knowledge discovery in databases – KDD).

Kot sem že omenil, se najpogosteje uporablja izraz podatkovno rudarjenje, vendar pa njegov pomen ni enotno priznan. Nekateri avtorji (npr. Fayyad, Piatetsky-Shapiro, Smyth, 1996) izraz uporabljajo izključno v ožjem pomenu kot poimenovanje tiste faze v procesu odkrivanja znanja v podatkih, v kateri s pomočjo izbranih algoritmov dejansko odkrivamo znanje, pogosteje pa se podatkovno rudarjenje razume kar kot sopomenka za odkrivanje znanja v podatkih (npr. Cabena et. al., 1997; Chapman et al., 1999; Klösgen, Żytkow, 2002). Nasprotno sta Berry in Linoff sprva zagovarjala razlago, da je odkrivanje znanja v podatkih sopomenka za tisto obliko podatkovnega rudarjenja, pri katerem se uporablja pristop »od spodaj navzgor«[1] (ang. bottom-up), tj. pristop, pri katerem vnaprej ne postavljamo nobenih predpostavk (Berry, Linoff, 1997), vendar pa sta pozneje svoje stališče spremenila in sedaj razlikujeta usmerjeno podatkovno rudarjenje, pri katerem vemo, kaj iščemo, in neusmerjeno podatkovno rudarjenje, pri katerem podatkom pustimo, da govorijo sami zase (Berry, Linoff, 2000).

V zadnjem času se je pokazalo, da sta izraza »podatkovno rudarjenje« in »odkrivanje znanja v zbirkah podatkov« preveč splošna in da potrebujemo izraze, ki bi natančneje opredeljevali tudi tip in obliko podatkov, v katerih iščemo znanje. Proces odkrivanja znanja v podatkih se namreč ne usmerja le na podatke, shranjene v strukturiranih zbirkah podatkov[2] (ang. database) ali skladiščih podatkov[3] (ang. data warehouse), ampak na primer tudi na nestrukturirane ali polstrukturirane podatke, ki so lahko shranjeni v dokumentih ali pa so dostopni prek svetovnega spleta. Eden pomembnih razlogov za ločevanje izrazov je tudi dejstvo, da tip in oblika podatka pomembno določata tudi postopke v posameznih fazah rudarjenja in zahtevata posebej prilagojene algoritme za odkrivanje znanja (v nadaljevanju

---

1 Obraten je pristop »od zgoraj navzdol« (angl. top-down), pri katerem v začetku procesa postavimo določene hipoteze in kasneje preverjamo njihovo veljavnost.

2 Zbirka podatkov je urejena zbirka medsebojno povezanih podatkov, ki je shranjena na nosilcu podatkov.

3 Podatkovno skladišče je zbirka podatkov, ki je namenjena odločanju na podlagi predhodne analize.

algoritme)[4].

V magistrskem delu bom zato za označevanje rudarjenja po podatkih, shranjenih v strukturiranih zbirkah, uporabljal kot sopomenki izraza *rudarjenje po strukturiranih podatkih* in *odkrivanje znanja v strukturiranih podatkih*. Za odkrivanje znanja v dokumentih bom uporabljal izraz *rudarjenje po besedilih* (ang. text mining – TM, knowledge discovery in text – KDT), za odkrivanje znanja v spletu pa *rudarjenje po spletu* (ang. web mining). Za označevanje procesa rudarjenja na splošno, v smislu pomena procesa *odkrivanja znanja v podatkih*[5], bom uporabljal tudi izraza *podatkovno rudarjenje*[6] in *rudarjenje po podatkih*.

2.1.1 Proces odkrivanja znanja v podatkih

Prve raziskave v okviru odkrivanja znanja in rudarjenja po podatkih so se večinoma omejevale zgolj na algoritme in tehnike. Kot sem že omenil, pa trenutno prevladuje prepričanje, da odkrivanje znanja ni le izbor pravih metod, ampak da gre za kompleksen, ponavljajoč in interaktiven proces.

Definicija odkrivanja znanja v podatkih je sicer dovolj splošna, da zajame večino povezanih aktivnosti, vendar pa ne ponuja natančnejše opredelitve posameznih korakov v procesu niti nasvetov za uspešno vodenje projektov za odkrivanja znanja. Iz literature poznamo veliko poskusov opisa procesa odkrivanja znanja v podatkih (npr. Fayyad, Piatetsky-Shapiro, Smyth, 1996; Cabena et al., 1997; Berry, Linoff, 1997; Klösgen, Żytkow, 2002a; Eirinaki, Vazirgiannis, 2003), ki v splošnem sicer sledijo enakim načelom, vendar pa se do določene mere razlikujejo v razumevanju in obravnavanju posameznih delov procesa ali pa posameznih korakov zaradi drugačnega razumevanja definicije sploh ne poznajo (npr. Fayyad, Piatetsky-Shapiro, Smyth, 1996; Eirinaki, Vazirgiannis, 2003).

Za zagotavljanje sistematičnosti in primerljivosti je nujno poenotenje pogleda na odkrivanje znanja v podatkih. Od različnih obstoječih metodologij na tem področju predstavlja najcelovitejši pristop projekt CRISP-DM[7] (CRoss-Industry Standard Process for Data Mining, v nadaljevanju CRISP-DM); v okviru tega projekta je bil razvit od orodij in industrijskih sektorjev neodvisen procesni model podatkovnega rudarjenja, ki po eni strani

---

[4] Algoritem za odkrivanje znanja (tudi algoritem podatkovnega rudarjenja) je opis končnega zaporedja dejanj in opravil, s katerim lahko v podatkih odkrivamo znanje in ga lahko zapišemo v nekem programskem jeziku.

[5] Primeren angleški izraz bi bil *knowledge discovery in data*.

[6] V angleški literaturi se izraz *data mining* uporablja tako za označevanje celotnega procesa podatkovnega rudarjenja na splošno kot tudi za rudarjenje po podatkih, shranjenih v strukturiranih zbirkah podatkov. Dodatno zmedo v razumevanje izraza vnaša še uporaba istega izraza za označevanje dela procesa, v katerem dejansko pride do odkrivanja znanja.

[7] Projekt CRISP-DM je nastal konec leta 1996 na pobudo podjetij DaimlerChrysler (takrat Daimler-Benz), SPSS (takrat ISL) in NCR, ki so imela že precej izkušenj s podatkovnim rudarjenjem. Projekt je delno financirala tudi Evropska komisija (program ESPRIT).

definira standardno metodologijo, po drugi pa predlaga tudi osnovne smernice za izvajanje projektov podatkovnega rudarjenja. Nekoliko drugačen pristop predstavlja metodologija SEMMA[8] družbe SAS; ta sicer ne ponuja celovitega metodološkega ogrodja, ampak je bolj logična organizacija nabora funkcij in orodij, s katerimi se v programskih rešitvah omenjene družbe izvajajo osnovne naloge podatkovnega rudarjenja (SAS, 2006). Po anketi družbe KDnuggets (Data Mining Methodology Poll, 2004) se omenjeni metodologiji uporabljata v 52 odstotkih projektov podatkovnega rudarjenja; v 34 odstotkih projektov organizacije uporabljajo metodologijo, ki so jo razvile same, v 6 odstotkih projektov organizacije uporabljajo druge metodologije, v 7 odstotkih projektov pa se ne uporablja nobena metodologija.

Metodologija CRISP-DM je predstavljena kot hierarhični procesni model, ki ga sestavljajo naloge na štirih ravneh abstrakcije (Chapman et al., 1999, str. 9). Na najvišji ravni je proces organiziran po stopnjah (ang. phase), znotraj katerih so na drugi ravni definirane splošne naloge (ang. generic task). Druga raven je splošna raven, saj mora po eni strani zajeti vse možne situacije podatkovnega rudarjenja, po drugi strani pa ostati celovita[9] in stabilna[10]. Na tretji ravni najdemo specializirane naloge (ang. specialized task), v katerih je določeno, kako naj se splošne naloge izvajajo v posameznih okoliščinah (npr. glede na tip podatkov, po katerih rudarimo). Na spodnji ravni so procesni primeri (ang. process instance) kot zapisi dejanj, odločitev in rezultatov dejanskega podatkovnega rudarjenja.

Življenjski krog projekta podatkovnega rudarjenja po metodologiji CRISP-DM je sestavljen iz šestih medsebojno povezanih faz: (1) razumevanje poslovnega problema, (2) razumevanje podatkov, (3) priprava podatkov, (4) modeliranje, (5) vrednotenje in (6) uporaba (za podrobnosti glej prilogo 1). Celoten proces je zasnovan ciklično, zaporedje izvajanja posameznih faz pa ni fiksno določeno (premiki naprej in nazaj so vedno potrebni, odločitev o naslednji fazi pa je močno pogojena z rezultati, ki jih dobimo v določeni fazi).

2.1.2 Interdisciplinarnost odkrivanja znanja v podatkih

Disciplina podatkovnega rudarjenja se je razvila in se razvija na presečišču različnih znanstvenih disciplin, katerih skupni imenovalec je usmerjenost k večanju uporabnosti golih podatkov in izkopavanju znanja iz njih (Fayyad, Piatetsky-Shapiro, Smyth, 1996, str. 39). Ob dejstvu, da se tehnike podatkovnega rudarjenja pri odkrivanju vzorcev in zakonitosti močno opirajo na znane tehnike s področja umetne inteligence (ang. artificial intelligence), strojnega

---

[8] Kratica SEMMA (Sample, Explore, Modify, Model, Asses) je sestavljena iz začetnih črk izrazov za faze v tem procesu – izbira vzorca, raziskovanje podatkov, preoblikovanje podatkov, modeliranje, ocenjevanje.

[9] Celovitost druge ravni pomeni, da zajema cel proces podatkovnega rudarjenja in vse možne uporabe podatkovnega rudarjenja.

[10] Stabilnost modela pomeni, da je model veljaven tudi, če bi prišlo v posameznih fazah procesa do izboljšav.

učenja (ang. machine learning), razpoznavanja vzorcev (ang. pattern recognition), vizualizacije podatkov (ang. data visualisation) in drugih, se marsikdo vpraša, v čem se podatkovno rudarjenje od omenjenih disciplin sploh razlikuje.

Odgovor je v tem, da posamezne discipline dejansko prispevajo le del metod, ki se uporabljajo in so potrebne za učinkovito odkrivanje znanja. Poleg tega je pri podatkovnem rudarjenju pomemben celoten proces odkrivanja znanja; to pomeni, da je treba rešiti tudi vprašanja dostopa do podatkov in načina njihovega shranjevanja, učinkovitega prilagajanja algoritmov veliki količini podatkov ter dileme glede interpretacije in vizualizacije rezultatov. Na proces podatkovnega rudarjenja lahko tako gledamo tudi kot na večdisciplinarno aktivnost, ki združuje različne tehnike na način, ki je zunaj dometa disciplin, iz katerih le-te izhajajo (Fayyad, Piatetsky-Shapiro, Smyth, 1996, str. 40).

Na podatkovno rudarjenje pa ne vplivajo samo izrazito tehnične discipline, ampak so nanj pomembno vplivali tudi razvoj znanosti na splošno ter filozofija, logika in predvsem statistika (Klösgen, Żytkow, 2002b, str. 22). Statistika namreč že več kot stoletje močno vpliva na razvoj metod za analizo podatkov in hkrati ponuja orodja, s katerimi lahko zajamemo in uporabljamo nedeterministično sestavino empiričnega znanja, s tem pa nadgrajujemo logični pristop k temu znanju (Klösgen, Żytkow, 2002b, str. 25). Glede na to, da je odkrivanje znanja v podatkih v osnovi statistična naloga, statistika ponuja tudi jezik in ogrodje za merjenje negotovosti, do katere pride, ko skušamo zakonitosti, ki smo jih našli ob preučevanju vzorca, prenesti na celotno populacijo (Fayyad, Piatetsky-Shapiro, Smyth, 1996, str. 40).

Ne glede na povezanost podatkovnega rudarjenja s statistiko, ali pa morda ravno zato, pa je imel izraz podatkovno rudarjenje med statistiki vse od šestdesetih let preteklega stoletja, ko so bile prvič predstavljene tehnike analize podatkov s pomočjo računalnikov, negativno konotacijo. Vzrok za to je bil v dejstvu, da so zgodnje tehnike podatkovnega rudarjenja temeljile le na metodah grobe sile (ang. brute-force methods) in se niso ozirale na statistične zakonitosti; to pa je ob dovolj velikem število ponovitev testiranja lahko privedlo tudi do potrditve dejansko neveljavnih hipotez (Klösgen, Żytkow, 2002a, str. 3). V zadnjih letih je bil storjen velik korak naprej k razumevanju statističnih vidikov testiranja (glej npr. Friedman, 1997) in med statistiki je vse manjkrat slišati »definicijo«, po kateri s podatkovnim rudarjenjem »*mučimo podatke, dokler le-ti ne priznajo«*. Če analitik ve, kaj dela, in pozna omejitve podatkovnega rudarjenja, lahko v procesu odkrivanja znanja v podatkih pride do povsem veljavnih in zanesljivih rezultatov; drugače se je rudarjenju boljše izogniti.

Tako kot za že omenjene tehnične discipline tudi za statistiko velja, da so njeni pristopi pri podatkovnem rudarjenju nadgrajeni in prilagojeni velikim količinam podatkov. Statistične tehnike same po sebi niso dovolj, da bi se z njimi lotili težjih izzivov podatkovnega rudarjenja, še posebej tistih, ki so povezani z analizo velikih naborov podatkov (Hand, Mannila, Smyth, 2001, str. 17). Vseeno pa odkrivanje znanja v podatkih ni zgolj raziskovalna

statistika (ang. exploratory statistics) na velikih količinah podatkov (Kardaun, Alanko, 1998). Pristopi podatkovnega rudarjenja se namreč v primerjavi s standardnimi statističnimi metodami izkažejo še posebej uporabni, kadar je v podatkih veliko število potencialno pomembnih spremenljivk, kadar se problem nanaša na analizo multidimenzionalnih in v posameznih subpopulacijah različnih odnosov, kadar za reševanje problema še ni veljavnega statističnega modela in kadar lahko za podskupine podatkov pričakujemo presenetljive rezultate[11] (Klösgen, Żytkow, 2002a, str. 4). Ne nazadnje se podatkovno rudarjenje razlikuje od statistike tudi v smislu namena zbiranja podatkov. Pri podatkovnem rudarjenju so podatki, ki jih analiziramo, tipično zbrani za kak drug namen (npr. za izdajo računov), zato tej analizi pogosto pravimo tudi *sekundarna analiza*. V nasprotju s tem se statistika ukvarja s primarno analizo podatkov, za katero je značilno, da za iskanje odgovora na določeno vprašanje določa tudi učinkovite strategije zbiranja podatkov (Hand, Mannila, Smyth, 2001, str. 2).

Naslednja pomembna disciplina za razvoj podatkovnega rudarjenja je področje podatkovnih baz in sistemov za podporo odločanju. Ne le da so podatkovne baze odprle pot množičnemu zbiranju, shranjevanju in obdelavi podatkov, ampak so uveljavile tudi logičen pogled na svet, ki predstavlja trdno podlago podatkovnemu rudarjenju (Klösgen, Żytkow, 2002b, str. 27). Baze podatkov predstavljajo tudi infrastrukturo, prek katere pri podatkovnem rudarjenju dostopamo do podatkov in jih upravljamo. Posebej uporabna so v tem smislu podatkovna skladišča, ki učinkovito rešujejo težavo količine in integracije podatkov ter njihovo preoblikovanje v obliko, ki je primerna za nadaljnje analize (Berry, Linoff, 2000, str. 16).

Ne nazadnje ne smemo pozabiti na pomemben prispevek sodobne informacijske tehnologije; ta je po eni strani v preteklosti omogočila zbiranje velike količine podatkov, po drugi strani pa eksplozivna rast računalniške moči in padanje razmerja med ceno in zmogljivostjo danes omogočata tudi vedno nove in nove načine analize podatkov (Berry, Linoff, 2000, str. 19). Vse bolj se v ospredje prebijajo tudi tehnike s področja pridobivanja informacij (ang. information retrieval) in obdelave naravnega jezika (ang. natural language processing), ki so uporabne za vedno pomembnejšo analizo vsebine dokumentov in medsebojno primerjavo besedil.

2.2 VRSTE ODKRIVANJA ZNANJA V PODATKIH

Že v prejšnjem poglavju sem omenil, da podatkovno rudarjenje ni enovit pojem, ki bi natančno opredeljeval vse najpomembnejše vrste in aplikacije rudarjenja; teh je namreč veliko ravno zaradi tesne povezanosti podatkovnega rudarjenja s številnimi drugimi znanstvenimi

[11] Vloga presenetljivih ali nepričakovanih rezultatov je pri podatkovnem rudarjenju pogosto preveč poudarjena. Dejansko lahko analitik, ki dobro pozna določeno področje, večino dobljenih rezultatov logično razloži, čeprav morda na začetku procesa o njih ni razmišljal.

disciplinami. Da bi lahko potencialni uporabniki rudarjenja prepoznali tiste sisteme, ki najbolj ustrezajo njihovim potrebam, jih je treba kategorizirati; za ta namen lahko uporabimo različna merila, kot so vrsta podatkov, vrsta podatkovnega modela, vrsta znanja, namen rudarjenja, uporabljene tehnike itd. (Han, Kamber, 2001, str 29; Klösgen, Żytkow, 2002, str. 6).

V zadnjem času se je najbolj uveljavila delitev podatkovnega rudarjenja na rudarjenje po strukturiranih podatkih, rudarjenje po besedilih in rudarjenje po svetovnem spletu (glej prilogo 2). Taka delitev sicer ni najbolj konsistentna, saj se za delitev ne uporablja enovito merilo; na eni strani gre za delitev glede na strukturiranost podatkov, po drugi pa tudi za delitev glede na mesto shranjevanja in način dostopa. Rudarjenje po spletu je namreč kombinacija rudarjenja po strukturiranih podatkih in rudarjenja po besedilih, vendar pa je svetovni splet v poslovanju organizacij in življenju uporabnikov že tako pomemben, da se podatke, dosegljive prek oz. s spleta, obravnava ločeno; hkrati velja tudi, da arhitektura svetovnega spleta zahteva uporabo posebej prilagojenih metod in tehnik.

2.2.1 Rudarjenje po strukturiranih podatkih in rudarjenje po besedilih

Rudarjenje po strukturiranih podatkih je najstarejša oblika rudarjenja, zato se izraz podatkovno rudarjenje najpogosteje povezuje prav s to vrsto rudarjenja. Vanjo uvrščamo tudi širši strokovni javnosti najbolj znane aplikacije podatkovnega rudarjenja, kot sta npr. analiza nakupovalne košarice (ang. market basket analysis) ali odkrivanje prevar v bančništvu (ang. fraud detection). Strukturirani podatki so v najširšem smislu podatki, pri katerih je natančno opredeljen nabor lastnosti posameznega objekta in njihova lokacija[12]. Pri tem je pomembno poudariti, da je lahko tudi posamezna lastnost objekta predstavljena v strukturirani ali nestruktirurani obliki, kar je razlog, da strukturiranih podatkov ne moremo preprosto enačiti z zbirko podatkov.

Večji del podatkov pa je pravzaprav shranjen v nestrukturiranih obliki, npr. v dokumentih. Razpoložljivost različnih zbirk dokumentov in elektronskih informacij raste tako hitro, da je treba za njihovo analizo, organiziranje, klasificiranje in označevanje ter za izkopavanje informacij iz njih uporabiti koncepte podatkovnega rudarjenja; v okviru rudarjenja po besedilih so se zato razvile posebne tehnike za dostop do nestrukturiranih informacij ter za analizo nabora značilnosti dokumentov, ki je navadno veliko večji in bolj raztresen (ang. sparse) kot nabor atributov strukturiranih podatkov (glej Feldman, 2002).

Značilni problem rudarjenja po dokumentih je kategorizacija dokumentov v skladu z vnaprej definirano taksonomijo. Besedila se za ta namen se običajno predstavi v obliki vektorja besed (ang. word-vector), pri katerem posamezno dimenzijo predstavlja utež, izračunana na osnovi

[12] Strukturirani podatki so danes največkrat shranjeni v relacijskih podatkovnih zbirkah ali skladiščih podatkov.

pogostnosti ponovitve posamezne besede v besedilu[13]. Podobnost med besedili se nato izmeri s kosinusno podobnostjo (ang. cosine similarity), pri kateri sta si besedili tem bolj podobni, kolikor manjši je kot med vektorjema, s katerima sta predstavljeni (Mladenić, Grobelnik, 2003).

Rudarjenje po besedilih ni uporabno le za kategorizacijo, ampak tudi za iskanje potencialno relevantnih podatkov v okviru posameznega področja preučevanja, za vsebinsko primerjavo dokumentov iz različnih virov, za identifikacijo potencialnih virov podatkov, za samodejno odgovarjanje na različne elektronske zahtevke ter za (spletno) iskanje v obliki prostega besedila (Sundgren, 2004).

2.2.2 Rudarjenje po spletu

Rudarjenje po podatkih o uporabi spleta lahko opredelimo kot uporabo tehnik podatkovnega rudarjenja, in sicer z namenom samodejnega odkrivanja in izločanja informacij iz spletnih dokumentov in storitev (Kosala, Blockeel, 2000, str. 2). Glede na to, katere podatke uporabljamo v procesu rudarjenja in kakšen je njegov namen, lahko razlikujemo *rudarjenje po podatkih o uporabi spleta* (ang. web usage mining – WUM), *rudarjenje po vsebini spleta* (ang. web content mining) in *rudarjenje po strukturi spleta* (ang. web structure mining). Medtem ko sta rudarjenje po podatkih o uporabi spleta in rudarjenje po vsebini spleta tesno povezana z rudarjenjem po strukturiranih podatkih oz. rudarjenjem po besedilih, črpa rudarjenje po strukturi spleta svoje metode predvsem iz raziskav socialnih mrež (ang. social network) in analize navedkov (ang. citation analysis) (Kosala, Blockeel, 2000, str. 9).

Med omenjenimi tremi (pod)disciplinami sicer ne moremo potegniti povsem jasne ločnice, saj so med seboj povezane; vseeno lahko rečemo, da se pri rudarjenju po vsebini in strukturi spleta bolj osredotočamo na odkrivanje uporabnih informacij v spletu in na analiziranje, kategoriziranje in klasificiranje dokumentov, pri rudarjenju po podatkih o uporabi spleta pa se usmerjamo k odkrivanju znanja o ljudeh, ki splet uporabljajo, njihovih interesih in pričakovanjih, težavah, s katerimi se spopadajo, in o njihovih zahtevah (Galeas, 2005).

2.2.2.1 Rudarjenje po podatkih o uporabi spleta

Rudarjenje po podatkih o uporabi spleta se usmerja na tehnike za napovedovanje vedenja uporabnikov spleta in v osnovi temelji na podatkih o obisku spletišča; ti se na različne načine shranjujejo ob vsakem zahtevku za prikaz določenega spletnega vira. Ker so podatki o uporabi največkrat shranjeni v dnevniških datotekah, ki jih samodejno ustvarjajo spletni

[13] Za določanje uteži se najpogosteje uporablja metoda TF-IDF (term frequency–inverse document frequency), pri kateri je vrednost uteži popravljena glede na pogostnost besede v celotni analizirani zbirki besedil.

strežniki, se za rudarjenje po podatkih o uporabi spleta včasih uporablja tudi izraz rudarjenje po spletnih dnevniških datotekah (ang. web log mining).

Rudarjenje po podatkih o uporabi spleta lahko uporabljamo za spremljanje splošnih vzorcev dostopa ali pa za spremljanje trendov in vzorcev individualnih uporabnikov. Prvi pristop je primeren za analiziranje učinkovitosti spletišča in izboljševanje njegove uporabnosti ter optimizacijo delovanja spletnih strežnikov, drugi pa za dinamično prilagajanje in poosebljanje spletišča na podlagi predvidenih interesov in želja uporabnikov (Galeas, 2005). Pridobljeno znanje nam pomaga, da predvsem boljše razumemo vedenje uporabnikov, in nam tako omogoča, da lažje zadovoljujemo potrebe uporabnikov in uresničujemo temeljni namen posameznega spletnega mesta.

2.2.2.2 Rudarjenje po vsebini spleta

Rudarjenje po vsebini spleta je proces izkopavanja znanja iz vsebine spletnih dokumentov ali njihovih metapodatkov. Tehnike za to vrsto rudarjenja lahko delimo na tiste, ki izvirajo iz področja pridobivanja informacij, in na tiste, ki izvirajo iz področja zbirk podatkov. Namen rudarjenja po vsebini spleta z vidika pridobivanja informacij je v osnovi podpreti in izboljšati iskanje informacij ali njihovo filtriranje v skladu s potrebami posameznih uporabnikov, namen rudarjenja po vsebini spleta z vidika zbirk podatkov pa je modeliranje podatkov na spletu z namenom, da se izboljša delovanje iskalnih strojev. Obe skupini tehnik se med seboj ne razlikujeta samo v metodah dela, ampak tudi v strukturiranosti in vrsti podatkov ter načinu njihove predstavitve (Kosala, Blockeel, 2002).

2.2.2.3 Rudarjenje po strukturi spleta

Namen rudarjenja po strukturi spleta je razkriti model organiziranosti svetovnega spleta, ki temelji na topologiji povezav med spletnimi objekti in je uporaben za kategorizacijo in rangiranje spletišč, pa tudi za iskanje podobnosti in povezav med njimi (Baglioni et al., 2003). Rudarjenje po strukturi spleta je tesno povezano z rudarjenjem po vsebini spleta, saj so hiperpovezave del vsebine, a se v primerjavi s slednjim usmerja na zunanjo strukturo spletnih dokumentov in ima zato širše področje delovanja. Tipičen primer odkrivanja znanja iz strukture spleta je identifikacija specifičnih tipov spletišč, kot so npr. zvezdišča (ang. hub) ali uveljavljena spletišča (ang. autohority), odkrivanje virtualnih skupnosti ali merjenje popolnosti in replikacij spletišč (Kosala, Blockeel, 2000, str. 9). Rudarjenje po strukturi spleta je zelo uporabno tudi znotraj posameznega spletišča, saj sama struktura skriva veliko pomembnih informacij. Tako je bilo npr. pokazano, da lahko iz strukture povezav med zadetki v iskalniku ugotovimo, ali je seznam zadetkov dober ali ne, in tudi precej natančno sklepamo, ali bo uporabnik iskalni pogoj nato spremenil (Leskovec, 2007).

3 RUDARJENJE PO PODATKIH O UPORABI SPLETA

Kot vsak drug način podatkovnega rudarjenja je tudi rudarjenje po podatkih o uporabi spleta proces, sestavljen iz več korakov. V nadaljevanju podrobno predstavljam tiste faze v procesu, ki jih je treba izvesti, ko so cilji rudarjenja že določeni (koraki 3, 4 in 5 po metodologiji CRISP-DM), saj se pri njih pokaže največja razlika med rudarjenjem po podatkih o uporabi spleta in drugimi vrstami podatkovnega rudarjenja.

3.1 ZBIRANJE PODATKOV

Kakovostni podatki so osnovni pogoj za kakovost in uporabnost vzorcev in pravil, ki jih najdemo v njih, zato je zbiranje podatkov eden izmed ključnih korakov pri odkrivanju znanja v podatkih. Osnovni vir podatkov za rudarjenje po podatkih o uporabi spleta so podatki o uporabi spletišča, za potrebe lažje in učinkovitejše priprave podatkov in odkrivanja vzorcev pa se poleg osnovnih podatkov vse bolj priznava tudi pomen podatkov o uporabnikih spletišča (Markellou, Rigou, Sirmakessis, 2005; Pierrakos et al., 2003), pa tudi podatkov o strukturi in vsebini spletišča ter semantičnega znanja (Eirinaki, Vazirgiannis, 2003; Mobasher, 2005).

Ena izmed osrednjih tem pri zbiranju podatkov o uporabnikih in njihovem vedenju je spoštovanje in varovanje zasebnosti uporabnikov. Zasebnost uporabnikov lahko definiramo kot *»moč nadzorovanja tega, kaj lahko drugi vedo o nas, in določanja vstopnih pogojev v naš osebni prostor«* (Europe's Information Society Thematic Portal – Privacy Protection, 2007). V skladu z zakonskimi in etičnimi načeli informacijske zasebnosti bi moral biti obiskovalec spletišča jasno obveščen o tem, katere podatke organizacija zbira o obiskovalcih, kako jih obdeluje in za kakšne namene jih uporablja, hkrati pa bi moral imeti možnost, da se s tem strinja ali ne (Wel, Royakkers, 2004, str. 129). V tem smislu je organizacija World Wide Web Consortium (v nadaljevanju W3C) razvila platformo za zasebnostne nastavitve (ang. Platform for Privacy Prefereces, v nadaljevanju P3P), ki je postala standard za obveščanje uporabnika o naboru in namenu zbiranja osebnih informacij na spletišču. P3P določa standardni zapis politike zasebnosti, ki jo lahko ustrezno podprte aplikacije samodejno preberejo in interpretirajo; s tem uporabnikom omogoča, da na konceptualni ravni določijo, katere informacije in pod kakšnimi pogoji bodo, če sploh, razkrite, in jim daje moč nad postavitvijo meje med zasebnim in javnim. Namen P3P je povečati zaupanje uporabnikov v svetovni splet, vendar pa zgolj določa standarde za objavo politike zasebnosti in primerjavo le-te z uporabnikovimi nastavitvami, ne more pa tudi nadzorovati verodostojnosti in pravilnosti izvedbe politike zasebnosti (P3P Project, 2006). Z vidika zbiranja podatkov za potrebe rudarjenja po podatkih o uporabi spleta je standard P3P pomemben, ker se uporablja v vseh priljubljenejših brskalnikih in lahko močno vpliva na uspešnost prepoznave uporabnika na podlagi uporabe tehnologije piškotkov (ang. cookies).

3.1.1 Podatki o uporabi spletišča

Podatki o uporabi spletišča so podatki o vseh aktivnostih, ki jih obiskovalci izvajajo na izbranem spletišču, in predstavljajo dragocen vir informacij o njihovem vedenju. Zbiramo jih lahko na ravni strežnika, na ravni odjemalca ali na posredniški ravni. Za potrebe rudarjenja po podatkih o uporabi spleta se največkrat uporabljajo podatki, ki se na strežniku shranjujejo v spletne dnevniške datoteke (ang. web log file, v nadaljevanju tudi dnevniška datoteka), zato za to zvrst rudarjenja včasih zasledimo kar ime »rudarjenje po spletnih dnevnikih« (ang. web log mining) (Chen, Fu, Tong, 2003; Yang, Zhang, 2001).

Večina spletnih strežnikov privzeto podpira splošno obliko zapisa dnevniške datoteke (Common Log File Format), pri kateri se v dnevniško datoteko shranjujejo osnovni podatki o zahtevi, ki jo odjemalec pošlje spletnemu strežniku. Spletni strežniki poleg tega povečini omogočajo tudi spremljanje zahtevkov v skladu z razširjeno obliko zapisa, največkrat v skladu z zapisom W3C[14] (za podroben opis glej prilogo 3); ta omogoča tudi beleženje podatkov, kot so smer prihoda, vsebina piškotka, vrednost poizvedovalnega niza (ang. query string) ipd. Z vidika zasebnosti je občutljiv predvsem podatek o smeri prihoda, zato protokol HTTP (RFC 2616, 1999) priporoča, da ima uporabnik možnost, da sam odloči o tem, ali bo dovolil pošiljanje tega podatka ali ne, česar pa večina brskalnikov ne upošteva.

Čeprav najbolj razširjeni spletni strežniki (npr. Microsoft Internet Information Server in Apache) upravitelju spletišča omogočajo izbor oz. določanje zapisa dnevniških datotek, je ta (avtomatizirani) način beleženja dokaj neprilagodljiv in kot tak v določenih pogledih manj primeren za spremljanje posebnih vrst podatkov o uporabi. Poleg tega podatki, shranjeni v dnevniških datotekah, že sami po sebi niso vedno zanesljiv vir. Osnovni razlog za nezanesljivost so različne stopnje predpomnjenja (ang. caching) v spletnem okolju (Srivastava et al., 2000, str. 13). Predpomnjenje je mehanizem, ki se lahko izvaja na strani odjemalca ali na ravni posredovalnega strežnika in je namenjen *»zmanjševanju obremenjenosti spletnih strežnikov in skrajševanju časa nalaganja podatkov ob identičnih prihodnjih zahtevkih«* (RFC 2616, 1999), ko se lahko stran prebere iz predpomnilnika, namesto da bi se prebrala iz izvornega spletnega strežnika. Posledično odjemalčeva zahteva sploh ne pride do spletnega strežnika, zato v dnevniški datoteki ni podatka o ogledu strani. Težavo lahko obidemo z uporabo direktive »cache-control« v HTTP glavi strani, ki jo morajo v skladu s protokolom HTTP upoštevati tako posredovalni strežniki kot odjemalci. S tem pristopom, ki se popularno imenuje tudi »razbijanje predpomnjenja« (ang. cache-busting) (Pitkow, 1997, str. 1348), lahko teoretično dosežemo tudi beleženje zahtevkov za ogled strani, ki jo je uporabnik v sklopu seje že videl, nato pa do nje dostopil prek gumbov »Nazaj« (ang. back) ali »Naprej« (ang. forward) oz. drugih mehanizmov za ogled zgodovine videnih strani; težje je ta

[14] Poleg zapisa W3C je precej razširjen tudi zapis NCSA, ki ga je razvilo Ameriško središče za uporabo superračunalnikov (National Center for Supercomputer Applications).

mehanizem uporabiti za preprečevanje predpomnjenja neznakovnih datotek (ang. non-text file) (npr. PDF datotek)[15]. Težava razbijanja predpomnjenja je tudi v tem, da povečuje obremenjenost spletnega strežnika in daljša odzivne čase, hkrati pa ne moremo zagotovo vedeti, ali se direktiva upošteva v vseh predpomnilnikih na poti med odjemalcem in spletnim strežnikom (Pitkow, 1997, str. 1348).

Kakovost podatkov v spletnih dnevniških datotekah lahko izboljšamo, če v analizo vključimo tudi dnevniške datoteke posredovalnih strežnikov, čeprav to zaradi številnosti teh strežnikov največkrat ni mogoče. Nekateri avtorji (npr. Pierrakos et al., 2003; Srivastava et al., 2000; Cunha, Bestavros, Crovella, 1995) zato navajajo, da so zanesljivejši tisti podatki o uporabi, ki so zbrani na strani odjemalca; zbiramo jih lahko z uporabo funkcij ukaznega jezika JavaScript[16] ali z javanskimi programčki (ang. Java applet)[17] ali pa tako, da uporabnikom v uporabo ponudimo za te namene prilagojen brskalnik[18]. V praksi se spremljanje obiska na strani odjemalca večinoma izvaja z jezikom JavaScript, katerega osnovna pomanjkljivost je v tem, da ga mora odjemalec najprej podpirati, nato pa tudi omogočati; odstotek uporabnikov, ki blokirajo JavaScript, znaša na svetovni ravni po zadnjih podatkih okrog 6 % (Browser Statistics, 2007), vendar pa se lahko ta delež med spletišči močno razlikuje, zato jih je treba preveriti tudi na konkretnem spletišču. Izključevanje določenega števila uporabnikov, ki ne podpirajo izvajanja skript na odjemalcu, je vsekakor ena izmed pomanjkljivosti tega načina, vendar pa s seboj hkrati prinaša tudi prednosti pri uporabi podatkov v procesu spletnega rudarjenja; večina pajkov (ang. spider)[19], ki s pregledovanjem spletišč močno izkrivljajo dejansko sliko vedenja uporabnikov, JavaScript ukazov ne podpira[20], zato njihovi ogledi v statistiki obiska niso zajeti, kar močno olajša izvajanje nalog v fazi čiščenja podatkov.

Med pomanjkljivosti odjemalskih skript lahko uvrstimo tudi razpoložljivost in dostopnost podatkov. Na strani strežnika se podatki o obisku v dnevniške datoteke shranjujejo, ne da bi bilo treba za ta namen pisati posebne programe, in so skrbnikom spletišča vedno na voljo v izvorni obliki. Nasprotno je treba odjemalske skripte pripraviti posebej in jih nato tudi posebej vključiti v vsako spletno stran; na svetovnem spletu lahko sicer najdemo veliko podjetij, ki

---

[15] Na straneh, ki se izvajajo dinamično, lahko HTTP glavo posredno dodamo tudi nebesedilnim datotekam, vendar pa povezava na datoteke ne sme kazati neposredno, ampak na posebno stran, ki samodejno odpre datoteko in ima v glavi ustrezno določene direktive za preprečevanje predpomnjenja.

[16] JavaScript je programski jezik, ki je namenjen pisanju skript, ki so del dokumentov HTML in se izvajajo v spletnih brskalnikih.

[17] Javanski programček je program, napisan v programskem jeziku Java in dodan dokumentu HTML.

[18] Glej npr. primera prilagoditve brskalnika NCSA Mosaic (Cunha, Bestavros, Crovella, 1995) in izgradnje samostojnega brskalnika Ginis Framework, ki je po videzu identičen brskalniku Internet Explorer 6 (Valeyathan, Yamada, 2007).

[19] Pajek je program, ki samodejno poišče povezave na druge spletne strani in jim nato sledi, zato lahko poišče veliko število spletnih strani. Poleg imena pajek se za ta tip programov uporabljajo tudi imena robot (ang. robot), gosenica (ang. crawler), črv (ang. worm) in popotnik (ang. wanderer).

[20] Med izjeme npr. štejemo robote, katerih namen je na spletišču iskati elektronske naslove (WebDeveloper.com, 2007).

(večinoma) brezplačno ponujajo potrebne skripte in skrbijo tudi za spremljanje vseh potrebnih podatkov ter za njihovo osnovno statistično obdelavo, vendar pa organizacije do teh podatkov ne morejo dostopati v obliki, ki jo potrebujejo za rudarjenje, niti nimajo nadzora nad zbiranjem podatkov in zagotavljanjem njihove kakovosti. V tej luči je večinoma najbolj smiseln razvoj lastnih rešitev, ki jih lahko organizacija povsem prilagodi svojim potrebam in zahtevam spletišča.

3.1.2 Podatki o uporabnikih

Podatke o uporabnikih pogosto imenujemo uporabniški profili, mednje pa prištevamo podatke o osebnih značilnosti obiskovalcev spletišča in njihovih aktivnostih. Uporabniški profili nam pri rudarjenju po podatkih o uporabi spleta in spletnemu poosebljanju po eni strani omogočajo natančnejše razločevanje vzorcev uporabe, po drugi strani pa omogočajo natančnejše in učinkovitejše poosebljanje spletne vsebine, saj je dodana vrednost vzorcev in pravil, ki nam jih v nadaljnjih korakih rudarjenja po podatkih o uporabi spleta uspe izluščiti iz podatkov o uporabi spletišča, veliko večja, če poznamo uporabnike, ki stojijo za posameznimi dejanji. Poznavanje uporabnikov nam tako omogoča natančnejše razločevanje vzorcev uporabe in nam pomaga pri njihovi analizi in pojasnjevanju. Markellou, Rigou in Sirmakessis (2005) prištevajo med najpogostejše podatke, ki sestavljajo uporabniški profil, demografske podatke, podatke o tem, kakšno je znanje uporabnikov o konceptih in povezavah med koncepti v določenem vsebinskem področju, podatke o drugih znanjih in sposobnostih uporabnikov, podatke o interesih in preferencah ter podatke o uporabnikovih načrtih in ciljih. Našteti podatki dejansko sestavljajo zgolj statični del profila; v dinamični del profila uporabnika se namreč shranjujejo podatki o njegovih aktivnostih.

Organizacije uporabljajo za pridobivanje podatkov o uporabnikih različne strategije. Statične podatke večinoma pridobivajo eksplicitno, tj. prek spletnih obrazcev in različnih vrst vprašalnikov, dinamične pa zbirajo implicitno medtem ko uporabniki brskajo po spletišču (Eirinaki, Vazirgiannis, 2003, str. 5). Nakupa demografskih podatkov o uporabnikih od specializiranih ponudnikov slovenske organizacije po večini ne uporabljajo, zaradi številnih težav z zasebnostjo pa je tudi v tujini ponudnikov takih podatkov vedno manj, saj uporabniki neradi dovolijo, da organizacija njihove podatke prodaja tretjim osebam. Uporabnikova percepcija zasebnosti močno vpliva tudi na kakovost in točnost eksplicitno zbranih podatkov; v strahu pred razkritjem ali zlorabo uporabniki v registracijskih obrazcih pogosto navajajo netočne podatke, kar močno zmanjša njihovo uporabnost. Eden izmed ključnih ukrepov, ki jih lahko organizacija v procesu zbiranja podatkov zato naredi, je, da povsod, kjer od uporabnika pričakuje, da bo vpisal svoje podatke, jasno opiše namen uporabe teh podatkov in dodano vrednost za uporabnika, hkrati pa za svojo verodostojnost skrbi tudi v navezi s P3P.

3.1.2.1 Povezovanje podatkov o uporabnikih s podatki o uporabi spletišča

Za uspešnost rudarjenja po podatkih o uporabi spleta je zelo pomembno, da lahko že v koraku zbiranja podatkov zagotovimo možnost spremljanja uporabe v okviru posamezne strežniške seje in možnost prepoznavanja uporabnikov pri ponavljajočih obiskih. Organizacije imajo za ta namen na voljo več metod, kot so npr. identifikacija na podlagi identifikacijskega protokola[21], sklepanje na podlagi številke IP[22] ali identifikacija z uporabo tehnologije piškotkov (Eirinaki, Vazirgiannis, 2003), ki pa niso popolnoma zanesljive.

Večina spletišč za identifikacijo uporabnikov uporablja tehnologijo piškotkov[23]. Piškotek je *»podatek, ki ga spletni strežnik pošlje spletnemu odjemalcu, ta pa ga ob naknadnih zahtevah pošlje nazaj strežniku«* (Web Characterization Terminology & Definitions Sheet). Tehnično gledano strežnik pošlje piškotek odjemalcu kot HTTP-glavo in njegovo vrednost zapiše v spletno dnevniško datoteko, brskalnik pa podatek glede na čas njegove veljavnosti shrani v delovni spomin računalnika ali zapiše na disk. Čas veljavnosti je pomemben z vidika identifikacije uporabnika; s piškotki sej (ang. session cookies) lahko uporabnika identificiramo samo znotraj enega obiska (seje), trajni piškotki (ang. persistent cookies) pa omogočajo identifikacijo znotraj poljubno določenega časovnega okvira.

Piškotki ne omogočajo kraje podatkov iz osebnega računalnika, vseeno pa so ena izmed tehnologij, ki je med uporabniki z vidika zasebnosti najmanj zaželena, saj se pogosto uporabljajo za namene, ki bolj kot uporabnikom služijo organizacijam (npr. za ciljno serviranje oglasov). Uporabniške varnostne nastavitve zato vedno pogosteje preprečijo shranjevanje trajnih piškotkov ali pa uporabniki piškotke redno brišejo[24]. Da bi se lahko organizacije kolikor se da izognile težavam, ki jih nezmožnost identifikacije uporabnika predstavlja pri rudarjenju po podatkih o uporabi spleta, bi morale vsaj:

- za identifikacijo uporabnika znotraj seje uporabljati piškotke sej, saj njihovo shranjevanje večina brskalnikov dovoljuje tudi pri zelo visoki stopnji deklarirane zasebnosti;

- povečati doseg trajnih in začasnih piškotkov z objavo politike zasebnosti v skladu s standardom P3P (pri dani ravni zasebnosti brskalniki spletiščem, ki imajo politiko zasebnosti definirano v obliki P3P, dovoli shranjevanje piškotka, drugim pa ne – za

---

[21] Identifikacijski protokol (RFC 1413, 1993) je protokol, ki omogoča ugotavljanje identitete uporabnika na podlagi povezave TCP. Težava pri uporabi protokola za identifikacijo je v tem, da mora za delovanje protokola odjemalec eksplicitno dovoljevati prenos podatkov prek vrat 113 (RFC 1413, 1993).

[22] Številka IP je naslov v obliki 32-bitne kode kot desetiškega zapisa s pikami, ki enolično označuje vsak računalnik, povezan v internet.

[23] Piškotke je razvilo podjetje Netscape Corporation, da bi tako pri e-nakupovanju omogočilo shranjevanje podatkov o vsebini nakupovalne košarice.

[24] Po podatkig podjetja Iprom naj bi okrog 19 % slovenskih uporabnikov interneta brisalo piškotke, 7 % uporabnikov pa naj bi jih blokiralo (Iprom, 2006).

primer glej prilogo 4).

3.1.3 Drugi pomembni podatki

Za učinkovito rudarjenje po podatkih o uporabi spleta so poleg podatkov o uporabi spletišča in podatkov o uporabnikih pomembni tudi drugi podatki, kot so podatki o strojnem in programskem okolju uporabnikov (Markellou, Rigou, Sirmakessis, 2005) ter podatki o vsebini in strukturi spletišča (Srivastava et al., 2000). Z njimi lahko izboljšamo učinkovitost uporabljenih algoritmov podatkovnega rudarjenja ali pa povečamo zanesljivost hevrističnih algoritmov, s katerimi podatke pripravimo za rudarjenje. Večina podatkov o okolju je na voljo v spletnih dnevniških datotekah (npr. resolucija zaslona, operacijski sistem, brskalnik itd.), nekatere pa morajo organizacije zbirati programsko (npr. razpoložljivost vtičnikov – ang. plug-in).

Podatki o vsebini so množica vseh objektov in povezav med njimi, ki jih spletišče posreduje obiskovalcem prek strani HTML, slik, videa ali zvočnih datotek. Mednje štejemo tudi semantične ali strukturne metapodatke, kot so opisne ključne besede (ang. descriptive keywords), semantične oznake, lastnosti dokumentov itd. Med podatke o vsebini štejemo tudi osnovno področno ontologijo (ang. underlying domain ontology), ki jo lahko zajamemo implicitno ali pa nam je na voljo eksplicitno v obliki konceptualnih hierarhij (ang. conceptual hierarchies), direktorijske strukture ali ontoloških jezikov (ang. ontology language) (Mobasher, Dai, 2004, str. 289).

Podatki o strukturi spletišča kažejo na to, kako je organizirana vsebina spletišča; mednje štejemo tako interno strukturo strani, ki se izraža prek HTML- in XML-oznak, kot tudi strukturo povezav med stranmi, ki jo zajamemo prek hiperpovezav (ang. hyperlink) (Eirinaki, Vazirgiannis, 2003, str. 2). Zbirka vseh povezav med stranmi na spletišču se imenuje topologija povezav (npr. Pirolli, Pitkow, Rao, 1996; Pitkow, 1997) ali tudi topologija spletnega grafa (ang. web site graph) (npr. Spilipoulou et al., 2003; Xing, Shen, 2004). Strukturo najpogosteje zajemamo z uporabo robotov, ki ustrezno obdelajo vse strani na spletišču (Cooley, Mobasher, Srivastava, 1999, str. 12).

3.2 PRIPRAVA IN PREDOBDELAVA PODATKOV

Namen priprave in predobdelave podatkov je odpraviti pomanjkljivosti zbranih podatkov in s tem zagotoviti, da kažejo točno sliko uporabniške aktivnosti na spletišču. Z različnimi

hevrističnimi metodami[25] neobdelane podatke preoblikujemo v obliko, ki ne omogoča le učinkovitejšega izvajanja podatkovnega rudarjenja, ampak je tudi osnova za vsakršno statistično analizo obiska spletišča. Način izvajanja posameznih nalog v procesu je treba zaradi razlik v strukturi in vsebini prilagoditi vsakemu spletišču posebej (Mobasher, 2005, str. 7); podatke je treba tipično očistiti neželenih vnosov, identificirati uporabnike in uporabniške seje, zaključiti poti in podatke združiti v ustrezne semantične enote (Markellou, Rigou, Sirmakessis, 2005, str. 32).

Za poenotenje načina spremljanja statistike obiska in primerljivosti rezultatov je organizacija W3C sprožila aktivnost za opredelitev spleta (W3C Web Characterization Activity, v nadaljevanju WCA), v okviru katere je nastal delovni osnutek definicij pojmov, povezanih s svetovnim spletom. Po WCA je *uporabnik* (ang. user) vsak posameznik, ki prek brskalnika dostopa do ene ali več datotek, shranjenih na spletnem strežniku. *Spletna stran* (ang. web page) je zbirka informacij, sestavljena iz enega ali več spletnih virov, ki so namenjeni hkratnemu prikazovanju in so dosegljivi prek enega URI-naslova[26]. *Ogled strani* (ang. page view) je vizualna predstavitev spletne strani v določenem odjemalcu v določenem trenutku in je sestavljen iz vseh datotek, ki sestavljajo spletno stran. *Uporabniška seja* (ang. user session) je časovno zaključeno zaporedje ogledov strani enega uporabnika na celotnem svetovnem spletu, *strežniška seja* (ang. server session, v nadaljevanju tudi seja) oz. *obisk* (ang. visit) pa podmnožica tega zaporedja znotraj posameznega spletišča; strežniška seja je v številnih aplikacijah in analizah poimenovana kot uporabniška seja (npr. WebTrends, Surfstats; Pierrakos et al., 2003), zato oba izraza uporabljam v magistrskem delu kot sopomenki. *Potek povezav* (ang. click-stream) je končno zaporedje ogledov strani na spletišču, *epizoda* (ang. episode) pa vsaka semantično pomembna podmnožica strežniške seje.

3.2.1 Čiščenje podatkov

Spletni strežniki v dnevniške datoteke beležijo podatke o zahtevkih vseh virov na danem spletišču, ne glede na njihov izvor. Namen čiščenja podatkov je izločiti iz dnevniških datotek podatke o vseh zahtevkih, ki pri rudarjenju po podatkih o uporabi spleta z vidika vsebine in strukture spletišča niso pomembni (Pierrakos et al., 2001) ali izkrivljajo sliko uporabniške aktivnosti na spletišču. Taki zahtevki se v dnevniške datoteke zapisujejo zaradi dveh razlogov: zaradi delovanja posebnih programov, imenovanih pajkov (ang. spider), ki z namenom indeksiranja vsebine ali iskanja informacij samodejno pregledujejo spletišča, in zato, ker se ob ogledu spletne strani v dnevniško ne zapišejo zgolj podatki o ogledu HTML-dokumenta, ampak podatki o vseh zadetkih (ang. page hit), tj. vseh datotekah, ki jo sestavljajo. Kot sem omenil v poglavju 3.1.1, je moč zahtevke pajkov dokaj učinkovito

[25] Hevristična metoda (tudi hevristika, ang. heuristics) je pristop na temelju zaporedja preizkusov, ki dajejo približne rezultate.

[26] URI-naslov (Uniform Resource Identificator) je niz znakov za enolično identifikacijo ali poimenovanje vira.

izločiti, če za spremljanje statistike uporabe uporabljamo odjemalske skripte, vendar pa tudi v tem primeru izločanje ni povsem zanesljivo, zato tega koraka čiščenja podatkov nikakor ne smemo izpustiti.

Beleženje nepomembnih zahtevkov je v osnovi posledica delovanja protokola HTTP; ta za prenos vsake datoteke zahteva posebno povezavo, zato se v dnevniške datoteke zapisujejo tudi prenosi vseh grafičnih in drugih elementov, datotek s stili in skriptami ter podobnih povezanih datotek (Cooley, Mobasher, Srivastava, 1999, str. 16). Ta naloga ne velja za problematično in večina sistemov za rudarjenje po podatkih o vsebini spleta lahko že samodejno briše datoteke z določenim imenom ali končnico (Cooley, Mobasher, Sirvastava, 1999; Koinotites, 2001). Vseeno se je treba brisanja lotiti previdno, saj pravilna identifikacija teh zapisov temelji na poznavanju notranje strukture in vsebine spletišča ter znanju o povezanem vsebinskem področju (Mobasher, 2005, str. 7).

Težja naloga je odstranitev podatkov o straneh, ki so jih zahtevali pajki, saj ni metode, po kateri bi lahko popolnoma zanesljivo vedeli, kdaj je ogled posledica zahtevka pajka, kdaj pa posledica konkretnega uporabnikovega dejanja. Nezanesljivost je posledica delovanja neetičnih robotov, saj ti pri dostopanju do spletišča ne upoštevajo dobre prakse, zato moramo za njihovo identifikacijo uporabiti hevristične metode. Pri teh odkrivanje robotov temelji na analizi njihovega navigacijskega vedenja (Tan, Kumar, 2002, str. 17):

- **Preverjanje dostopa do datoteke robots.txt.** V skladu s predlaganim *Standardom za izključevanje robotov* (Koster, 1994; Kolar, Leavitt, Mauldin, 1996) morajo roboti ob obisku spletišča najprej preveriti vsebino datoteke *robots.txt*, v kateri je v standardnem formatu zapisano, katere strani na spletišču lahko obiščejo. Ob predpostavki, da roboti standard upoštevajo, lahko iz dnevniške datoteke izločimo vse seje, ki vsebujejo tudi zahtevek za omenjeno datoteko. Ta pristop je dokaj učinkovit, vendar pa se nanj ne smemo zanašati, saj standard ni obvezen in se ga mnogi roboti ne držijo.

- **Preverjanje uporabniškega agenta in naslova IP.** Druga priljubljena metoda za identifikacijo robotov temelji na prepoznavanju imena uporabniškega agenta in IP-naslova znanih robotov. Pristop ima dve pomanjkljivosti: zahteva natančen seznam robotov, ki ga je treba nenehno dopolnjevati, poleg tega pa ne omogoča identifikacije neznanih in zakritih robotov. Za reševanje prve težave na svetovnem spletu obstaja več strani, ki naj bi vzdrževale sezname robotov (glej npr. The Web Robots Database, 2007; Identification des robots, 2007), vendar pa je njihova ažurnost vprašljiva[27]. V osnovi je ta pristop zato primeren le za predhodno že identificirane robote.

[27] Na dan 26. 1. 2007 ni nobeden izmed seznamov vseboval podatka o uporabniškem agentu robota, ki ga uporablja najpriljubljenejši slovenski iskalnik Najdi.si – »Mozilla/5.0 (compatible; Najdi.si/3.1)«.

Omenjeni metodi sta učinkoviti za odstranjevanje zahtevkov robotov, vendar pa temeljita na predpostavki, da se vsi roboti obnašajo v skladu z dobro prakso in da upoštevajo neformalna pravila obnašanja. Žal se na to ne moremo zanašati, zato je treba za natančnejšo identifikacijo uporabljati različne hevristične pristope, pri katerih odkrivanje temelji na analizi predvidenega vedenja robotov (Tan, Kumar, 2002, str. 17). Izbor meril za hevristične algoritme in njihovih mejnih vrednosti (ang. threshold) je odvisen od vsebine in strukture spletišča. Pomanjkljivost hevrističnih pristopov je v tem, da temeljijo na pravilih vedenja in kot taki niso popolnoma zanesljivi; čeprav z njimi identificiramo veliko dodatnih robotskih zahtevkov, pa lahko z njimi izločimo tudi povsem veljavne uporabniške zahtevke. V raziskavah, ki so se usmerjale na identifikacijo robotov (Kumar, Tan, 2002; Almeida et. al, 2001; Sen, Hansen, 2000), so se kot učinkoviti izkazali pristopi, ki temeljijo na kombinaciji parametrov, kot so število zahtevkov po metodi HEAD[28] (ang. HEAD request method), število zahtevkov brez podatka o napotitelju, dolžina obiska posamezne strani, število obiskanih strani v okviru ene seje, zaporedje obiskanih strani, število ponovljenih zahtevkov, število nočnih zahtevkov ipd. Nekateri metode temeljijo tudi na identifikaciji dejanj, ki jih roboti ne morejo izvesti (Almeida et. al, 2001), kot so npr. dejanja, ki temeljijo na uspešnem reševanju različnih Turingovih testov[29] (Inaccessibility of Visually-Oriented Anti-Robot Tests, 2003).

3.2.2 Identifikacija uporabnikov in sej

Za analizo uporabe spletišča nam ni treba poznati identitete uporabnikov, moramo pa iz podatkov v dnevniških datotekah znati razločiti različne uporabnike in različne obiske posameznih uporabnikov. Ker standard HTTP takega razlikovanja sam po sebi ne omogoča, moramo za ta namen uporabiti različne proaktivne (ang. proactive, a priori) in reaktivne (ang. reactive, a posteriori) pristope. Medtem ko je namen proaktivnih strategij nedvoumno povezovanje vsakega zahtevka s posamezno osebo pred ali med njegovim obiskom spletišča, skušajo reaktivne strategije te povezave vzpostaviti pozneje na podlagi obstoječih nepopolnih podatkov (Spilipoulou et al., 2003, str. 172). Pri obeh pristopih najprej identificiramo uporabnike in nato zaporedje aktivnosti posameznega uporabnika (tudi dnevnik uporabniških aktivnosti, ang. user acitivty log) razdelimo na posamezne seje (za slednji postopek se v angleški literaturi pogosto uporablja izraz *sessionization*), načeloma pa velja, da so proaktivne strategije natančnejše (Spilipoulou et al., 2003, str. 178).

Vnaprejšnja identifikacija uporabnikov najpogosteje temelji na uporabi piškotkov. Ob upoštevanju pomislekov in težav, predstavljenih v prejšnjem poglavju, so trajni piškoti ustrezna rešitev za razlikovanje različnih uporabnikov, vendar pa na njihovi podlagi ne

---

[28] Metoda HEAD (glava) je način klica spletne strani, pri kateri strežnik vrne samo metapodatke o strani (podatke v glavi strani), ne ta tudi vsebine strani.

[29] Turingov test je preizkus razuma oziroma inteligence, ki ga je predlagal Alan Turing in omogoča razlikovanje med človekom in računalniškim sistemom.

moremo ugotoviti, kdaj se začne in konča posamezni obisk. Ta podatek lahko (v grobem) dobimo z uporabo začasnih piškotkov (ang. session cookies), s pomočjo katerih vsaki seji dodelimo unikatni identifikator (Berendt et al., 2002, str. 181). Identifikator poteče, ko uporabnik zapre brskalnik, ko je povezava prekinjena ali ko se izteče čas trajanja seje[30].

Identifikacija uporabnikov in rekonstrukcija sej pri reaktivnih strategijah temelji na hevrističnih postopkih ob določenih predpostavkah o vedenju uporabnikov ali značilnostih spletišča (Mobasher, 2005, str. 10). Njihov namen je oceniti, ali posamezni ogledi strani pripadajo posameznemu uporabniku in ali so bile strani ogledane v okviru enega ali več zaporednih obiskov. Pri tem želimo, da z izbranim hevrističnim postopkom h čim natančneje sestavimo množico konstruiranih sej (ang. constructed session), C_h; za idealni hevristični postopek h^* velja $C_{h^*} \equiv R$, pri čemer je R množica pravih (realnih) sej (ang. real session) (Berend et al., 2002, str. 181; Spilipoulou, 2003, str. 177). Največkrat se pri reaktivnih strategijah uporabljajo hevristični postopki, ki temeljijo na analizi podatkov o IP-naslovu in uporabniškem agentu ter predpostavkah o trajanju ogledov (časovno orientirani postopki) ali predpostavkah o navigacijskem vedenju obiskovalcev (navigacijsko usmerjeni postopki).

Identifikacija uporabnikov na podlagi naslova IP in uporabniškega agenta je najbolj trivialen postopek, a daje zelo nenatančne rezultate, saj lahko naslov IP zaradi uporabe posredovalnih strežnikov pri mnogih uporabnikih enak, hkrati pa ima lahko isti uporabnik zaradi dinamičnega prirejanja IP-naslovov ob ponovnih obiskih dodeljeno drugačno IP-številko (Srivastava et al., 2000, str. 14). Kombinacija IP-številke s podatkom o uporabniškem agentu natančnost načeloma izboljša, vendar pa je raznolikost uporabniških agentov premajhna, da bi se kakovost prepoznave bistveno izboljšala, hkrati pa postaja predpostavka, da uporabniki za brskanje po internetu vedno uporabljajo isti brskalnik (Berendt, Spilipoulou, 2000, str. 57), zaradi pojavljanja novih različic in tipov brskalnikov vse bolj šibka. Pirolli, Pitkow in Rao (1996) zato predlagajo nadgradnjo algoritma z navigacijskim algoritmom, ki temelji na predpostavki, da uporabniki do strani na spletišču dostopajo s klikanjem na povezave, ne pa tako, da vpisujejo URL-naslove strani. Algoritem uporablja podatke o napotitelju (ang. refferer) in topologijo spletišča, da ugotovi, ali je do izbrane strani mogoče priti prek strani, ki so bile v okviru seje že obiskane. Cooley et al. (1999) predlagajo enostavnejšo različico tega algoritma; pri tej različici se zahtevek pripiše obstoječemu uporabniku oz. seji, če je napotitelj v množici strani, ki so bile v okviru seje že ogledane. Ena izmed osnovnih ugotovitev navigacijskih algoritmov je tudi ta, da zahtevek, ki nima podatka o napotitelju, predstavlja začetek nove seje; to načeloma drži, vendar pa je lahko podatek o napotitelju nedoločen tudi, če uporabnik klikne na gumb »Nazaj« ali pa če spletišče uporablja okvirje, zato Berendt et al. (2001) predlagajo, da se zahtevek, ki nima napotitelja, vendar pa je bil izveden znotraj dovolj

[30] Trajanje seje na strežniku ni povezano z obstojem povezave; seja se konča toliko minut po zadnjem uporabnikovem dejanju, kolikor je določena dolžina seje, pri čemer ni pomembno, ali uporabnik pred tem zapre brskalnik.

majhnega časovnega intervala Δ, ne šteje kot začetek nove seje (10 sekund v Spilipoulou et al., 2003). Za razmejevanje sej je bila predlagana tudi uporaba modelov na osnovi statističnega jezika (Huang et al., 2004), ki izhajajo iz teorije informacij in so robustnejši od tradicionalnih načinov, vendar pa zahteva uporaba teh modelov tudi več truda in predpostavk.

Časovno usmerjeni algoritmi na drugi strani za razmejevanje uporabnikov in sej uporabljajo največje predvideno trajanje obiska ali največje trajanje ogleda posamezne strani. Če je čas od začetka ogleda prve strani oz. ogleda posamezne strani daljši od določenega časovnega intervala, se zahtevek samodejno pripiše novemu uporabniku oz. novi seji. Na podlagi empirične raziskave, opisane v Catledge, Pitkow (1995), je povprečni časovni interval med dvema dejanjema uporabnika 9,3 minute. Ob predpostavki, da se večina statistično značilnih dogodkov zgodi v območju do 1,5 standardnega odklona od povprečne vrednosti, je bil za dolžino seje predlagan interval v dolžini 25,5 minute; v večini standardnih aplikacij se na tej osnovi za dolžino obiska uporablja približek, ki znaša 30 minut (Berend et al., 2001, str. 8), za dolžino ogleda strani pa približek 10 minut (Spilipoulou et al., 2003). Primerna dolžina obiska in ogleda strani je sicer močno odvisna od strukture, vsebine in namena spletišča, zato je treba pred določitvijo teh parametrov proučiti, kako uporabniki uporabljajo spletišče. Pri določanju dolžine ogleda strani je smiselno razlikovati tudi kategorije strani, saj ogled strani, namenjene poglobljenemu branju, traja dlje kot ogled strani, katere namen je preusmerjanje na druge vire (za primer kategorizacije strani glej Cooley, Mobasher, Srivastava, 1999).

Učinkovitost posameznih hevrističnih algoritmov je odvisna od strukture spletišča, načina zbiranja podatkov in namena analize. Izbira pravilnega pristopa, nabora algoritmov in določanje vhodnih parametrov tako še vedno ostaja v domeni strokovnjakov, ki so v konkretnem primeru odgovorni za analizo dnevniških datotek. V raziskavi učinkovitosti posameznih reaktivnih strategij za razmejevanje sej ob predpostavki, da so uporabniki že identificirani (Spilipoulou et al., 2003), je bilo ugotovljeno, da so v tem primeru časovne hevristike primernejše od navigacijskih; hkrati avtorji ugotavljajo, da so slednje uporabnejše, če za identifikacijo uporabnikov ne moremo uporabiti proaktivnih strategij. Ponujajo sicer nabor kategoričnih in postopnih meritev za merjenje učinkovitosti in napak posameznih hevristik, žal pa v svoji raziskavi učinkovitost izbranih algoritmov preučujejo samostojno in ne predvidevajo kombiniranja pristopov. Da bi bila izbira še težja, večina avtorjev algoritme in njihovo uporabnost analizira le v zvezi z identifikacijo uporabnikov ali v zvezi z identifikacijo sej (npr., Pitkow, 1997; Pierrakos et al., 2003; Mobasher, 2005; Cooley, Mobasher, Srivastava, 1999) ali pa se povsem osredotočajo na algoritme za razmejevanje sej ob predpostavki, da so uporabniki že identificirani (Spilipoulou et al., 2003; Berendt et al., 2002). Konkretnih podatkov o najprimernejšem celovitem pristopu tako nimamo, prav tako pa nam niso na voljo informacije o algoritmih, ki jih za ta namen uporabljajo naprednejša komercialna orodja za analizo obiska (in naj bi bili »izredno natančni«; glej npr. NetGenesis, 2007; Iprom, 2006). Vseeno lahko trdimo, da najboljše delujejo algoritmi, ki izkoriščajo proaktivno zbrane podatke, kadar teh ni pa uporabijo pristop, ki je v dani situaciji

najprimernejši. Primer algoritma, ki ga lahko uporabimo za splošno identifikacijo sej, podajam v poglavju 5.3.

3.2.3 Identifikacija transakcij

Za nekatere algoritme je množica ogledov strani v okviru seje preširok koncept za učinkovito prepoznavanje vzorcev uporabe, saj lahko uporabnik v okviru enega obiska spletišča zasleduje več ciljev in opravlja več različnih nalog. Po zgledu nekaterih drugih aplikacij podatkovnega rudarjenja (predvsem analize nakupne košarice) je bilo zato predlagano nadaljnje razmejevanje sej v semantično pomembne enote, imenovane transakcije (Chen, Park, Yu, 1996); te se ujemajo z definicijo epizode po WCA.

Identifikacija transakcij večinoma temelji na modelu uporabniškega vedenja, pri katerem so za odkrivanje asociacijskih pravil in vzorcev zaporedij dejansko pomembne samo vsebinske strani (ang. content page), tj. strani, na katerih obiskovalci najdejo informacije, ki jih zanimajo. Ostale strani uporabniku le pomagajo pri iskanju informacij in jih lahko imenujemo pomožne strani (ang. auxiliary page) (Cooley, Mobasher, Srivastava, 1999, str. 13). Kategorizacija strani na vsebinske in pomožne v okviru spletišča ni absolutna, saj je lahko stran, ki je za nekega uporabnika pomožna, za drugega vsebinska in obratno. Dinamična identifikacija transakcij zato temelji na predpostavkah o tem, kako posamezni uporabnik dojema posamezni tip strani. Chen, Park in Yu (1996) predlagajo pristop na podlagi največje napredujoče reference (ang. maximal forward reference), pri kateri je vsaka transakcija definirana kot množica strani od začetka obiska do vključno strani, za katero se zgodi vzvratna referenca (ang. backward reference); pri tem je napredujoča referenca definirana kot stran, ki je še ni v množici strani trenutne transakcije. Za največjo slabost pristopa se omenja dejstvo, da ni nujno, da so v dnevniških datotekah zapisane vse vzvratne reference (Pierrakos et al., 2003, str. 331; Huang et al., 2004, str. 1291), vendar lahko to dokaj učinkovito razrešimo z že omenjenim (delnim) razbijanjem predpomnjenja ali s spremljanjem uporabe na strani odjemalca. Analitičen pogled razkrije drugo, pomembnejšo pomanjkljivost, in sicer da identifikacija transakcij na podlagi največje napredujoče reference temelji na zelo poenostavljenem razumevanju načina navigacije po spletišču, pri katerem obiskovalec natančno pozna pot do želene vsebinske strani in se na že ogledane strani ne vrača zato, da bi našel drugo (pravilno) pot do iskane informacije, ampak zato, da bi našel še kako drugo informacijo.

Drugačen je pristop na podlagi dolžine reference (ang. reference length) (Cooley, Mobasher, Srivastava, 1999), ki temelji na predpostavki, da je kategorizacija posamezne strani odvisna od tega, koliko časa obiskovalec porabi za ogled strani; kvantitativne raziskave namreč razkrijejo, da je čas ogleda pomožnih strani bistveno krajši od časa ogleda vsebinskih strani (Cooley, Mobasher, Srivastava, 1999; Byrne et al., 1999). Če lahko ocenimo, kolikšen je

delež pomožnih strani med vsemi stranmi na spletišču, lahko na osnovi histograma izračunamo dolžino reference, ki predstavlja oceno optimalne mejne vrednosti za razločevanje pomožnih in vsebinskih referenc (Huang et al., 2004, str. 1291). Tudi ta pristop ima svoje slabosti, predvsem pa natančnost zmanjšuje dejstvo, da lahko motnje pri brskanju po spletišču, ki jih povzročajo zunanji dejavniki (npr. odhod uporabnika na kosilo), povzročijo napačno kategorizacijo referenc (Pierrakos et al., 2003, str. 330). Avtorji to slabost priznavajo, vendar pa menijo, da je malo verjetno, da bi se za neko stran taka napaka redno ponavljala, in dodajajo, da bi jih lahko pri aplikaciji algoritmov podatkovnega rudarjenja izločili z ustrezno nastavitvijo parametra minimalne podpore (ang. support) (Cooley, Mobasher, Srivastava, 1999, str 23). Druga težava je v tem, da je treba za pravilno kategorizacijo dokaj natančno oceniti delež pomožnih referenc na spletišču, kar ni lahka naloga. Woon, Ng in Lim (2005) so zato predstavili nov pristop, temelječ na dolžini reference, pri katerem kategorizacija posamezne reference pri posameznem uporabniku temelji na primerjavi časa ogleda strani s povprečnim časom ogleda vseh strani v okviru obiska; če je čas ogleda krajši od povprečja, se referenca kategorizira kot pomožna, drugače pa kot vsebinska.

3.2.4 Druga opravila

Med obvezne naloge v okviru predobdelave podatkov spadajo tudi postopki predobdelave vsebine in strukture, saj te podatke potrebujemo za učinkovito izvajanje vseh drugih nalog v tej fazi, še posebej pa pri identifikaciji uporabnikov, sej in transakcij. Predobdelava vsebine je v tej zvezi postopek preoblikovanja podatkov v merljivo obliko, ki jo lahko obdelamo z izbranimi algoritmi podatkovnega rudarjenja (Srivastava et al., 2000, str. 14); za ta namen se večinoma uporabljajo pristopi na podlagi modela vektorskega prostora (ang. vector space model).

Odvisno od strukture spletišča, načina zbiranja podatkov in namena analize lahko v okviru predobdelave podatkov izvedemo še dodatne postopke, kot so npr. zaključevanje poti, oblikovanje podatkov ali izračun različnih meritev. Zaključevanje poti (ang. path completion) je postopek, v katerem zapolnimo manjkajoče korake v navigacijskih poteh, do katerih lahko pride, če spletišče ne uporablja postopkov za razbijanje predpomnjenja. Za ta namen se uporabljajo podobni postopki kot za identifikacijo uporabnikov in sej, tj. postopki na podlagi analize podatkov o napotiteljih in topologije spletišča, pri čemer čas in trajanje obiska na strani ocenimo z dodatnimi algoritmi (Cooley, Mobasher, Srivastava, 1999, str. 19). Oblikovanje podatkov in njihovo združevanje se večinoma nanaša na različne sintaktične spremembe; te ne spreminjajo njihovega pomena, ampak so potrebne zaradi značilnosti uporabljenih orodij in algoritmov (Chapman, 1999, str. 52).

3.3 ODKRIVANJE VZORCEV UPORABE

Rezultat predobdelave podatkov sta množica *n* ogledov strani $P = \{p_1,p_2,...p_n\}$ in množica *m* transakcij $T = \{t_1,t_2,...,t_m\}$, pri čemer je vsak t_i podmnožica p_i. Ogledi strani so semantično pomembne enote, ki jih analiziramo v procesu podatkovnega rudarjenja, transakcijo pa si lahko predstavljamo kot zaporedje urejenih parov dolžine *l*: $t=\langle(p_1,w(p_1)), (p_2,w(p_2)), ..., (p_l,w(p_l))\rangle$, kjer p_i predstavlja ogled strani, $w(p_i)$ pa utež, ki določa pomembnost ogleda v dani transakciji – vrednost uteži večinoma temelji na dolžini ogleda strani ali pa je predstavljena v binarni obliki, kjer je utež ogledanih strani enaka 1, utež ostalih strani pa je enaka 0. Tako pojmovanje nam omogoča, da transakcije predstavimo kot vektorje v *n*-dimenzionalnem prostoru ogledov strani: $\vec{t} =\langle(w_{p1},w_{p2},...w_{pn}\rangle$, kjer v primeru, da je bila stran ogledana v okviru transakcije, za vsak $i \in \{1, 2,..., n\}$ velja $w_{pi}=w(p_i)$, za ostale strani pa velja $w_{pi}=0$. Množico vseh transakcij si lahko tako predstavljamo kot matriko *TP* velikosti $m \times n$ (Mobasher, Dai, 2004, str. 291). Taka predstavitev transakcij je osnova za številne tehnike neusmerjenega odkrivanja znanja, kot sta npr. oblikovanje gruč ali odkrivanje asociacijskih pravil, s katerimi lahko odkrijemo pomembne uporabniške segmente ali pa na podlagi navigacijskih vzorcev najdemo pomembne povezave med posameznimi predmeti na strani[31].

3.3.1 Klasificiranje

Klasifikacija je ena izmed najpogostejših nalog v okviru rudarjenja po podatkih, med najbolj znanimi primeri uporabe klasifikacije v poslovnem svetu pa so analiza izgubljanja strank, prepoznavanje zlorab, ciljno oglaševanje idr. Pogostost te oblike rudarjenja lahko po eni strani pripišemo načinu človeškega razumevanja sveta, za katero je zelo pomembno nenehno razvrščanje, kategoriziranje in ocenjevanje, po drugi strani pa tudi dejstvu, da lahko zelo veliko problemov podatkovnega rudarjenja preoblikujemo v klasifikacijski problem.

Vsebinsko gledano pri klasificiranju domnevamo, da obstaja nabor objektov, ki jih označujejo določeni atributi oz. značilnosti in pripadajo različnim razredom (ang. class). Oznake razredov so diskretne spremenljivke, za vsak objekt pa vemo, kateremu razredu pripada. (Chapman, 1999, str. 74). Na tej osnovi je klasificiranje proces izgradnje klasifikacijskih modelov (imenovanih tudi klasifikatorji, ang. classifiers), ki vrednost razrednega atributa (ang. class attribute) določajo kot funkcijo ostalih (vhodnih) atributov (Tang, MacLennan, 2005, str. 6) in jih lahko uporabljamo za napovedovanje razreda objektov, pri katerih vrednost razrednega atributa (še) ni poznana. Klasificiranje je tehnika usmerjenega odkrivanja znanja, saj temelji na razvrščanju objektov v razrede, ki smo jih identificirali že v učni množici podatkov, in se kot tako konceptualno razlikuje od razvrščanja v skupine. To znanje lahko

[31] Predmet (ang. item) na spletni strani je lahko ogled strani, posamezna stran, obisk ipd.

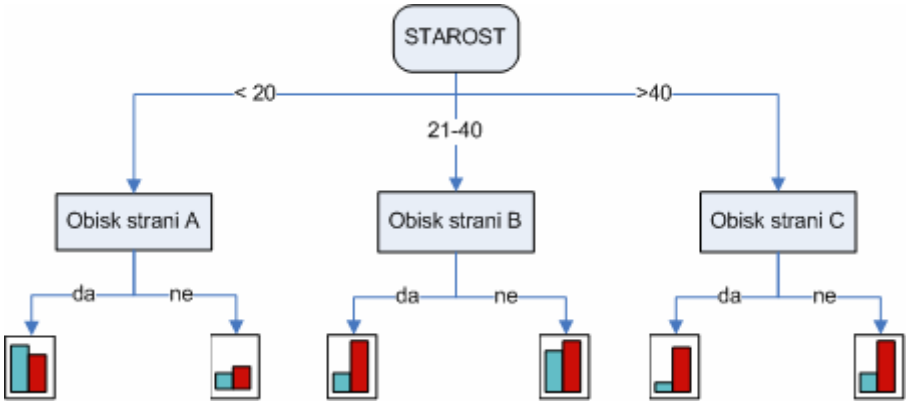
uporabimo tako za razumevanje obstoječih podatkov kot tudi za napovedovanje vedenja prihodnjih primerov (Pierrakos et al., 2003, str. 337). V praksi se vseeno zgodi, da bi želeli namesto razredov napovedovati zvezne (numerične) vrednosti podatkovnih atributov (npr. dohodek strank) – za izdelavo takih modelov se uporabljajo tehnike regresijske analize, postopek pa se za razliko od klasifikacijskega napovedovanja imenuje preprosto napovedovanje (ang. prediction) ali ocenjevanje.

V okviru spletnega poosebljanja se lahko klasificiranje uporablja za modeliranje vedenja obiskovalcev in značilnosti spletnih strani ali za razvrščanje strani in uporabnikov v posebne skupine, pomembne z vidika poosebljanja (glej npr. Srivastava et al., 2000; Chan, 2000; Pierrakos et al., 2003; Smyth, Berry, 2004), lahko pa se uporablja tudi pri določanju manjkajočih demografskih atributov obstoječih uporabnikov (Murray, Durrell, 2000). Med klasifikacijskimi tehnikami so se za potrebe poosebljanja doslej največkrat uporabljala odločitvena drevesa in Bayesovi klasifikatorji[32], vendar pa je njihova uporaba veliko bolj omejena, kot je uporaba neusmerjenih tehnik; pri podatkih, povezanih s spletom (npr. transakcije ali strani), namreč pogosto ni mogoče zagotoviti množice vnaprej klasificiranjih podatkov (Pierrakos et al., 2003, str. 337).

Odločitvena drevesa so grafikonom poteka podobne strukture (ang. flow-chart), pri katerih vsako vozlišče (ang. node) predstavlja test vrednosti atributa, vsaka veja (ang. branch) rezultat testa, posamezni listi (ang. leaves) pa predstavljajo razrede ali porazdelitve razredov. Njihova osnovna ideja je rekurzivno razdeljevanje podatkov v podmnožice, tako da vsaka podmnožica vsebuje bolj ali manj homogene vrednosti ciljne spremenljivke, tj. razrednega atributa, pri čemer se pri vsaki razdelitvi drevesa vse vhodne spremenljivke oceni glede na njihov vpliv na razredni atribut (Tang, MacLennan, 2005, str. 146). Rezultat izgradnje odločitvenega drevesa je množica pravil *če-potem* (ang. if-then rules), ki predstavljajo možne poti od korenskega vozlišča do končnih listov in jih lahko zelo preprosto uporabimo za klasificiranje in napovedovanje (Han, Kamber, 2000, str. 290); na sliki 1 je prikazan primer odločitvenega drevesa, ki kaže razdelitev uporabnikov po spolu glede na njihovo starost in podatek o zadnji obiskani strani (modri stolpci v listih označujejo moške, rdeči pa ženske).

[32] Med ostale tehnike, ki jih lahko uporabljamo za klasificiranje in napovedovanje prištevamo nevronske mreže, regresijsko in diskriminacijsko analizo, genetske algoritme (angl. genetic algorithms), tehnike iskanja najbližjih sosedov (angl. nearest neighbour classifiers) in sklepanja na podlagi primerov (angl. case-based reasoning) ter pristope na osnovi mehkih in neobdelanih množic (angl. fuzzy sets, rough sets) (Chapman et al., 1999; Han, Kamber, 2000).

Slika 1: Primer odločitvenega drevesa



Vir: prirejeno po Tang, MacLennan, 2005, str. 150

Podobno enostavna je uporaba Bayesovih klasifikatorjev, ki jih uporabljamo za napovedovanje verjetnosti razredne pripadnosti in temeljijo na Bayesovem teoremu; ta določa ogrodje za naknadno (a posteriori) revizijo znanja in pravil na osnovi novih spoznanj (Bayes' theorem, 2007). Poenostavljeno rečeno teorem pravi, da lahko verjetnost hipoteze *A* v primeru, da imamo na voljo podatke o hipotezi *B*, izračunamo na osnovi podatkov o verjetnosti posamezne hipoteze in verjetnosti hipoteze B ob dani hipotezi A. V okviru podatkovnega rudarjenja se večinoma uporablja poenostavljena različica teorema, imenovana naivni Bayesovi klasifikatorji (ang. naive Bayesian classifiers), pri katerih predvidevamo, da je učinek posameznega atributa na posamezni razred neodvisen od vrednosti ostalih atributov, kar imenujemo pogojna verjetnost razreda (Han, Kamber, 2000, str. 296). Naivni klasifikatorji so bolj enostavni za izračun, vseeno pa je njihova učinkovitost podobna učinkovitosti odločitvenih dreves in nevronskih mrež.

Med bolj znane primere uporabe tehnik odločitvenih dreves in naivnih Bayesovih klasifikatorjev pri spletnem poosebljanju spada poskus izračunavanja cenilk zanimivosti strani (ang. Page Interest Estimator, PIE) (Chan, 2000), ki se jih lahko uporablja za pomoč pri izbiranju strani, ki se jih priporoča posameznemu uporabniku. Pri teh cenilkah je funkcija zanimivosti strani za posameznega uporabnika odvisna od absolutnega in relativnega števila njegovih obiskov na določeni strani, trajanja in svežosti obiska ter binarnega atributa, ki označuje, ali ima uporabnik stran shranjeno med priljubljenimi stranmi.

Za potrebe spletnega poosebljanja je klasificiranje uporabno tudi v številnih nalogah predobdelave vsebine, saj lahko s temi tehnikami učinkovito razvrščamo spletne strani glede na njihov tip ali namen uporabe (npr. navigacijske in vsebinske strani), lahko razvrščamo posamezne seje ali transakcije, lahko pa klasificiranje uporabimo tudi za razvrščanje uporabnikov v skupine, ki jih organizacija določi z vidika svojega poslovnega modela. V povezavi z uporabniki lahko tehnike klasificiranja uporabimo tudi za napovedovanje vrednosti demografskih atributov uporabnikov ali za določanje manjkajočih podatkov.

3.3.2 Razvrščanje v skupine

Razvrščanje v skupine (ang. clustering) spada med neusmerjene tehnike odkrivanja znanja, natančneje med tehnike učenja na podlagi opazovanja, in je največkrat uporabljena metoda za odkrivanje vzorcev iz podatkov o spletu (Pierrakos et al., 2003, str. 333). Namen razvrščanja je opazovane objekte razvrstiti v skupine oz. gruče, tako da so si objekti znotraj gruče med seboj zelo podobni, hkrati pa se kar se da razlikujejo od objektov v drugih gručah.

Osnovne metode za razvrščanje v skupine lahko razdelimo v 3 skupine (prirejeno po Han, Kamber, 2000, str. 346–348), ki se med seboj razlikujejo glede na to, kakšne metode uporabljajo za razvrščanje objektov:

- **Delitvene metode (ang. partitioning methods)[33].** Te metode množico *n* objektov razdelijo na *k* skupin[34]. Metode v tej skupini delujejo tako, da po začetni razvrstiti objektov uporabljajo iterativni pristop, v katerem skušajo delitev optimizirati s premeščanjem objektov med skupinami.

- **Hierarhične metode (ang. hierarchical methods).** Hierarhične metode delujejo tako, da skupino objektov hierarhično razstavijo na manjše skupine. Postopek lahko poteka od spodaj navzgor (ang. bottom-up approach, tudi agglomerative approach) ali od zgoraj navzdol (ang. top-bottom approach, tudi divisive approach).

- **Metode, ki temeljijo na modelu (ang. model-based methods).** Pri metodah, ki temeljijo na modelu, se za vsako skupino postavi model, na osnovi katerega se skupini dodelijo ustrezajoči objekti.

Metode razvrščanja v skupine lahko razdelimo tudi v trde in mehke (Tang, MacLennan, 2005, str. 191). Pri trdem razvrščanju (ang. hard clustering) se za razvrščanje največkrat uporabljajo mere razdalje, meje med posameznimi skupinami pa so trdno določene, kar pomeni, da posamezni objekt pripada natanko eni skupini. Tipični predstavnik te skupine algoritmov je *k-means* algoritem; pri njem je objekt dodeljen tisti skupini, katere središče (centroid) mu je na podlagi preproste evklidske razdalje najbližji. Na drugi strani pri metodah mehkega razvrščanja le-to ne temelji na razdalji, ampak na verjetnostnih merah; to omogoča, da je objekt z različno verjetnostjo dodeljen več skupinam hkrati. Med slednje metode se uvršča tudi algoritem EM (Expectation Maximization), katerega delovanje temelji na predpostavki, da so vrednosti v vsaki dimenziji (tj. vrednosti vsakega atributa) normalno porazdeljene; če

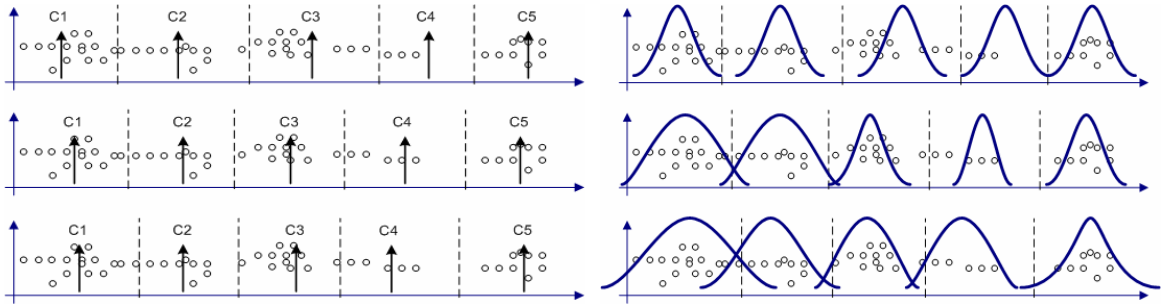
---

[33] V to skupino metod lahko uvrščamo tudi metode, ki temeljijo na gostoti, in metode, ki temeljijo na mreži, čeprav so lahko obravnavane tudi kot samostojne skupine metod.

[34] Za delitev se največkrat uporablja razdalja med objekti. Poleg evklidske razdalje (angl. euclidean distance) spadajo med najpogosteje uporabljene metode za izračun razdalje tudi kosinusna podobnost (angl. cosine similarity) ter evklidska in Pearsonova korelacija (angl. Pearson correlation) (Mobasher, Dai, 2005, str. 299).

pade točka znotraj krivulje, se pripiše skupini z verjetnostjo, ki je določena s porazdelitvijo. Ker se lahko krivulje posameznih skupin prekrivajo, je lahko tako objekt razvrščen v več skupin hkrati (Tang, MacLennan, 2005, str. 192). V zvezi s spletnim poosebljanjem se večkrat pojavlja mnenje, da je primernejše mehko razvrščanje, saj se lahko uporabniki oz. strani, ki so predmet razvrščanja, nahajajo v več skupinah (npr. Paliouras et al., 2000; Mobasher et al., 2002; Perkowitz, Etzioni, 2000; Pierrakos, 2003).

Slika 2: Primerjava trdega razvrščanja v skupine in mehkega razvrščanja



Vir: Tang, MacLennan, 2005, str. 191; Sarka, 2006

Pri spletnem poosebljanju se razvrščanje v skupine največkrat uporablja za iskanje skupin podobnih uporabnikov in za iskanje skupin podobnih strani. Za razvrščanje uporabnikov so zelo uporabni vektorji transakcij $\vec{t}$, definirani v začetku tega poglavja (matrika TP), ki predstavljajo strani, ki so jih uporabniki obiskali v okviru posameznega obiska (oz. transakcije); gruče transakcijskih vektorjev (tudi transakcijske gruče), ki so si med seboj podobni, na tej osnovi predstavljajo uporabnike ali skupine uporabnikov, ki izkazujejo podobno navigacijsko vedenje ali so si podobni na podlagi katerih drugih lastnosti, ki so bile izražene s transakcijskimi vektorji (Mobasher, 2005, str. 20). Pogled na transakcije lahko tudi obrnemo in transponiramo matriko TP in tako dobimo vektorje strani, ki so bile obiskane v posamezni transakciji (s strani posameznega uporabnika); z razvrščanjem teh vektorjev v skupine lahko dobimo skupine strani, ki so si podobne glede na to, kateri uporabniki jih obiskujejo, vendar pa je število dimenzij v tem primeru za večino tradicionalnih algoritmov preveliko, zato je pri tem pristopu potrebno predhodno zmanjšanje razsežnosti (glej npr. algoritem ARHP v Mobasher et al. (2002)).

Strani lahko razporejamo v gruče tudi na podlagi njihove vsebinske podobnosti; uteži vektorja p v tem primeru predstavljajo pogostnosti pojavljanja posameznih vsebinskih značilnosti[35] na strani p, pri čemer je vrednost uteži običajno normalizirana TF.IDF[36] vrednost posamezne besede. Tako dobljene vsebinske profile strani lahko kombiniramo s profili uporabe in jih

[35] Vsebinska značilnost je skupni izraz za ključne besede, fraze, imena kategorij in druge dele besedila.

[36] Po TF.IDF shemi je vrednost uteži besede določena funkcija pogostosti besede v besedilu (term frequency - TF) in števila dokumentov, ki vsebujejo besedo (inverse document frequency – IDF). Ta shema daje večjo vrednost besedam, ki se pogosto pojavljajo v manjšem številu dokumentov.

uporabimo za učinkovitejše poosebljanje. Z uporabo semantičnega znanja, ki je del vsebinskih profilov, lahko npr. tudi učinkovito rešimo *težavo novega predmeta* (ang. new item problem), značilno za pristope na podlagi sodelovanja in za pristope, pri katerih se uporabljajo samo profili uporabe (Mobasher, Dai, 2005, str. 287): čeprav se stran še ne pojavlja v profilih uporabe, jo lahko zaradi njene semantične povezanosti s »starejšimi predmeti« vseeno priporočamo uporabnikom.

Razvrščanje transakcij v gruče je zelo uporabno za odkrivanje skupin uporabnikov, vendar pa transakcijske gruče same po sebi niso učinkovit način za predstavljanje agregatnega pogleda na splošne uporabniške značilnosti; v njih je lahko zbranih na tisoče transakcij s številnimi ogledi strani, to pa močno otežuje možnost njihove analize za namen pridobivanja novega znanja ali izvedbo nalog, kot je spletno poosebljanje. Težavo lahko enostavno rešimo tako, da za vsako gručo izračunamo njen centroid ali t. i. povprečni transakcijski vektor, pri katerem so vrednosti posameznih dimenzij (ogledov strani) izračunane kot povprečne vrednosti dimenzij vseh vektorjev v gruči. Če so uteži v transakcijskih vektorjih binarne, potem dimenzijska vrednost ogleda *p* v centroidu *c* predstavlja delež transakcij v gruči, v katerih se pojavlja ogled, s tem pa tudi relativno pomembnost ogleda. Oglede strani v centroidu lahko nato razvrstimo glede na te vrednosti in izločimo manj značilne oglede ter tako dobimo skupinski profil uporabe (ang. group usage profile), ki predstavlja interese ali vedenje uporabnikov v gruči (glej sliko 3).

Slika 3: Primer izdelave skupinskih profilov uporabe (pu_c) iz transakcijskih gruč

		p_0	p_1	p_2	p_3	p_4	p_5
Gruča 0	Transakcija 1	0	0	1	1	0	0
	Transakcija 4	0	0	1	1	1	0
	Transakcija 7	0	0	1	1	0	0
Gruča 1	Transakcija 0	1	1	0	0	0	1
	Transakcija 3	1	1	0	0	0	1
	Transakcija 6	1	1	0	0	0	1
	Transakcija 9	0	1	1	0	0	1
Gruča 2	Transakcija 2	1	0	0	1	1	0
	Transakcija 5	1	0	1	1	1	0
	Transakcija 8	1	0	1	1	1	0

Spletni pogledi

Profili uporabe	p_0	p_1	p_2	p_3	p_4	p_5
pu_0	0,0	0,0	1,0	1,0	0,3	0,0
pu_1	0,8	1,0	0,3	0,0	0,0	1,0
pu_2	1,0	0,0	0,7	1,0	1,0	0,0

Vir: prirejeno po Mobasher, 2005

Ker je razvrščanje v skupine neusmerjeno, je težko oceniti, kateri algoritem in pripadajoči nabor parametrov sta v posamezni situaciji vsebinsko najboljša. Poleg tega je bilo malo storjenega tudi na področju primerjave njihovih učinkov, kar lahko pripišemo pomanjkanju objektivnih meril, neodvisnih od področja uporabe. Načeloma lahko za ocenjevanje kakovosti modelov, dobljenih pri razvrščanju, uporabimo dva pristopa: rezultate lahko primerjamo s podatki z znano strukturo ali jih uporabimo pri nalogah, katerih rezultati so vnaprej znani

(npr. za napovedovanje) (Pierrakos et al., 2003), lahko pa njihovo kakovost ocenjujemo tudi na osnovi njihovih želenih lastnosti[37]. Paliouras et al. (2000) za ta namen predlagajo dve osnovni meritvi: *raznolikost* (ang. distinctiveness), ki kaže, kako močno se modeli med seboj razlikujejo, ter *pokritje* (ang. coverage), ki pove, koliko strani na spletišču je vključenih v model. Podobno bi lahko za primerjavo kakovosti uporabili statistične meritve, kot so npr. skupna entropija gruč ali standardni odklon in povprečna verjetnost pripadnosti gručam (Sarka, 2006), čeprav se pojavlja mnenje, da so te mere preveč podobne statističnim merilom, ki jih uporabljajo algoritmi sami, in so zato lahko pristranske (Pierrakos et al., 2003, str. 336).

3.3.3 Asociacijska pravila

Asociacijska pravila (ang. association rules) so tehnika za iskanje pogostih vzorcev, asociacij in korelacij med predmeti (ang. item) v opazovani množici. Tehnika odkrivanja asociacijskih pravil je bila zasnovana za analizo nakupne košarice (Agrawal, Imielinski, Swami, 1993), pri rudarjenju po podatkih o uporabi spleta pa lahko ta pravila uporabimo za povezovanje strani, ki se najpogosteje pojavljajo skupaj v posamezni strežniški seji. Metoda asociacijskih pravil se največkrat uporablja za napovedovanje najzanimivejše naslednje strani pri posameznem uporabniku v priporočilnih sistemih (glej npr. Lin, Alvarez, Ruiz, 2002; Nakagawa, Mobasher, 2003), uporabna pa so tudi kot vodilo pri prestrukturiranju in organizaciji spletišča (npr. tako da dodamo povezave med stranmi, ki so pogosto obiskane v posamezni transakciji) (Spilipoulou, Pohle, 2001; Cooley, Mobasher, Srivastava, 1997), ali za izboljšanje odzivnosti sistema prek vnaprejšnjega nalaganja spletnih strani in podatkov (Eirinaki, Vazirgiannis, 2003, str. 10).

Asociacijsko pravilo je pravilo v obliki $X \Rightarrow Y$, kjer X in Y predstavljata podmnožici predmetov (ang. itemset) v transakciji T in jih imenujemo tudi telo pravila (ang. body) (X) in glava pravila (ang. head) (Y) (Lin, Alvarez, Ruiz, 2002, str. 88). Pomen pravila je v tem, da navzočnost vseh predmetov v X z določeno verjetnostjo namiguje na navzočnost vseh predmetov v Y. Veljavnost vsakega asociacijskega pravila lahko ocenimo s podporo (ang. support) in zaupanjem (ang. confidence), pri čemer podpora, $\sigma_r = \sigma(X \cup Y)$, predstavlja verjetnost, da se X in Y v posamezni transakciji pojavljata skupaj, zaupanje $\alpha_r = \sigma(X \cup Y)/\sigma(X)$ pa predstavlja pogojno verjetnost navzočnosti Y v posamezni transakciji v primeru navzočnosti množice X ($P(Y|X)$). Meri podpore in zaupanja nam omogočata, da iz nabora odkritih pravil izločimo samo najuporabnejša pravila. Hkrati si z njima pomagamo tudi pri omejevanju iskalnega prostora pravil, katerih število raste eksponentno s številom predmetov, in tako zmanjšujemo računsko kompleksnost algoritmov.

---

[37] Npr. homogenost znotraj gruče, heterogenost med gručami, število gruč, povprečna velikost gruč (Paliouras et al., 2000).

Težava z uporabo merila minimalne podpore je v tem, da odkrite povezave večkrat ne vsebujejo redkih, a vseeno pomembnih predmetov; prav pri podatkih o uporabi spletišča se pogosto zgodi, da so strani na globlji ravni veliko manj obiskane kot strani na višjih ravneh. Težavo lahko deloma odpravimo z določanjem različnih vrednosti podpore za različne predmete (Mobasher, 2005, str 16). Dodatno težavo predstavlja določitev take vrednosti podpore, da dobimo želeno število uporabnih pravil. Lin, Alvarez in Ruiz (2002) predlagajo za ta namen uporabo algoritma ASARM (Adaptive-Support Association Rule Mining), saj ta ob dani meri zaupanja najde želeno število pravil z največjo podporo.

3.3.4 Odkrivanje zaporednih vzorcev

Asociacijska pravila v spletnem okolju niso najustreznejša tehnika odkrivanja znanja, saj pri analizi povezanosti med predmeti ne upoštevajo časovne razsežnosti (tj. zaporedja predmetov v posamezni transakciji), ki je še posebej pomembna za analizo navigacijskih vzorcev (ang. navigation pattern) ali analizo poti prečenja (ang. traversal path). Tehnike odkrivanja zaporednih vzorcev (ang. sequence pattern discovery) v tem smislu predstavljajo nadgradnjo tehnik odkrivanja asociacijskih pravil in omogočajo natančnejše rezultate pri aplikacijah, kot so npr. priporočilni sistemi ali sistemi za prednalaganje vsebine. Osnovna težava pri tem je velika množica različnih možnih poti skozi spletišče.

Zaporedni vzorci imenujemo tista zaporedja predmetov, ki se pojavljajo v dovolj velikem številu transakcij. Zaporedje $<s_1, s_2, \ldots, s_n>$ obstaja v transakciji $t = <p_1, p_2, \ldots, p_m>$ (kjer je $n \leq m$), če obstaja n pozitivnih celih števil, tako da velja: $1 < a_1 < a_2 < \ldots < a_n \leq m$ in za vsak i velja $s_i = p_{ai}$. Pravimo, da je zaporedje $<cs_1, cs_2, \ldots, cs_n>$ stično zaporedje (ang. contiguous sequence) v transakciji, če obstaja pozitivno celo število $0 \leq b \leq m - n$ in velja $cs_i = p_{b+1}$ za vse $i = 1$ do n . V stičnem zaporednem vzorcu se mora v transakciji, ki podpira vzorec, vsak par sosednjih elementov (s_i in s_{i+1}) pojaviti zaporedno, medtem ko lahko zaporedni vzorec predstavlja nestična pogosta zaporedja v dani množici transakcij. Zaporedje s k elementi imenujemo tudi k-zaporedje (Agrawal, Srikant, 1995, str. 5).

Metode za odkrivanje zaporednih vzorcev lahko razdelimo v dve skupini: na deterministične[38], ki temeljijo na spremljanju vseh možnih kombinacij navigacijskih poti, in stohastične[39], pri katerih se zaporedje že obiskanih spletnih strani uporablja za napovedovanje najverjetnejših naslednjih obiskov na osnovi verjetnosti (Pierrakos et al., 2003, str. 343).

---

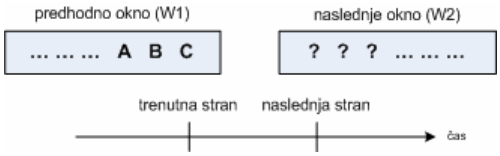
[38] Deterministični modeli (sistemi) so modeli, katerih delovanje je vnaprej določljivo, kar pomeni, da dajo ob danih podatkih vedno enak rezultat.

[39] Stohastični modeli (sistemi) so nedoločljivi modeli, ki na enake vhode ne reagirajo vedno z enakimi izhodi. Vedenje stohastičnih modelov opisujemo z verjetnostjo možnih reakcij na določene vplive.

3.3.4.1 Deterministične metode

Skupna značilnost determinističnih metod je v tem, da na zaporedne vzorce gledajo kot na poseben primer strožje definiranih asociacijskih pravil, kjer ni pomembna zgolj navzočnost predmetov v podmnožici, ampak tudi njihov vrstni red. Kot taki zadovoljujejo tudi merilo zapiranja navzdol (ang. downward closure) (Mobasher, 2005, str. 17), zato lahko za njihovo odkrivanje uporabimo algoritem *Apriori*, pri katerem spremenimo definicijo podpore tako, da temelji na pogostosti pojavljanja zaporedij, in ne le podmnožic predmetov (Agrawal, Srikant, 1995). Vendar pa ti algoritmi v osnovi niso zasnovani za uporabo v spletnem okolju, zato jih ne moremo neposredno uporabiti za napovedovanje. Za napovedovanje jih je potrebno namreč najprej pretvoriti v razvrščevalce (ang. classifiers), kar pomeni, da se moramo pri vsakem posameznem opazovanju znati odločiti, katerega izmed možnih vzorcev izbrati za napovedovanje dejanj, ki sledijo. Yang, Li in Wang (2004) so za oblikovanje optimalnega napovedovalnega modela sistematično analizirali različne metode za gradnjo napovedovalnih modelov na osnovi asociacijskih pravil, pri čemer so podrobno raziskovali dve pomembni dimenziji pri gradnji takih modelov: vrsto predhodnikov v pravilih (ang. antecedent) ter merila za izbor napovedovalnih pravil. Za ta namen so definirali dve okni, ki vsebujeta množico strani: predhodno okno (ang. antecedent window), ki vsebuje določeno število strani, ki jih je uporabnik do določene časovne točke že obiskal, in naslednje okno (ang. consequent window), v katerem so strani, ki jih bo obiskal po danem trenutku. Z uporabo pravila premikajočega se okna (ang. moving window) so oblikovali dnevniško tabelo (ang. log table), v kateri predstavlja vsaka vrstica strani, ki jih zajema vsak par premikajočih se oken, število stolpcev pa ustreza velikosti teh oken. Primer premikajočega se okna in dnevniške tabele, ki ustreza paru premikajočih se oken velikosti [3,2], je prikazan na sliki 4.

Slika 4: Premikajoča se okna in dnevniška tabela



W_1			W_2	
A_1	A_2	A_3	P_1	P_2
A	B	C	A	C
B	C	A	C	D
C	A	C	D	G

Vir: Yang, Li, Wang, 2004, str. 257

Analizirali so pet različnih metod za oblikovanje zaporednih asociacijskih pravil v obliki LS→DS[40], kjer množica LS (leva stran) vsebuje strani iz predhodnega okna, množica DS (desna stran) pa strani iz naslednjega okna (Yang, Li, Wang, 2004, str. 258-261):

- **Pravila podmnožice** (ang. subset rule) so pravila iz področja tradicionalnih asociacijskih pravil, pri katerih zaporedje in sosedstvo predmetov niso pomembni. Ta pravila odsevajo razmišljanje, da relativni vrstni red strani za napovedovanje ni vedno

[40] Ang. LHS (Left-Hand Side) → RHS (Right-Hand Side)

pomemben.

- **Pravila zaporedja** (ang. subsequent rule) upoštevajo tudi zaporedje ogledanih strani. Zaporedje v okviru predhodnega okna je oblikovano iz niza strani, ki se pojavljajo v takem vrstnem redu, kot so bile dejansko obiskane. Pri tem ni zahtevana stičnost zaporedja niti ni potrebno, da se zaporedje zaključi v predhodnem oknu.

- **Pravila zadnjega zaporedja** (ang. latest-subsequence rule) ne upoštevajo zgolj vrstnega reda, ampak tudi podatke o času obiska zadnje strani v zaporedju. V tem načinu predstavitve pravil se lahko na levi strani pravila znajdejo samo zaporedja, ki so se do danega trenutka že zaključila.

- **Pravila podniza** (ang. substring rule) so pravila, pri katerih je zahtevana tudi stičnost zaporedja v predhodnem oknu.

- **Pravila zadnjega podniza** (ang. latest-substring rule) so najstrožje definirana pravila in zahtevajo, da so zaporedja v predhodnem oknu stična, hkrati pa se morajo končati v danem trenutku. Na pravila zadnjega podniza lahko gledamo tudi kot na Markove modele reda N, kjer N zajema različne rede do dolžine W_1.

Da bi iz nabora pravil izbrali tisto, ki omogoča najboljšo napoved, lahko tako kot pri asociacijskih pravilih izberemo pravilo z najvišjim zaupanjem. Druga metoda izbora temelji na ugotavljanju dolžine ujemanja in izmed vseh pravil izbere tisto, pri katerem je dolžina LS največja (Pitkow, Pirolli, 1999). Slabost obeh mer je v tem, da zahtevata, da analitik vnaprej določi zahtevano podporo, to pa je težka naloga, saj lahko ob previsoko postavljeni vrednosti podpore prezremo manj pogoste, vendar zelo uporabne vzorce, prenizko postavljena vrednost pa povzroča predobro prileganje (ang. overfitting), pri katerem daje model dobre rezultate samo pri podatkih iz množice za učenje. Zaradi navedenega Yang, Li, Wang (2004) predlagajo novo mero, t. i. pesimistično zaupanje (ang. pessimistic confidence), ki združuje meri zaupanja in podpore ter temelji na opazovani stopnji napake in podpori posameznega pravila in tako odpravlja potrebo po umetnem določanju najnižje vrednosti podpore. Metoda pesimističnega zaupanja je vedno pesimistična glede natančnosti modela, zato vrednost zaupanja izračuna ob predpostavki najvišje možne stopnje napake. Pokažemo lahko, da sta med omenjenimi petimi vrstami pravil najnatančnejši[41] merili zadnjega podniza in zadnjega zaporedja, med merami pa je najnatančnejša mera pesimističnega zaupanja (za podrobno analizo glej Yang, Li, Wang, 2004).

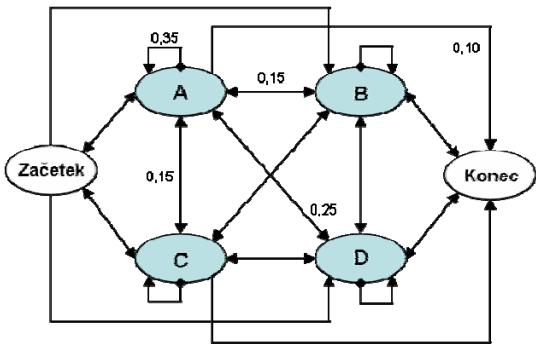
[41] Natančnost je definirana kot delež pravilnih napovedi izmed vseh napovedi.

3.3.4.2 Stohastične metode

Tipični primer stohastične metode je markovski model (tudi markovske verige, ang. Markov chains), ki je najpogosteje uporabljena metoda za odkrivanje zaporednih vzorcev za napovedovanje povezav (ang. link prediction). Osnovna prednost markovskih modelov, ti temeljijo na trdni matematični podlagi, je v tem, da lahko z njihovo pomočjo generiramo navigacijske poti, ki jih lahko brez dodatnega obdelovanja uporabimo za napovedovanje in so tako zelo uporabni za spletno poosebljanje. Njihova največja pomanjkljivost je, da niso enostavno berljivi[42] in nam zato ne omogočajo hitrega vpogleda v uporabo spletišča (Pierrakos, 2003, str. 345), vendar pa je bilo tudi na tem področju opravljenih nekaj raziskav (Cadez et al., 2003, str. 410), na podlagi katerih so se razvili preprosti pregledovalniki vzorcev zaporedij.

Markovska veriga je zaporedje $X_1, X_2, \ldots, X_s$ slučajnih spremenljivk. Množico možnih vrednosti spremenljivke imenujemo prostor stanj (ang. state space), pri čemer X_t označuje stanje procesa v času t. Po definiciji je torej markovska veriga tak diskretni stohastični model, za katerega velja, da je verjetnostna porazdelitev stanj modela v trenutku $t + 1$ odvisna le od stanja modela v času t, ne pa od predhodnih stanj, ki jih je model dosegal na poti do stanja i_t. Poenostavljeno lahko rečemo, da je prehod v novo stanje modela v naslednjem koraku odvisen le od tega, v katerem stanju se model trenutno nahaja, ne pa od tega, kako je do tega stanja prišel (Peterle, 2002, str. 44). Verjetnost prehoda iz stanja i v stanje j ($p_{i,j}$) imenujemo prehodna verjetnost (ang. transition probability), matriko vseh prehodnih verjetnosti pa prehodno matriko (ang. transition probability matrix) (glej sliko 6). Velja, da je verjetnostna porazdelitev stanj v markovski verigi neodvisna od časa – ne glede na to, v katerem trenutku gledamo prehod iz stanja i v stanje j, bo verjetnost vedno enaka (Grinstead, Snell, 1997).

Slika 5: Usmerjeni graf markovske verige



Vir: Tang, MacLennan, 2005, str. 211

[42] Markovske verige lahko sicer predstavimo v obliki usmerjenega grafa (ang. directed graph) (glej sliko 5), kjer robove označimo z verjetnostmi prehoda med vozliščema na posameznih koncih roba, vendar pa je pri velikem številu različnih stanj tak prikaz zelo nepregleden.

Slika 6: Prehodna matrika markovske verige in izračun verjetnosti prehoda iz stanja *j* v stanje *i*

$$\begin{bmatrix} p_{1,1} & p_{1,1} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,1} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix} \qquad p_{i,j} = \sum_{k=1}^{n} p_{i,k} p_{k,j}$$

Vir: Grinstead, Snell, 1997; Tang, MacLennan, 2005, str. 211

V povezavi s spletnimi transakcijami lahko markovske verige uporabimo za modeliranje verjetnosti prehodov med posameznimi stranmi[43], funkcionalno pa predstavljajo enako rešitev kot N–gram modeli[44]. Napovedna natančnost teh modelov je zelo odvisna od števila preteklih stanj, ki so upoštevana pri napovedovanju naslednjega stanja; število upoštevanih stanj določa red (ang. order) markovskega modela oz. vrednost *n* v N-gram modelih[45]. Z modeli višjega reda lahko v splošnem dosežemo večjo natančnost napovedi, vendar pa običajno na račun manjšega pokritja in večje kompleksnosti, ki jo prinaša veliko število možnih stanj. Empirične študije o natančnosti modelov so pokazale, da so modeli višjega reda natančnejši, saj daljše poti vsebujejo več informacij (Zukerman, Albrecht, Nicholson, 1999; Su et al., 2000), vendar pa ta pristop vodi v izredno povečanje računske kompleksnosti. Ena izmed možnih rešitev tega problema je pristop na osnovi markovskih modelov vseh k-redov (ang. all-*k*th-order), ki ga kombiniramo z eno izmed metod za zmanjševanje števila stanj (Pitkow, Pirolli, 1999). Za večino spletišč naj bi sicer najboljše rezultate ponujali modeli reda 3 ali 4 (Su et al., 2000), čeprav so lahko tudi modeli prvega in drugega reda v določenim primerih zadovoljivo natančni (Sen, Hansen, 2000; Pierrakos et al., 2003, str. 345).

Nekoliko drugačen pristop k odkrivanju zaporedij z uporabo markovskih modelov je njihovo kombiniranje z algoritmi na razvrščanje v skupine (Cadez et al., 2003). Razbijanje populacije na skupine, ki ima vsaka svoj markovski model, temelji na podmeni, da so današnji uporabniki spleta preveč raznoliki[46], da bi lahko ob upoštevanju danih omejitev glede kompleksnosti in zahtevane hitrosti razvili markovski model takega reda, ki bi dajal dovolj natančne rezultate. Cadez et al. (2003) predlagajo uporabo na modelu temelječega razvrščanja, pri katerem se vsakega uporabnika ob prihodu na spletišče z določeno verjetnostjo dodeli v eno izmed *k* gruč, nato pa se njegovo navigacijsko vedenje generira iz modela, ki je značilen za to gručo. Če pri tem predpostavljamo, da je vedenje vsakega

---

[43] Vsaka stran predstavlja stanje v okviru zaporedja.

[44] N-gram je podzaporedje *n* stanj v danem zaporedju. Idejo je predstavil Claude Shannon v okviru teorije informacij. Beseda »gram« je okrajšava za »grammatics« (gramatika) in izvira iz dejstva, da so se ti modeli razvijali predvsem v okviru preučevanja besedil in slovničnih značilnosti.

[45] *n*-gram model ustreza markovskemu modelu reda (*n*-1).

[46] Raznolikost obiskovalcev pomeni, da so ti po navigacijskem vedenju zelo različni, zato je število različnih stanj – različnih obiskanih strani – v markovskem modelu izredno veliko.

uporabnika neodvisno od vedenja ostalih uporabnikov[47], potem ta model ustreza t. i. mešanim modelom (ang. mixture model) s *k* komponentami in ga lahko v zgoraj opisanem kontekstu definiramo kot

$$p(x \mid \theta) = \sum_{k=1}^{K} p(c_k \mid \theta) p_k(x \mid c_k, \theta)$$

kjer je $p(c_k|\theta)$ mejna verjetnost gruče *k* (Σ_k $p(c_k|\theta) = 1$), $p_k(x|c_k,\theta)$ je statistični model, ki opisuje spremenljivke za uporabnike v gruči *k*[48], θ pa označuje parametre modela. Če predpostavljamo, da predstavlja *X* multivariatno naključno spremenljivko, ki je naključno dolgo zaporedje spremenljivk, ki opisujejo pot uporabnika skozi spletišče, potem lahko vsako komponento modela predstavimo kot markovski model prvega reda

$$p(x \mid c_k, \theta) = p(x_1 \mid \theta_k^I) \prod_{i=2}^{L} p(x_i \mid x_{i-1}, \theta_k^T)$$

kjer θ^I_k predstavlja parametre verjetnostne porazdelitve prve obiskane strani za uporabnike iz gruče *k*, θ^T_k pa parametre njihove verjetnostne porazdelitve prehodov med stranmi. Za učenje modela Cadez et al. (2003) predlagajo znani dvostopenjski proces maksimiranja pričakovanja (ang. expectation maximization).

Tak model imenujemo mešanica markovskih modelov prvega reda (ang. mixture of first-order Markov models) in ima v primerjavi s »klasičnim« markovskim modelom prvega reda precej večjo napovedno moč, saj je od njega vsebinsko bogatejši, ker lahko prek mehanizma mešanja zajame tudi nekatere odvisnosti višjega reda (Cadez et al., 2003, str. 416). Formalno gledano vodi mešanica markovskih modelov prvega reda do napovedovalne porazdelitve, ki sama po sebi ni markovski model prvega reda, ampak so verjetnosti prehoda pravzaprav funkcije uteži članstva v posameznih gručah; te so nadalje funkcija zgodovine zaporedja in so tako tipično močno odvisne od vzorca vedenja pred ogledom posamezne strani. Večja napovedna moč mešanice modelov je bila pokazana tudi empirično (Anderson, Domingos, Weld, 2002; Cadez et al., 2003).

3.4 ANALIZA ODKRITIH VZORCEV UPORABE

Faza analize odkritega znanja je v procesu podatkovnega rudarjenja izredno pomembna, saj nam omogoča, da ocenimo veljavnost in uporabnost odkritih vzorcev in primerjamo natančnost posameznih modelov, ki smo jih razvili v predhodni fazi. Hkrati v tej fazi tudi preverimo, v kolikšni meri generirani modeli ustrezajo postavljenim poslovnim ciljem, in

[47] Gre za tradicionalno IID-predpostavko, po kateri so naključne spremenljivke med seboj neodvisne in imajo enako verjetnostno porazdelitev. Kratica IID je sestavljena iz angleških besed, ki opisujejo osnovne značilnosti predpostavke (Independent and Identically-Distributed).

[48] Ta statistični model predstavlja komponento mešanega modela.

skušamo ugotoviti, ali obstajajo kakršni koli poslovni razlogi, zaradi katerih so modeli nepopolni. Ker je podatkovno rudarjenje cikličen proces, navadno potrebujemo nekaj ponovitev, da najdemo model ali modele, ki ustrezajo vsem zahtevam.

Eden izmed pomembnih razlogov za analizo odkritih vzorcev je pojasnjevanje odkritega znanja in izločanje nezanimivih vzorcev in pravil, saj je za končne uporabnike tipično zanimivo le majhno število odkritih pravil. Pravilo je zanimivo in predstavlja znanje, če ga lahko uporabniki enostavno razumejo, če je z določeno stopnjo verjetnosti veljavno tudi na novih ali testnih podatkih, če je potencialno uporabno[49] in če prej še ni bilo poznano (Han, Kamber, 2000, str. 27). Včasih se zgodi tudi, da model ne vsebuje nobenih zanimivih vzorcev; en možni razlog za tako situacijo je lahko v tem, da so podatki povsem naključni in zato v njih ni bogatih informacij, verjetnejši razlog pa je navadno v tem, da izbor spremenljivk v modelu ni bil pravilen (Tang, MacLennan, 2005, str. 16). Za ocenjevanje zanimivosti vzorcev obstaja kar nekaj objektivnih meril, kot so npr. v prejšnjih poglavjih opisane mere podpore in zaupanja (asociacijska pravila, zaporedni vzorci)[50], ki temeljijo na strukturi vzorcev in različnih statističnih metodah, vendar pa večinoma niso primerna za ocenjevanje vzorcev, odkritih z neusmerjenimi metodami odkrivanja znanja. Na drugi strani temeljijo subjektivne metode ocenjevanja na uporabnikovih prepričanjih, da so vzorci zanimivi in nepričakovani ali pa da mu nudijo strateške informacije, na podlagi katerih lahko ukrepa; tak način ocenjevanja je primeren zlasti pri ocenjevanju primernosti rezultatov razvrščanja v skupine.

Kakovost napovedovalnih modelov moramo preveriti tudi v smislu njihove napovedovalne natančnosti in moči. Za ta namen lahko uporabimo mero, imenovano dvig (ang. lift), ki je definirana kot razmerje med pogostostjo pojava v vzorčni populaciji in pogostostjo pojava v celi populaciji; grafično jo lahko prikažemo v t. i. grafikonu dviga (ang. lift chart), s katerim lahko enostavno primerjamo kakovost izbranih modelov z idealnim modelom in »modelom« z naključnim ugibanjem. V kontekstu spletnega poosebljanja lahko npr. primerjamo, koliko uporabnikov, ki bi jih izbrali na podlagi njihovega navigacijskega vedenja, bi v primerjavi z naključno izbranimi uporabniki obiskalo stran A. Med druge grafikone natančnosti (ang. accuracy chart), ki jih lahko uporabimo za preverjanje kakovosti modelov, spadajo tudi grafikon dobička (ang. profit chart), s katerim preverjamo finančne učinke modelov, in raztreseni grafikon natančnosti (ang. scatter accuracy plot), ki grafično prikazuje odmike napovedi od dejanskih vrednosti v primeru, da je ciljna spremenljivka zvezna. Pri diskretnih ciljnih spremenljivkah lahko za ta namen uporabimo klasifikacijsko matriko (ang. classiffication matrix; tudi confusion matrix), v kateri so prikazani podatki o pravilnih in napačnih napovedih (Tang, MacLennan, 2005; Seth et al., 2005).

---

[49] Določena pravila so sicer popolnoma veljavna, vendar pa z vidika poslovnega modela nepomembna. Primer takega pravila bi bilo npr. pravilo »če je uporabnik moški in je poročen, potem je s 100-odstotnim zaupanjem njegov zakonski partner ženska«.

[50] Na zahtevane vrednosti zaupanja in podpore lahko vplivamo že v fazi modeliranja, saj lahko večini algoritmov določimo njihove najnižje vrednosti.

Sama metodologija analize odkritih vzorcev uporabe je močno odvisna od namena uporabe rudarjenja po spletu, zato je pomembno, da ima analitik na voljo tudi orodja, ki omogočajo analizo po meri. Najpogostejša oblika take analize je uporaba tehnik vizualizacije, različnih jezikov za raziskovanje odkritega znanja in različnih analitičnih sistemov, kot je npr. OLAP[51]. Vizualizacija je preverjen način za pomoč človeku pri njegovem razumevanju različnih realnih in abstraktnih pojavov, saj je človeški sistem zaznavanja naravno prilagojen za razpoznavanje razlik in povezav med objekti, ki so prikazani v različnih barvah, oblikah, velikostih in postavitvah (Klösgen, Lauer, 2002, str. 509). Za posamezne metode podatkovnega rudarjenja (predvsem za odločitvena drevesa, asociacijska pravila, Bayesove mreže in razvrščanje v skupine) so bila razvita posebna vizualizacijska orodja (glej npr. Klösgen, Lauer, 2002; Cadez et al., 2003), prav tako pa obstaja več različnih orodij za vizualizacijo navigacijskih poti in vzorcev navigacijskega vedenja. Nabor slednjih orodij sega od enostavnejših pristopov, pri katerih gre zgolj za grafično predstavitev spletnih poti (glej npr. sistem WebViz, Pitkow, Bharat, 1997), pa do naprednejših metod, pri katerih se poleg priljubljenosti povezav in poti med posameznimi stranmi prikazujejo tudi njihova vsebinska podobnost in drugi atributi (glej npr. Herder, Weinreich, 2005; Chi, 2002).

Za uspešno analizo vzorcev in njihovo uporabo v procesu podpore odločanju je zelo pomembno, da ima analitik proste roke, da raziskuje v smer, ki se mu zdi primerna in potrebna. Za podporo tej aktivnosti obstaja več jezikov za raziskovanje vsebine izdelanih modelov podatkovnega rudarjenja; ti so večinoma podobni jeziku SQL, vendar pa splošni standard še ni sprejet; večina jezikov je bila namreč razvita posebej za različne sisteme za podatkovno rudarjenje ali pa v okviru različnih raziskovalnih projektov (npr. MINT, Spilipoulou, Pohle, 2001 ali DMQL, Han, Kamber, 2000). Vseeno se tudi na tem področju pojavlja vse več pobud za standardizacijo, med katerimi velja omeniti SQL/Multimedia for Data Mining[52], Java Data Mining API[53], PMML[54] in OLEDB[55] for DM; prav slednja pobuda je izredno zanimiva, saj je njen namen približati tehnologijo podatkovnega rudarjenja širši razvijalski skupnosti. OLEDB for DM je industrijski standard, katerega razvoj vodi Microsoft in ki definira splošne koncepte in API[56] za podatkovno rudarjenje ter na konceptih relacijskih baz temelječi jezik za izdelavo in učenje modelov, pregledovanje njegove vsebine in

---

[51] OLAP (ang. OnLine Analytical Processing, sprotna analitična obdelava podatkov) je programska oprema za takojšnjo analizo podatkov v podatkovni bazi. Praviloma prikazuje podatke večrazsežnostno.

[52] SQL/Multimedia (SQL MM) je ISO standard, zasnovan kot SQL razširitev za različne podatkovne strukture. Pobudnik razširitve za podatkovno rudarjenje je predlagal IBM (ISO/IEC 13249-6).

[53] Java Data Mining API je javanski paket za podatkovno rudarjenje in javanskim aplikacijam omogoča komunikacijo z modeli podatkovnega rudarjenja. Razvoj paketa vodi Oracle (JSR-73).

[54] PMML (Predictive Model Markup Language) razvija industrijska organizacija Data Mining Group, katere člani so največji ponudniki paketov za podatkovno rudarjenje (SAS, SPSS, IBM, Microsoft, Oracle, KXEN idr.). Namen jezika PMML je definirati standardni XML-format za povezovanje z modeli podatkovnega rudarjenja na način, ki je neodvisen od aplikacij, sistemov in arhitekture (DMG).

[55] OLEDB je Microsoftov API za enoten dostop do različnih tipov tabelarnih podatkov.

[56] API (Application Programming Interface) je vmesnik, ki aplikacijam omogoča klicanje funkcij v operacijskem sistemu ali drugih programih.

napovedovanje (DMX – Data Mining Extensions). S standardom OLEDB for DM je močno povezan tudi standard XML for Analysis, za katerega skrbi skupina XML/A Council in ki prek protokola SOAP[57] omogoča namenskim programom od platforme neodvisno povezovanje s strežniki za podatkovno rudarjenje in s strežniki OLAP.

3.4.1 Uporaba odkritih vzorcev za izboljševanje učinkovitosti spletišča

Eden izmed pogostih razlogov za uporabo rudarjenja po podatkih o uporabi spleta je uporaba tako pridobljenega znanja za ocenjevanje strukture in uporabnosti spletišča, in sicer z namenom, da bi povečevali njegovo uspešnost (glej npr. Jorge et al., 2003; Chi, 2002; Spilipoulou, Pohle, 2001; Srikant, Yang, 2001). Osnovno vprašanje, ki ga je treba v tej zvezi razrešiti, je, kdaj je spletišče sploh uspešno; pri tem je treba izhajati iz strategije spletišča in najprej ugotoviti, kakšni so poslovni cilji spletišča (zakaj spletišče obstaja), kakšno vrednost ustvarja za uporabnike (zakaj naj bi ga uporabniki obiskali) in kako trajna za uporabnike je dodana vrednost (čemu naj bi se uporabniki vrnili). Če želimo, da je spletišče uspešno, mora zadovoljevati tako potrebe njegovih lastnikov kot tudi potrebe samih obiskovalcev; pri tem običajno zadovoljstvo lastnikov merimo s poslovno usmerjenimi merili, zadovoljstvo uporabnikov pa z uporabniško usmerjenimi merili za vsako uporabniško skupino posebej (Berendt, Mobasher, Spillipoulou, 2002).

V številnih programih za merjenje statistike obiska je uspešnost spletišča izražena z osnovnimi statističnimi podatki, kot so npr. trajanje obiska, število obiskov ali uvrstitev v iskalnikih, vendar pa so ti podatki preveč splošni in površinski, hkrati pa so lahko tudi zavajajoči (veliko število obiskovalcev na našem spletišču namreč še ne pomeni, da so uporabniki z njim zadovoljni in da izvajajo želena transakcijska dejanja). Podobno nepopolne so metrike, kot so npr. dostopnost strani, odzivi časi ipd., ki uspešnost povezujejo s kakovostjo posameznih strani (glej npr. Sullivan, 1997; Alpar, 1999), saj jih je zaradi razlik v pomembnosti posameznih strani težko prenesti na raven celotnega spletišča. Boljše rezultate dajejo metrike za merjenje uporabnosti spletišča (ang. web usability) (glej npr. Eighmey, 1997); ta je visoka, kadar lahko uporabniki svoje cilje dosegajo hitro, enostavno in z majhno stopnjo napake ter pri tem izkusijo visoko subjektivno zadovoljstvo (Nielsen, 2003). Pomanjkljivost slednjih metod merjenja uspešnosti je v tem, da so močno odvisne od izbora reprezentativnega vzorca uporabnikov in njihovega odziva, hkrati pa še vedno ne nudijo enostavnega odgovora glede uspešnosti spletišča z vidika lastnikov. Lastnike namreč bolj zanimajo mere, kot sta učinkovitost stika (ang. contact efficieny) in učinkovitost pretvorbe (ang. conversion efficiency) (Berthon, Pitt, Watson, 1996; Spilipoulou, Pohle, 2001); prva mera pove, kolikšen je delež uporabnikov, ki so na spletišču ostali dlje od najmanjšega predvidenega časa, druga mera pa, kolikšen je delež uporabnikov, ki so po raziskovanju

[57] SOAP (Simple Object Access Protocol) je protokol za izmenjavo XML-sporočil prek protokola HTTP.

spletišča tudi izvedli želena transakcijska dejanja (npr. nakup, prijava na e-novice ipd.).

Izračun zgoraj opisanih mer lahko temelji zgolj na statistiki obiska spletišča ali njegovih posameznih delov, lahko pa si pri merjenju in povečevanju uspešnosti pomagamo tudi z rezultati in vzorci, ki jih odkrijemo v procesu rudarjenja po podatkih o uporabi spleta. Rudarjenje nam omogoča, da uspešnost spletišča izmerimo bolj podrobno kot na ravni celotne populacije uporabnikov, saj lahko z njim identificiramo pomembne skupine uporabnikov in izločimo vzorce, ki prispevajo k uspešnosti spletišča, ter vzorce, ki opisujejo, kako k temu prispevajo posamezni deli spletišča (Berendt, Mobasher, Spilipoulou, 2002).

Metode rudarjenja po podatkih o uporabi spleta lahko po drugi strani uporabimo tudi za izboljšanje navigacije po spletišču. Navigacijska zasnova je eden izmed ključnih elementov uspešnosti spletišča in bi morala biti enostavna in konsistentna, hkrati pa bi morala podpirati uporabnikove cilje in njegovo vedenje ter mu omogočati prihranek časa pri izvajanju nalog (Fleming, 1998). Realnost je v večini primerov drugačna, saj je sistem krmarjenja po spletišču zasnovan na organizacijsko zasnovanih predpostavkah o tem, kako bi morali uporabniki uporabljati spletišče. Te predpostavke niso vedno usklajene s potrebami in znanjem uporabnikov, hkrati pa ne upoštevajo posebnosti posameznih uporabniških skupin in se ne spreminjajo tako pogosto, kot se spreminjajo navade obiskovalcev spletišča. Vzorci uporabe, ki jih lahko odkrijemo z analizo podatkov o obisku spletišča, nam v tej zvezi pomagajo identificirati razlike v predvidenem in želenem navigacijskem vedenju (Chi, 2002; Berendt, Mobasher, Spilipoulou, 2002) ter izvesti spremembe za njuno zbliževanje (npr. dodajanje povezav med stranmi, ki sicer niso povezane, a so pogosto obiskane hkrati v okviru posameznih uporabniških sej). Za večanje uspešnosti spletišča po drugi strani seveda ni nujno, da skušamo vedenje uporabnikov prilagoditi našim predstavam, saj lahko pogosto večji učinek dosežemo tako, da opaženo vedenje sprejmemo in ga izkoristimo za povečevanje učinkovitosti stika in pretvorbe (Spilipoulou, Pohle, 2001). Drug, še perspektivnejši, vidik uporabe rudarjenja po spletu je uporaba znanja o vedenju obiskovalcev za spletno poosebljanje; le-to obravnavam v naslednjem poglavju.

4 SPLETNO POOSEBLJANJE

Svetovni splet (ang. world wide web) še vedno raste s presenetljivo hitrostjo, tako z vidika števila obiskov na spletnih straneh kot tudi z vidika velikosti in zapletenosti spletišč (ang. web site). Hkrati do dostopajo svetovnega spleta vse boj heterogeni in globalno razporejeni uporabniki, ki imajo različne interese in cilje in za brskanje uporabljajo različne naprave in vmesnike. Da bi organizacije lahko maksimirale uporabnost objavljenih informacij, jih morajo filtrirati in jih prilagoditi potrebam individualnih uporabnikov, zaradi česar postaja

oblikovanje spletnega mesta, upravljanje spletnega strežnika in zagotavljanje enostavnosti pri navigaciji skozi spletno mesto vse bolj zapletena naloga.

Pomembnost uporabniške prijaznosti spletišč in enostavne navigacije močno poudarja teorija iskanja informacij (ang. information foraging theory)[58] (Pirolli, Card, 1995); ta pravi, da uporabnik porabi tem manj časa na posameznem spletišču, kolikor več je spletišč, ki mu ponujajo informacije. To pomeni, da mora imeti spletišče intuitivno navigacijo in čim manj drugih ovir za odkrivanje informacij, saj se v nasprotnem primeru hitro povečuje tveganje, da bo uporabnik spletišče zapustil. V skladu s konceptom informacijske sledi (ang. information scent), ki je osrednji koncept teorije iskanja informacij, bo uporabnik namreč informacije na spletišču iskal, vse dokler bo imel občutek, da je na pravi poti; če sled z vsakim klikom ne postaja močnejša, uporabniki odnehajo; pri tem se mora zdeti napredek pri iskanju dovolj hiter, da je vreden pričakovanega napora za dostop do želene strani.

V luči teorije iskanja informacij, velikega števila spletišč, (pre)obilja informacij in naraščajoče heterogenosti uporabnikov je eden izmed najučinkovitejših pristopov, ki jih lahko podjetja uporabijo za vzpostavljanje in poglabljanje odnosa s svojimi strankami, spletno poosebljanje (ang. web personalization).

4.1 OPREDELITEV SPLETNEGA POOSEBLJANJA

Natančna opredelitev spletnega poosebljanja, tj. poosebljanja uporabniške izkušnje na spletnih straneh, je vse prej kot lahka naloga, saj funkcionalnosti, ki so kategorizirane kot »poosebljanje« segajo od prikazovanja imena uporabnika na spletišču do možnosti spreminjanja navigacije in prilagajanje izdelkov na osnovi modelov uporabniških potreb in vedenja. V povezavi s tem magistrskim delom lahko spletno poosebljanje definiramo kot *»proces spreminjanja funkcionalnosti, vmesnika, vsebine ali značilnosti sistema, in sicer z namenom, da se poveča njegova relevantnost za posameznega uporabnika«* (Blom, 2000, str. 313), oz. kot *»vsako dejanje, s katerim se informacije ali storitve, ki jih ponuja spletišče, ponuja posameznemu uporabniku ali skupini uporabnikov na osnovi znanja o navigacijskih navadah uporabnikov in njihovih individualnih interesih ter v povezavi le-teh z vsebino in strukturo spletišča«* (Eirinaki, Vazirgiannis, 2003, str. 1).

Osnovni namen poosebljanja je uporabnikom ponuditi tisto, kar si želijo, ne da bi jim bilo treba to izrecno zahtevati, pri čemer je uspešnost tega procesa odvisna od treh vidikov spletišča: ponujene vsebine, postavitve posameznih strani in strukture celotnega spletišča

[58] Teorija iskanja informacij izhaja iz teorije optimalnega iskanja v biologiji in antropologiji. Omenjena teorija analizira izmenjavo (ang. trade-off) med vrednostjo pridobljene informacije in stroški izvajanja iskalnih aktivnosti s pomočjo računalnika (Pirolli, Card, 1995).

(Mulvenna et al., 2000). Relevantnost vsakega objekta na spletni strani za posameznega uporabnika močno vpliva na raven njegovega zadovoljstva, prav tako pa struktura spletišča, ki je definirana z obstojem povezav med različnimi stranmi, uporabnikovo navigacijo omejuje na nabor vnaprej definiranih poti in tako določa uporabnikovo zmožnost, da relativno enostavno dostopa do njemu pomembnih strani. Vendar pa je definicija pomembnosti zelo subjektivna, zato pogosto prihaja do razkoraka med tem, kako uporabniške potrebe dojemajo načrtovalci spletišča, in dejanskimi potrebami uporabnikov, kar lahko vpliva na učinkovitost spletišča.

Proces poosebljanja lahko začne kar sistem ali pa uporabnik sam, pri čemer slednji postopek imenujemo prikrojevanje (ang. customization)[59]. Osnovna razlika med poosebljanjem in prikrojevanjem je v tem, da pri prikrojevanju uporabnik samostojno izvede vnaprej določena dejanja, s katerimi spremeni vsebino, postavitev, stil in druge lastnosti uporabniškega vmesnika spletišča (Boiko, 2005, str. 740), medtem ko se prilagajanje vsebine pri poosebljanju izvaja dinamično, brez obveznih posegov uporabnika. Formalno gledano so sistemi, ki omogočajo prikrojevanje, adaptabilni sistemi (ang. adaptable systems), sistemi za poosebljanje pa prilagodljivi sistemi (ang. adaptive systems) (Markellou, Rigou, Sirmakessis, 2005, str. 29). Med prikrojevanjem in poosebljanjem sicer ne moremo potegniti povsem jasne ločnice, saj v aplikacijah pogosto soobstajata, končno razmerje med njima pa se določi z upoštevanjem koristi, zahtev in potrebne vpetosti uporabnika ter posledic napačnih prilagoditev. Pogosto je primerno, da ima uporabnik pri prilagajanju možnost, da prilagajanje dovoli ali onemogoči ali vsaj potrdi posamezne prilagoditve in druge spremembe.

Začetni poskusi spletnega poosebljanja so bili dejansko omejeni zgolj na prikrojevanje, pri katerem so si lahko uporabniki ročno izbrali vsebine, ki so jih želeli imeti prikazane na svojih »osebnih« straneh. Poleg tega, da je bila uspešnost poosebljanja pri tem pristopu močno odvisna od interesa uporabnikov, da sporočijo vse zahtevane podatke, ima ta pristop še eno pomembnejšo pomanjkljivost – namreč to, da mora uporabnik že vnaprej poznati vsebino, ki ga zanima. Zaradi omenjenih pomanjkljivosti so se pozneje pojavili naprednejši pristopi k poosebljanju, med katerimi velja omeniti zlasti naslednje (Mobasher, 2007, str. 21):

- **Sistemi na osnovi odločitvenih pravil** (ang. manual decision rule systems) so sistemi, pri katerih poosebljanje storitev, ki jih ponuja spletišče, temelji na pravilih, ki jih definira načrtovalec spletišča, in podatkih, ki jih sporočijo uporabniki. Tipično se pri teh sistemih pri postopku registracije pridobi statične profile posameznikov, ki se jim nato na podlagi različnih pravil in modelov prikaže ustrezna vsebina[60].

---

[59] Prikrojevanje je osnova t. i. »my« fenomena. Številni portali (npr. Yahoo) besedo »my« (moj) uporabljajo za označevanje storitve za prikrojevanje nekaterih objavljenih vsebin.

[60] Primer takega sistema je storitev MyYahoo!, ki omogoča samodejno poosebljanje vsebine športnih rezultatov, vremenske napovedi in drugih vsebin na osnovi podatkov o prebivališču uporabnika.

- **Sistemi za filtriranje na osnovi vsebine** (ang. content-based filtering systems) so sistemi, pri katerih se za priporočanje novih vsebin uporabnikom uporabljajo podatki iz njihovih profilov in podatkov o zgodovini njihovih obiskov. Pri tem proces priporočanja temelji na vsebinski podobnosti med že obiskanimi in novimi vsebinami (za izračun podobnosti se uporabljajo tehnike strojnega učenja in rudarjenja po besedilih) ali podobnosti med samim uporabniškim profilom in vsebino[61].

- **Sistemi za filtriranje na osnovi sodelovanja** (ang. collaborative filtering systems) so sistemi, pri katerih poosebljanje vsebine temelji na primerjavi skupnih značilnosti v preferencah različnih uporabnikov, pri čemer so te preference običajno izražene eksplicitno ali pa so pridobljene na osnovi uporabnikovega ocenjevanja različnih vsebin. Za razliko od filtriranja na osnovi vsebine, pri katerem je glavni poudarek na tem, kaj uporabnika zanima, je pri filtriranju s sodelovanjem poudarek na iskanju oseb, ki jih zanimajo podobne stvari kot danega uporabnika (Pierrakos et al., 2003, str. 319). Najbolj znana aplikacija tega pristopa je uporaba priporočanja v spletni knjigarni Amazon.com (Linden, Smith, York, 2003).

Sistemi, ki temeljijo na nesamodejnih odločitvenih pravilih, imajo enako pomanjkljivost kot drugi neavtomatizirani kompleksni sistemi, tj. zahtevajo veliko truda, da se jih namesti in vzdržuje. Poleg tega večinoma zahtevajo intenzivno sodelovanje uporabnika, kar je velik minus za njihovo uporabo. Oba omenjena pristopa na osnovi filtriranja skušata to težavo premostiti z uporabo tehnik strojnega učenja, s pomočjo katerih se lahko analizirajo podatki in sestavijo zahtevani uporabniški modeli, vendar pa tudi nista brez pomanjkljivosti. Osnovna težava pri filtriranju na osnovi vsebine je v analizi vsebine spletnih strani in iskanju semantičnih podobnosti; kljub velikemu napredku v zadnjih letih je taka analiza še vedno omejena na različne statistične metode[62], ki pa se jih ne da uporabljati pri večpredstavnostnih vsebinah (ang. multimedia) in drugih vsebinah z malo besedila. Pri filtriranju na osnovi sodelovanja teh težav sicer ni, vendar pa se tu pojavlja težava s skalabilnostjo, poleg tega pa je kakovost priporočil pri tej vrsti poosebljanja močno odvisna od števila ocen posameznega uporabnika; hkrati mora imeti posamezni objekt vsaj eno oceno, da se ga lahko sploh priporoči uporabnikom (Pierrakos et al., 2003, str. 319). Za odpravljanje težav s skalabilnostjo je bilo predlaganih več optimizacijskih strategij (glej Linden, Smith, York, 2003), ki so dovolj učinkovite, da se filtriranje na osnovi sodelovanje uspešno uporablja v številnih e-trgovinah, vendar pa ni primerno za integriranje poosebljene izkušnje uporabnikov po celem spletišču in v proces poosebljanja težko vključi različne tipe objektov in različne uporabniške dogodke.

---

[61] Med primere teh sistemov uvrščamo orodja, kot so PRES, WebWatcher, Letizia idr. (glej Meteren, Someren, 2007).

[62] Med omenjene statistične metode štejemo uporabo že omenjene TF.IDF vektorske predstavitve besedil, latentno semantično analizo (Jin, Zhou, Mobasher, 2004) idr.

Zaradi teh težav se je raziskovanje obrnilo k t. i. opazovalnemu poosebljanju (ang. observational personalization); to temelji na predpostavki, da lahko namige za uspešno poosebljanje najdemo v preteklem navigacijskem vedenju obiskovalcev, zato uporabnikom ni treba razkrivati nobenih osebnih informacij (Mulvenna et al., 2000, str. 124). Na opazovalno poosebljanje lahko gledamo tudi kot na nadgradnjo filtriranja na osnovi sodelovanja; algoritmi za odkrivanje vzorcev lahko pri opazovalnem poosebljanju na osnovi ocenjevalnih in navigacijskih profilov uporabnikov generirajo agregatne uporabniške modele, ki jih lahko v kombinaciji s profilom aktivnega uporabnika uporabimo za učinkovito napovedovanje njegovega vedenja in priporočanje vsebine (Mobasher, 2007, str. 6).

Količina podatkov, ki jo je treba analizirati pri opazovalnem poosebljanju, je tako velika, da je ni mogoče obvladati brez močnih orodij in metod, hkrati pa klasične metode preiskovanja in analiziranja podatkov niso več uporabne. Zato Mobasher, Cooley in Srivastava (2000) za ta namen predlagajo uporabo spletnega rudarjenja, natančneje rudarjenja po podatkih o uporabi spleta, ki lahko pomaga izboljšati skalabilnost, natančnost in fleksibilnost sistemov za poosebljanje.

4.1.1 Možnosti uporabe spletnega poosebljanja

Tehnologije za poosebljanje segajo od vsakdanje uporabe podatkovnih baz, piškotkov in dinamičnega generiranja strani do iskanja ujemajočih se vzorcev, uporabe algoritmov s področja strojnega učenja, sklepanja na osnovi različnih pravil in, kot sem že večkrat omenil, tudi podatkovnega rudarjenja. Že to kaže, da lahko sistemi za spletno poosebljanje omogočajo zelo širok nabor funkcij poosebljanja in jih lahko z vidika izbrane politike poosebljanja (ang. personalization policy) klasificiramo na osnovi zelo različnih meril; ta merila so prikazana v tabeli 1.

Tabela 1: Klasifikacija sistemov za poosebljanje z vidika politike poosebljanja

Klasifikacija	Merilo
Individualni, sodelovalni (ang. collaborative)	Izbor podatkov, ki se uporabljajo pri poosebljanju
Proaktivni, reaktivni (tudi konzervativni)	Vpetost uporabnika v postopek poosebljanja
Temelječi na uporabnikih, temelječi na proizvodih	Vrsta podatkov, ki se uporablja pri priporočilih
Temelječi na spominu, temelječi na modelu	Tehnična izvedba
Odjemalski, strežniški	Lokacija izvajanja poosebljanja
Enouporabniški, večuporabniški	Število uporabnikov, na katerih temelji model, ki se uporablja za poosebljanje pri posameznem uporabniku
Občutljivi na okoliščine, neobčutljivi na okoliščine	Prilagajanje načina poosebljanja glede na prikazano vsebino
Statični, dinamični	Število uporab funkcij poosebljanja
Pojasnjevalni, nepojasnjevalni	Razpoložljivost pojasnil o vzrokih poosebitev

Vir: Anand, Mobasher, 2005; Pierrakos et al., 2003

Z vsebinskega vidika lahko funkcije poosebljanja razdelimo v štiri razrede: pomnjenje (ang. memorization), vodenje (ang. guidance), prilagajanje (ang. customization) in podpora za učinkovitejše izvajanje nalog (ang. task performace support) (Blom, 2000; Kobsa, Koeneman, Pohl, 2001; Pierrakos et al., 2003); vsak razred je podrobneje predstavljen v nadaljevanju.

4.1.1.1 Pomnjenje

Pomnjenje je najenostavnejša oblika poosebljanja; zanjo je značilno, da sistem spremlja in shranjuje podatke o uporabniku, kot so npr. ime ali obiskane strani. Ko se obiskovalec vrne na spletišče, se ti podatki uporabljajo za to, da se ga spomni na njegovo preteklo vedenje. Spletišča pomnjenja največkrat ne ponujajo kot samostojno obliko poosebljanja, ampak kot del večjih rešitev za poosebljanje. Primeri funkcije pomnjenja so npr.:

- **Pozdravljanje uporabnika.** Sistem za poosebljanje ob obisku prepozna uporabnika in izpiše njegovo ime (običajno skupaj s pozdravnim besedilom). Čeprav je ta oblika poosebljanja zelo enostavna, predstavlja običajno prvi korak za povečevanje zvestobe in odzivnosti strank (Joinson, 2007).

- **Ustvarjanje zaznamkov.** Sistem shranjuje podatke o straneh, ki jih je uporabnik obiskal v preteklosti, in mu jih ponudi za lažje iskanje informacij ob ponovnih obiskih. Ena izmed možnih izvedb prikaza takih zaznamkov je tudi označevanje povezav v zemljevidih spletišč.

- **Poosebljene pravice za dostop.** Sistem prepozna raven pravic posameznega uporabnika in mu ponudi vse informacije, do katerih lahko dostopa.

4.1.1.2 Vodenje

Namen vodenja pri spletnem poosebljanju je pomagati uporabnikom, da bi na spletišču hitreje našli informacije, ki jih zanimajo, in jim hkrati ponuditi alternativne možnosti za pregledovanje vsebine spletišča. Poleg povečevanja zaupanja strank rešuje vodenje tudi težavo informacijskega preobilja in je v primerjavi s tehnikami pomnjenja, ki so v pomoč predvsem uporabnikom, ki velikokrat obiskujejo posamezno spletišče, posebej koristno za nove in nevešče uporabnike. Primeri funkcij vodenja vključujejo:

- **Priporočanje vsebine.** Sistemi za priporočanje vsebine spadajo med najbolj priljubljene sisteme za poosebljanje; njihov namen je, da uporabniku priporočijo seznam vsebin, za katere se v skladu z izbranim modelom poosebljanja predvideva, da ga zanimajo. Vsebinsko gledano se lahko uporabniku priporoča različne tipe objektov (npr. izdelke, storitve, informacije itd.), tehnično pa so taka priporočila običajno povezave na ustrezne spletne strani; te povezave se večinoma prikazujejo v posebnem oknu ali okviru ali v posebej označenemu delu strani.

- **Učenje uporabnikov.** Poosebljeno spletišče lahko uporabniku nudi v skladu z njegovim znanjem in interesi na vsakem koraku pomoč pri uporabi spletišča. Za učenje so primerne skoraj vse tehnike za poosebljanje, med drugim tudi priporočanje vsebine, poudarjanje ali prilagajanje določenih delov vsebine, čarovniki za prikaz vsebine ipd.

4.1.1.3 Prilagajanje

V razredu prilagajanja najdemo funkcije, ki prilagajajo vsebino, strukturo in postavitev spletišča ali posameznih spletnih strani, pri čemer se upoštevajo želje, interesi, znanje in možnosti uporabnikov. Tovrstno prilagajanje spletišča je zelo primerno za učinkovito podajanje vsebine, saj po eni strani uporabnikom lajša interakcijo s spletiščem, po drugi strani pa lastnikom spletišča omogoča, da uporabnike lažje pripeljejo do ciljnih strani.

- **Prilagajanje vsebine.** Za spreminjanje vsebine spletnih strani z namenom, da se poveča njena relevantnost za posameznega uporabnika, se lahko uporablja več tehnik; najpogostejše so različne verzije strani, različne verzije posameznih delov strani, barvanje posameznih delov strani in prilagodljivo raztegljivo besedilo (ang. stretchtext)[63].
- **Prilagajanje strukture.** V skupino prilagajanja strukture lahko štejemo tudi funkcionalnost spreminjanja postavitve določenih strani (ang. page layout), večinoma pa v to skupino prištevamo funkcije za prilagodljivo manipulacijo spletnih povezav, kot so sortiranje, komentiranje, skrivanje (odkrivanje), onemogočanje (omogočanje) ter odstranjevanje (dodajanje) povezav. Tehnike prilagajanja strukture se največkrat uporabljajo v kombinaciji s tehnikami prilagajanja vsebine in so pogoste pri izgradnji poosebljenih pogledov in kotičkov.

- **Prilagajanje načina predstavitve.** Prilagoditve načina predstavitve in formata so prilagoditve, pri katerih vsebina običajno ostane enaka, spremeni pa se način njihovega posredovanja posameznemu uporabniku – slike se npr. spremenijo v besedilo, videoposnetki v slike, besedilo v zvok ipd. Te vrste prilagoditev so zelo koristne za uporabnike, ki si vsebine ne morejo optimalno ogledati na običajen način, bodisi zaradi omejitev programske ali strojne opreme bodisi zaradi različnih telesnih hib (npr. slabovidnost). Prilagajanje načina predstavitve lahko temelji na uporabnikovem ravnanju v preteklosti ali pa na podatkih o opremi, ki jo uporabnik uporablja (namesto osebnega računalnika lahko npr. za dostop do spletišča uporablja mobilni telefon) (Smyth, Berry, 2004).

---

[63] Raztegljivo besedilo je besedilo, ki ga lahko uporabnik ali sistem sam raztegne (razširi) oz. skrči.

4.1.1.4 Podpora za učinkovitejše izvajanje nalog

Poosebljanje kot podpora za učinkovitejše izvajanje nalog je najnaprednejši način poosebljanja in se je razvilo iz posebne vrste prilagodljivih sistemov, imenovanih osebni pomočniki. Tovrstno poosebljanje se največkrat uporablja za:

- **avtomatizirano izvajanje opravkov**, kot so poosebljeno prenašanje različnih datotek, pošiljanje pošte, iskanje informacij ipd.;

- **poosebljeno dokončevanje vnosov** pri izpolnjevanju obrazcev ali iskanju informacij;

- **poosebljeno iskanje**, ki omogoča samodejno prilagajanje iskalnih poizvedb in rezultatov iskanja, uporabniku prilagojeno oblikovanje skupin rezultatov in njihovo razvrščanje. Uporabnik lahko tako lažje najde informacije, ki ga zanimajo, hkrati pa je poosebljeno iskanje zelo učinkovito pri podpori raziskovalnemu iskanju (ang. exploratory search) (White et al., 2006; Kožuh, 2006).

4.1.2 Izbira ustreznih tehnik in oblik spletnega poosebljanja

Velika raznolikost oblik spletnega poosebljanja se kaže tudi v številnosti raziskav s tega področja; Pierrakos et al. (2003) v svoji študiji uporabe tehnik podatkovnega rudarjenja za spletno poosebljanje navajajo primere več kot trideset različnih aplikacij, v okviru katerih so bili za poosebljanje z različnim uspehom uporabljeni različni pristopi in algoritmi. Vendar pa je ta raznolikost lahko tudi nevarna – spletno poosebljanje je kot mlada in perspektivna disciplina trenutno še vedno dokaj neurejen in nepreizkušen nabor orodij, ki načrtovalce spletišč hitro premamijo, da poskušajo najti primerne uporabe zanjo, pri čemer pozabljajo, da bi se morali osredotočiti na povečevanje dodane vrednosti za končnega uporabnika. Pri vsem tem pa ni pomembno zgolj to, kakšno vrednost ima neka oblika poosebljanja za obiskovalce spletišča in kako dobro podpira sprejeto strategijo organizacije, ampak tudi, kako pomembna je v primerjavi z drugimi funkcionalnostmi, ki bi jih lahko organizacije uvedle, in kako poosebljanje vključiti v celotno strategijo spletnega nastopa.

Kramer, Noronha in Vergo (2000) za ugotavljanje, katere oblike poosebljanja so v posameznem primeru potrebne, predlagajo proces, ki je sestavljen iz šestih korakov in poteka ob sodelovanju z uporabniki; z modeliranjem uporabniških ciljev, prepričanj in obnašanja zagotavlja, da sistem za poosebljanje uporabnikom prinaša resnično vrednost. Poudarjajo celostni pristop, ki temelji na natančni opredelitvi in analizi nalog, ki se izvajajo na spletišču, ter njihovi analizi v skladu s specifikami posameznega projekta. Njihov pristop je vsekakor koristen in omogoča identificiranje pasti, ki se jim je potrebno izogniti pri načrtovanju poosebljenih storitev, a je tako dolgotrajen in zapleten, da ga organizacije redko izvajajo. Zato se zdi pomembnejši namen njihovega procesa in njihova osredotočenost na potrebe

uporabnikov, ki morajo biti pri načrtovanju rešitev spletnega poosebljanja vedno v ospredju.

4.1.3 Primernost uporabe spletnega poosebljanja

Mnenja o koristnosti in učinkovitosti poosebljanja so zelo različna. Številne raziskave govorijo poosebljanju v prid (ChoiceStream, 2006; Johnson, 2006; WebTrends, 2006, Sarner, 2004; Kobsa, 2001), saj naj bi imelo pozitiven vpliv na dobičkonosnost strank; kupci, ki imajo vsebino poosebljeno, preživijo v povprečju na spletišču dlje časa, obiščejo več strani in prinašajo več prihodka. Poleg tega ima poosebljanje pozitiven vpliv na zvestobo strank (Johnson, 2006; WebTrends, 2006; Sarner, 2004,), hkrati pa povečuje strokovno znanje, kot ga dojemajo posamezni uporabniki, in njihovo zaupanje v organizacijo, s čimer večajo njeno verodostojnost (Fogg, 2003). Temu stališču pritrjujejo tudi podatki o vlaganjih v tehnologijo za poosebljanje[64], ki naj bi se po predvidevanjih družbe Datamonitor od leta 2001 dalje povečevali za povprečno 84 odstotkov na leto in naj bi v letu 2006 dosegli 2,1 milijarde ameriških dolarjev (The global outlook for personalization applications, 2001). Podobne so ugotovitve raziskave globalnih ekonomskih trendov razvoja do leta 2020, po kateri naj bi bilo poosebljanje v najrazličnejših oblikah gonilna sila t. i. »interakcijskega gospodarstva« (ang. interaction economy)[65] (Foresight 2020, 2006).

Na drugi strani se pojavljajo tudi raziskave, ki opozarjajo na pretirano poveličevanje poosebljanja in spodbijajo podatke o njegovi učinkovitosti. Raziskava družbe Jupiter Research (Jupitermedia, 2003) je pokazala, da samo 14 odstotkov kupcev poosebljena priporočila prepričajo k dodatnim nakupom, za le 8 odstotkov obiskovalcev spletne strani pa so poosebljene storitve razlog za vnovični obisk. Raziskava trdi, da so vlaganja v poosebljanje spletnih strani večinoma ekonomsko neupravičena, saj je mogoče večino ciljev, ki jih želi spletišče doseči s poosebljanjem, doseči na cenovno učinkovitejše načine. Podobno je stališče Jacoba Nielsena, vodilnega strokovnjaka za spletno uporabnost (ang. web usability), ki trdi, da je potreba po spletnem poosebljanju pogosto zgolj izgovor, da spletišče ne sledi osnovnim smernicam uporabnosti (Nielsen, 1998). Opozorila, da poosebljanje ni vedno prava rešitev, se zdijo upravičena, še posebej zaradi strahu uporabnikov, da bo spletišče zlorabilo njihove podatke in posledičnemu izogibanju poosebljanja.

Različni pogledi na ustreznost uporabe spletnega poosebljanja narekujejo posebno pozornost pri odločanju za njegovo implementacijo. Pri tem je izjemno pomembno, da se strokovnjaki za spletno načrtovanje zavedajo, kakšne so zmožnosti in omejitve orodij za poosebljanje, kakšne so njihove koristi in pomanjkljivosti za končne uporabnike storitev, ki jih ponuja

---

[64] Tehnologija za poosebljanje zavzema najrazličnejše rešitve za poosebljanje izkušnje, izdelkov ali storitev. Največkrat se poosebljanje obravnava v povezavi s svetovnim spletom in CRM rešitvami.

[65] Interakcijsko gospodarstvo (tudi virutalno gospodarstvo) je gospodarstvo, pri katerem je zelo pomembna uporaba elektrosnkih storitev za interakcijo med organizacijami in njenimi uporabniki.

spletišče, ter kakšne so morebitne alternativne rešitve. Pri tem si lahko pomagamo tudi z analizo motivacijskih dejavnikov za poosebljanje, ki jih lahko razdelimo na tiste, ki omogočajo oz. lajšajo dostop do informacij, tiste, ki pomagajo pri doseganju delovnih ciljev, in tiste, ki so namenjene podpori osebnim razlikam[66] (Blom, 2000).

4.2 SPLETNO POOSEBLJANJE IN RUDARJENJE PO SPLETU

Internet je postal v preteklih letih za večino organizacij osrednja tema pri razvoju in nadgradnji poslovnih strategij, svetovni splet pa v tem kontekstu predstavlja organizacijam osnovno sredstvo za komunikacijo z uporabniki. Omogoča hiter, enostaven, nemoten in stroškovno učinkovitejši dostop do informacij, vendar pa mnoge uporabnike postavlja pred izziv, kako iz množice podatkov izluščiti tiste, ki jih v danem trenutku potrebujejo. Problem je namreč v tem, da je lahko v preteklosti uporabnik z brošurami, publikacijami in drugim tiskanim gradivom naenkrat dobil celotno informacijo, ki mu jo je avtor publikacije pripravil z namenom, da bi mu hkrati omogočil pregled in vse potrebne podrobnosti o določeni vsebini (Stražišar, 2005, str. 43); za razliko od tiskanih publikacij je vsebina na spletu usmerjena v ekran, kar velikokrat pomeni, da vsebine informacije ni vedno enostavno prepoznati in jo je potrebno oblikovati med pregledovanjem spletišča, torej od strani do strani. To težavo se da do določene mere omiliti z upoštevanjem načel spletne uporabnosti, povsem pa je ni mogoče odpraviti, saj je povezana z osnovnimi kognitivnimi značilnostmi spletnih uporabnikov (Tam, 2005) in z naraščajočo heterogenostjo spletnih uporabnikov.

Povsem tehtno vprašanje, ki se ob vsem tem zastavlja, je, zakaj bi torej vlagali toliko truda v načrtovanje spletišča in pripravo spletnih vsebin, če so današnji spletni uporabniki z vidika njihovega znanja in razumevanja ter interesov tako raznoliki, da je vsem nemogoče ustreči. Eden izmed pomembnih razlogov za to je vsekakor globalni doseg svetovnega spleta, zaradi katerega lahko uporabnik enostavneje kot kadarkoli prej zamenja svojega ponudnika informacij, storitev in izdelkov. Številne organizacije zato vlagajo velike napore v izboljševanje spletne izkušnje svojih uporabnikov, s tem pa obenem dvigujejo tudi pričakovanja in želje uporabnikov storitev neprofitnih in vladnih organizacij, kot je Statistični urad RS, kjer je uporabnikovo zadovoljstvo prav tako pomembno.

Eden izmed najpogostejših načinov, ki jih organizacije uporabljajo za povečevanje kakovosti in prijaznosti svojih spletišč, je uporaba metod za testiranje uporabnosti (ang. usability testing) – le-te so sicer učinkovit in primeren način za pridobivanje informacij o težavah, s katerimi se uporabniki srečujejo na spletišču, vendar pa imajo po mojem prepričanju dve veliki pomanjkljivosti, če jih ne kombiniramo z drugi pristopi. Prva se nanaša na objektivnost

[66] Osebne razlike so razlike v programski in strojni opremi posameznih uporabnikov, razlike, ki izhajajo iz morebitnih zdravstvenih omejitev, razlike, ki izhajajo iz različne ravni znanja, itd.

in celovitost metod za merjenje uporabnosti; uspešnost uporabljene metode je namreč po eni strani odvisna od pravilno zastavljenih nalog ali vprašanj izvajalca testa in primernosti njegove interpretacije rezultatov, po drugi stani pa tudi od verodostojnosti uporabnikovega obnašanja v danih testnih okoliščinah in reprezentativnosti uporabnikov, ki sodelujejo v testiranju. Drugi očitek, ki ga lahko pripišemo testom uporabnosti, je njihova osredotočenost na povprečnega uporabnika (Kožuh, 2005a). Osnovni namen testov je namreč, da bi spletišče zgradili oz. nadgradili tako, da bi čim bolj zadovoljevalo potrebe vseh uporabnikov, kar pa ob njihovi naraščajoči heterogenosti vodi do vse večjega zanemarjanja specifik v ciljih in znanju posameznih uporabniških skupin. Ena izmed možnih rešitev te težave je sicer, da pripravimo različne verzije spletišča in različne storitve za različne tipe uporabnikov, vendar pa se organizacije za tako rešitev zaradi težavnosti vzdrževanja odločajo izjemno redko.

Rudarjenje po spletu rešuje prvo omenjeno težavo, saj ponuja metode za učinkovito analizo velikih količin podatkov o obisku spletišča; ti v osnovi predstavljajo navigacijsko vedenje in navigacijske težave najširšega spektra uporabnikov v realnih okoliščinah. Z rudarjenjem lahko odkrijemo značilne uporabniške skupine, odkrivamo skrite vzorce uporabe ter potrjujemo ali zavračamo obstoječe hipoteze o uporabniškem obnašanju, s katerimi osebe, odgovorne za načrtovanje spletišča, utemeljujejo posamezne tehnične, strukturne in vsebinske značilnosti spletnih strani. Rešitev za drugo pomanjkljivost se skriva v spletnem poosebljanju, pri katerem izziv uporabniške raznolikosti rešujemo tako, da poskušamo spletišče čim bolj prilagoditi potrebam individualnega uporabnika ali specifičnih uporabniških skupin, pri čemer se vse te spremembe in prilagoditve izvajajo samodejno.

S širjenjem svetovnega spleta se je za vse organizacije močno povečal delež uporabnikov, o katerih nimajo zelo zanesljivih podatkov, zato vedno težje ugotavljajo, ali so vsebine uspele predstaviti v obliki, kot jih uporabniki potrebujejo v danem trenutku, ter kakšni so vzroki za morebiten neuspeh na tem področju. Rudarjenje po spletu kot samostojna oblika analize, predvsem pa v navezi s spletnim poosebljanjem, je tako ena izmed osnovnih metod, ki jih lahko organizacije uporabijo zato, da bi lahko izkoristile ves potencial, ki ga ponuja svetovni splet, in da bi s ponujanjem boljše uporabniške izkušnje koncept informacijske sledi (glej poglavje 4) obrnile sebi v prid.

Spletno poosebljanje na osnovi rudarjenja po spletu je bilo dejansko večkrat označeno kot ena izmed najperspektivnejših aplikacij (ang. killer application) podatkovnega rudarjenja (Mobasher, 2006; Markellou, Rigou, Sirmakessis, 2005; Kohavi, Provost; 2001; Mulvenna et al., 2000, idr.). Vzrok za to ni zgolj naraščajoča priljubljenost svetovnega spleta in s tem povezana naraščajoča kompleksnost spletnega poslovnega okolja, ampak tudi to, da je pri takem spletnem poosebljanju izpolnjenih vseh pet zahtev za uspešnost aplikacije podatkovnega rudarjenja (Kohavi, Provost, 2001):

- na voljo so podatki z bogatimi opisi, ki vsebujejo bolj zapletene vzorce, kot je zgolj

korelacija;

- podatkov je dovolj za izgradnjo zanesljivih modelov;
- zbiranje podatkov je nadzorovano, zanesljivo in avtomatizirano;
- rezultate lahko ovrednotimo;
- obstaja možnost integracije procesa podatkovnega rudarjenja z obstoječimi procesi.

4.3 OPREDELITEV PROCESA SPLETNEGA POOSEBLJANJA Z UPORABO RUDARJENJA PO PODATKIH O UPORABI SPLETA

Kot sem že omenil, se je zaradi težav pri poosebljanju na osnovi sodelovanja ali vsebine in drugih tradicionalnih oblik poosebljanja, raziskovanje sistemov za poosebljanje obrnilo k opazovalnemu poosebljanju, ki temelji na predpostavki, da lahko vse podatke, ki jih potrebujemo za uspešno poosebljanje, najdemo v podatkih o preteklem navigacijskemu vedenju obiskovalcev na spletišču; osnovna prednost tega pristopa je v tem, da uspešnost poosebljanja ne temelji izključno na pripravljenosti uporabnikov, da zaupajo svoje osebne preference, ampak se te podatke lahko pridobi implicitno s spremljanjem njihovega vedenja.

Prednost tega pristopa pa s seboj prinaša tudi težavo, saj je količina podatkov, ki jo je treba analizirati za učinkovito opazovalno poosebljanje, izredno velika. Tako je npr. povprečna velikost dnevnih spletnih dnevniških datotek spletišča Statističnega urada Republike Slovenije v letu 2006 znašala kar 96,1 MB oz. povprečno nekaj več kot 143 KB na vsakega obiskovalca, skupna velikost vseh dnevniških datotek v letu 2006 pa je znašala 34,2 GB[67]; povprečno dnevno število zapisov je znašalo 326.047, letno število zapisov pa dobrih 119 milijonov. Količina podatkov je sicer močno odvisna od načina spremljanja podatkov in števila atributov, ki jih spremljamo, vseeno pa je prevelika, da bi jo analizirali s klasičnimi orodji in tehnikami, zato Srivastava et al. (2000) za ta namen predlagajo uporabo rudarjenja po podatkih o uporabi spleta.

Namen rudarjenja po podatkih o uporabi spleta je zajeti vzorce obnašanja in profile obiskovalcev spletišča ter na njihovi osnovi izdelati model, ki omogoča, da se vsakemu obiskovalcu vsebina poosebi v skladu z njegovimi interesi, kot so zaznani z njegovimi vzorci

[67] V spletnih dnevniških datotekah Statističnega urada Republike Slovenije se shranjujejo vsi podatki, ki so definirani v W3C Extended Log format.

uporabe spletišča. Celoten proces lahko v grobem[68] razdelimo na tri faze (Mobasher, Cooley, Srivastava, 1999; Markellou, Rigou, Sirmakessis, 2005; Pierrakos et al., 2003):

1. priprava podatkov (zbiranje in čiščenje podatkov),

2. modeliranje (odkrivanje vzorcev iz podatkov),

3. priporočanje (poosebljanje).

Prvi dve fazi rudarjenja se tipično izvajata v zalednih sistemih, brez povezave s spletiščem (ang. offline); natančneje sem jih predstavil v poglavju 3. Nasprotno se priporočanje izvaja v stvarnem času (ang. real-time) in predstavlja povezano (ang. online) komponento sistema za poosebljanje. Pri tem pod procesom priporočanja ne razumemo le klasičnega priporočanja v obliki seznamov strani, izdelkov, storitev itd., ki se lahko ponudijo obiskovalcu z namenom, da se bo lažje odločal[69], ampak gre v tem procesu za izbor različnih objektov spletišča na osnovi priporočil, ki izhajajo iz uporabnikovega vedenja in modela, ter za njihovo predstavitev v obliki poosebljenega pogleda na spletišče. Priporočila kot seznami so v tem kontekstu torej samo ena izmed možnih oblik priporočanja vsebine.

Samodejno poosebljanje vsebine zgolj na osnovi izdelanih uporabniških modelov ne pomeni nujno, da je (samoprilagodljivo) spletišče sposobno povsem samostojno odločati o vrstah in načinu priporočanja ter povsem samostojno sestavljati vsebino strani. Tako prilagajanje je teoretično sicer mogoče[70], z vidika lastnikov spletišča in samih obiskovalcev pa ni priporočljivo. Perkowitz in Etizioni (2000) v eni izmed temeljnih študij samoprilagodljivih sistemov tako npr. zagovarjata pristop k prilagajanju, pri katerem se pri poosebljanju že v osnovi omejimo na zgolj neškodljive (ang. nondestructive) spremembe, hkrati pa uporabniku tudi pustimo možnost nadzora nad vsemi netrivialnimi spremembami. Z njunim pristopom se lahko strinjamo, saj ni rečeno, da bi lahko s popolnoma avtomatiziranim pristopom izpolnjevali vse cilje poosebljanja in zagotovili, da s spremembami ne bi preveč vznemirjali obiskovalce.

Kot lahko vidimo, je nadzor nad poosebljanjem potreben. Model, ki ga dobimo v fazi modeliranja, moramo pred njegovo uporabo zato temeljito pregledati z vidika njegove primernosti, nato pa odkrita pravila, vzorce in povezave pretvoriti v znanje glede analiziranega spletišča; šele nato sledi preoblikovanje odkritega znanja v niz uporabnih pravil,

---

[68] Proces rudarjenja po podatkih o uporabi spleta v ožjem pomenu zajema »tehnični« del procesa rudarjenja po podatkih po metodologiji CRISP-DM.

[69] Taka oblika priporočanja je najpogostejša oblika poosebljena in je značilna predvsem za spletne trgovine in podobna spletišča (npr. Amazon, iTIVI idr.).

[70] Popolnoma samoprilagodljiv sistem za poosebljanje bi moral imeti vgrajeno sposobnost učenja na lastnih napakah in odzivih obiskovalcev; na osnovi teh povratnih informacij bi se lahko njegov odziv s časom spreminjal. Primerna tehnika za tak način učenja so nevronske mreže.

ki omogočajo že opisano prilagajanje spletišča v smislu vsebine, strukture, načina predstavitve in medijskega formata (Markellou, Rigou, Sirmakessis, 2005). V praksi se poosebljanje praviloma izvaja tako, da se spremlja, katere strani je uporabnik obiskal, nato pa se poskuša na tej osnovi uporabnika bodisi dodeliti kakemu posplošenemu uporabniškemu profilu in vsebino poosebiti v skladu z značilnostmi dodeljenega profila, ali pa se vsebina pooseblja na osnovi vnaprej definiranih pravil; tudi ta pravila morajo seveda temeljiti na analizi pričakovanih potreb, zato lahko rečemo, da v tem primeru poosebljanje temelji na implicitnih uporabniških modelih. Pregled značilnosti in funkcionalnosti nekaj obstoječih sistemov za poosebljanje podajam v tabeli 2.

Tabela 2: Pregled nekaterih obstoječih sistemov za poosebljanje

	Način zbiranja podatkov	Identifikacija uporabnikov	Identifikacija sej	Metoda odkrivanja vzorcev	Vrsta poosebljanja
SETA	A, U	da	V	A	P, V, S
TELLIM	A	ne	V	K	S
UM2001	D	da	T	A	P, V, S
Oracle10iAS Personalization	A, U	da	V	K, A	P, V
NETMIND	D	da	T	S, K	V, S
Re:action	A	da	V	S, Z	V, S
WebPersonalizer	D, P	da	T, V	S, A	V
Yan et al.	D	ne	T	S	V
Kamdar, Joshi	D, P	da	T	S	V
SiteHelper	D	ne	T	K	V
WUM	D	ne	T	Z	S

Legenda:

- *Stolpec 2:* A = različni agenti za zbiranje podatkov, D = dnevniške datoteke, P = piškotki, U = podatki o uporabnikih
- *Stolpec 4:* T = na osnovi trajanja seje, V = na osnovi pregledane vsebine
- *Stolpec 5:* A = asociacijska pravila, K = klasificiranje, S = razvrščanje v skupine, Z = zaporedni vzorci
- *Stolpec 6:* P = pomnjenje, S = spreminjanje vsebine, V = vodenje

Vir: prirejeno po Pierrakos, 2003, str. 357–358

4.4 INTEGRACIJA SEMANTIČNEGA ZNANJA IN RUDARJENJA PO PODATKIH O UPORABI SPLETA

Vzorci uporabe, ki jih lahko odkrijemo z uporabo rudarjenja po podatkih o uporabi spleta, so učinkoviti pri zajemanju povezav na ravni uporabnikov, strani ali izdelkov in podobnosti na ravni uporabniških sej, vendar pa nam brez pomoči podrobnega znanja o področju (ang. domain knowledge) ne nudijo veliko vpogleda v temeljne razloge za opazovano vedenje in povezovanje med posameznimi objekti ali uporabniki. Nekatere pomenske vidike lahko sicer zajamemo z integracijo rudarjenja po podatkih o uporabi spleta in različnih pristopov za filtriranje vsebine na osnovi ključnih besed, vendar pa tako ne moremo zajeti povezav na

globlji semantični ravni, ki so bolj zapletene in temeljijo na atributih strukturiranih objektov.

4.4.1 Semantični splet in ontologije

Integracija semantičnega znanja za učinkovitejše poosebljanje z uporabo rudarjenja po podatkih o uporabi spleta temelji na ideji semantičnega spleta. Semantični[71] splet (ang. semantic web) je nadgradnja obstoječega spleta, v katerem lahko računalniki sami obdelujejo podatke, kar brskalnikom in drugim programom omogoča, da lažje najdejo, izmenjujejo in združujejo informacije. Semantični splet temelji na viziji Tima Berners-Leeja[72] o spletu kot univerzalnem mediju za izmenjavo podatkov, informacij in znanja. Trenutno je na spletu objavljenih ogromno podatkov, ki pa jih računalniški sistemi ne znajo zadovoljivo interpretirati, če programska oprema ni napisana z namenom izrabe podatkov na točno določenem spletišču; cilj semantičnega spleta je tako svetovni splet obogatiti z informacijami, ki bi jih lahko računalniki sami obdelovali in s tem pomagali uporabnikom pri izvajanju njihovih nalog[73].

Pri semantičnem spletu imajo ključno vlogo ontologije. Ontologije (ang. ontologies) so formalne in soglasno sprejete konceptualizacije, ki omogočajo skupno razumevanje vsebinskega področja (ang. domain); to se lahko izmenjuje med ljudmi in aplikacijskimi sistemi (Fensel, 2004, str. 4), pri čemer je konceptualizacija izraz za abstraktni model določenega fenomena, s katerim so določeni vsi pomembni pojmi (koncepti) tega fenomena in povezane hierarhije. Ontologije so torej formalna strukture, ki podpirajo izmenjavo in ponovno uporabo znanja in jih uporabljamo za pomensko (semantično) predstavitev (delno) strukturiranih dokumentov (Horvat, 2003, str. 32).

4.4.1.1 Standardi

Ideja semantičnega spleta temelji na skupnih formatih za integracijo in združevanje podatkov iz različnih virov ter na jeziku za opisovanje povezav med podatki in objekti iz vsakdanjega sveta. Formati in jezik so določeni s tremi standardi (W3C Semantic Web Activity, 2007):

- **Resource Description Framework** (RDF) je preprosti model za predstavitev in izmenjavo informacij na spletu[74];

---

[71] Semantika ali pomenoslovje je nauk o pomenu, ki je izražen z jezikom, kodo ali kako drugo predstavitveno obliko.

[72] Timothy John "Tim" Berners-Lee je izumitelj svetovnega spleta in direktor konzorcija World Wide Web (W3C).

[73] Primer takega sodelovanja računalnika in uporabnika je natančnejše delovanje spletnih iskalnikov na osnovi semantičnih informacij.

[74] RSS je ena izmed najpriljubljenejših aplikacij standarda RDF.

- **RDF Schema** (RDFS) je jezik za predstavitev ontologije oz. znanja in opisuje lastnosti (relacije) in razrede RDF, omejitve za območja in zaloge vrednosti (pri lastnostih) ter relaciji podrazred in podlastnost;

- **Web Ontology language** (OWL)[75] kot jezik nadgrajuje RDFS in s svojim besednjakom omogoča tudi zapletenejše relacije za opisovanje lastnosti. Načrtovan je tako, da omogoča preslikavo na bogato opisno logiko (ang. description logic), kar omogoča formalni opis semantike.

Podatkovni model RDF pozna tri tipe objektov: osebke (ang. subjects), povedke (ang. predicates) in predmete (ang. objects). Osebek je entiteta, na katero se lahko sklicujemo prek spletnega naslova (URL ali URI)[76], in je element, ki ga opisujemo z izjavami RDF. Povedek vzpostavlja povezavo med osebkom in predmetom ter vrednost predmeta določa kot značilnost osebka (Fensel, 2004, str. 19). RDF-podatke običajno predstavljamo s sintakso XML, možno pa je tudi njihovo predstavljanje v relacijski obliki trojic (npr. N-Triples) (glej prilogo 6).

4.4.2 Semantika spletišča

Pristop k spletnemu poosebljanju na podlagi rudarjenja po podatkih o uporabi spleta se je pokazal za učinkovit način poosebljanja v primerjavi s starejšimi pristopi na osnovi filtriranja (Anand, Mobasher, 2005; Mobasher, 2005; Markellou, Rigou, Sirmakessis, 2005; Eirinaki, Vazirgiannis, 2003; Pierrakos et al., 2003; Mobasher, Cooley, Srivastava, 2000; Mulvenna et al., 2000, idr.). Vseeno pa ima tak pristop k poosebljanju nekaj slabosti (Dai, Mobasher, 2005; Markellou, Rigou, Sirmakessis, 2005); med temi velja omeniti problem novega predmeta (ang. new item problem), težave, ki nastanejo, če je podatkov malo in če se vsebina spletišča pogosto spreminja, ter vsebinske slabosti, ki izhajajo iz poosebljanja zgolj na osnovi transakcijskih podatkov. Dai in Mobasher (2005) trdita, da uporaba semantičnega znanja v sistemih za spletno poosebljanje lahko vodi do globljega odnosa med obiskovalcem in spletiščem, saj integracija takega znanja sistemom omogoča izvajanje dodatnih poosebitev na osnovi vsebinskih značilnosti objektov, ki se priporočajo, hkrati pa ponuja zmožnost pojasnjevanja dejanj obiskovalcev. Da pa bi lahko uporabnikom priporočali različne tipe kompleksnih objektov na osnovi njihovih osnovnih lastnosti in atributov, mora biti sistem sposoben označevati uporabnikove segmente in objekte na globlji pomenski ravni z uporabo področnih ontologij, in ne zgolj na osnovi ključnih besed, ki se že uporabljajo v nekaterih aplikacijah spletnega poosebljanja.

---

[75] Poleg jezika OWL obstajajo še drugi ontološki jeziki, kot so XOL, OIL in DAML+OIL (Fensel, 2004).

[76] Za uporabnost izjav RDF je nujen dogovor o semantiki identifikatorjev virov (ang. resource identifiers). RDF te semantike sam po sebi ne vključuje, najpogosteje pa se uporablja besednjak v skladu z *Dublin Core Metadata Initiative* (Resource Description Framework, 2007).

Kot sem že omenil, je semantika spletišča formalen opis vsebine različnih spletnih strani, ki sestavljajo spletišče, in ga lahko predstavimo z ontologijami in metapodatki (Berend, Hotho, Stumme, 2002, str. 265). Ontologije spletišča običajno vključujejo koncepte, urejevalne relacije med njimi (konceptualne hierarhije, ang. conceptual hirerachies) ter druge relacije med koncepti, ki obstajajo v področju, ki ga predstavlja spletišče (Dai, Mobasher, 2005, str. 279) in jih lahko sestavimo na osnovi različnih meril; v literaturi (za podrobnosti glej Berendt, 2002) so tako opisani primeri spletnih ontologij, ki temeljijo zgolj na vsebinskem področju, na tipični vsebini strani, na strukturi in funkciji strani ali pa na področnem modelu dogodkov (ang. domain event model)[77].

Največji izziv v sistemih za poosebljanje naslednje generacije je učinkovita integracija semantičnega znanja iz področnih ontologij v različne dele procesa rudarjenja po podatkih o uporabi spleta, vključno s pripravo podatkov, odkrivanjem vzorcev in fazo priporočanja. Da bi bilo to mogoče, je treba v sam proces vključiti tri pomembne dodatne aktivnosti (Berendt, Hotho, Stumme, 2002; Dai, Mobasher, 2005; Mobasher, 2005):

- **Pridobivanje področne ontologije.** Proces pridobivanja, vzdrževanja in nadgrajevanja ontologije imenujemo grajenje ontologije (ang. ontology engineering) ali učenje ontologije (ang. ontology learning); za majhna spletišča je možno ta proces izvajati ročno ali polavtomatsko, pri večjih pa ga je nujno avtomatizirati. Za spletišča, pri katerih se vsebina generira dinamično iz podatkovne baze, je mogoče ontologije sestaviti dokaj enostavno[78], pri ostalih pa je treba uporabiti metode rudarjenja po vsebini in rudarjenja po strukturi spleta (glej Berend, Hotho, Stumme, 2002).

  Rezultat te faze procesa je nabor formalno definiranih ontologij, ki natančno predstavljajo semantiko spletišča. Ontologije lahko predstavimo na različne načine, pri čemer izbira neposredno vpliva na različne faze rudarjenje; najpogosteje se za predstavljanje uporabljajo modeli vektorskega prostora (Loh et al., 2000), deskriptivna logika (OWL, DAML+OIL idr.) (Fensel, 2004), logika prvega reda (ang. first order logic) (Craven et al., 2000), relacijski modeli (Dai, Mobasher, 2005), verjetnostni relacijski modeli (Getoor et al., 2001) in verjetnostni markovski modeli (Anderson, Domingos, Weld, 2002).

- **Sestavljanje baze znanja.** Na sestavljanje baze znanja lahko gledamo kot na grajenje množice preslikav med koncepti ali relacijami in objekti spletišča (Dai, Mobasher, 2005, str. 281). Namen te faze je med stranmi, ki sestavljajo spletišče, poiskati primerke konceptov in relacij[79], tako da jih lahko uporabimo za izvajanje različnih nalog rudarjenja. Pri tem se lahko le deloma zanesemo na ročno označevanje

---

[77] Iskanje, izbiranje, dodajanje v košarico, plačevanje …

[78] Relacijska shema podatkovne baze, ki je sestavljena iz več tabel in tujih ključev, ki semantično povezujejo relacije med njimi, je sama po sebi že primer ontologije in jo je treba zgolj pretvoriti v želeno obliko.

[79] Faza sestavljanja baze znanja se od tu imenuje tudi učenje primerkov (ang. instance learning).

dokumentov[80], zato so najprimernejši pristopi za klasificiranje besedil, ki so poznani v okviru disciplin rudarjenja po vsebini in strukturi spleta.

- **Z znanjem izboljšano rudarjenje po podatkih.** Končni cilj uporabe semantičnega znanja v procesu rudarjenja po podatkih je povečanje njegove učinkovitosti. Z uporabo ontologij lahko povečamo natančnost klasificiranja in razvrščanja v skupine ter najdemo zanesljivejše vsebinske vzorce (Dai, Mobasher, 2005, str. 281), hkrati pa nam pomagajo pri interpretaciji in analizi odkritih vzorcev. Podrobnejši pregled koristi semantičnega znanja pri spletnem poosebljanju podajam v naslednjem razdelku.

4.4.3 Uporaba semantičnega znanja v procesu rudarjenja po podatkih o uporabi spleta

Napisal sem že, da lahko semantično znanje izboljša kakovost poosebitev in prilagoditev spletišča, saj nam pomaga bolje razumeti vzroke za navigacijsko vedenje obiskovalcev. Semantično znanje, shranjeno v bazi znanja, pa ni koristno samo v fazi poosebljanja, ampak ga lahko uporabimo v vseh treh osnovnih korakih odkrivanja znanja o uporabi spleta.

4.4.3.1 Faza priprave podatkov

Osnovna naloga v fazi priprave podatkov je iz podatkov izločiti nepotrebne podatke in jih preoblikovati v obliko, ki je primerna za izvajanje izbranih nalog podatkovnega rudarjenja. Če v ta proces uvedemo semantično znanje, lahko v podatkih o uporabi npr. lažje identificiramo robote, seje in semantično pomembne transakcije, kar lahko izboljša učinkovitost spletnega poosebljanja, kot je bilo to pokazano v Mobasher et al. (2002). V tej fazi lahko izvajamo tudi preslikavo ravni konceptov v raven ogledov strani (Dai, Mobahser, 2005, str. 290) in tako preoblikujemo vektorje transakcij (glej poglavje 1.3); rezultat je vektor $\vec{t}' = \langle w^t_{o_1} w^t_{o_2}, \ldots, w^t_{o_k} \rangle$, kjer je o_j semantični objekt, ki se pojavlja v j-tem ogledu strani, $w^t_{o,j}$ pa je utež, ki določa pomen objekta v transakciji. Omenjeni semantični objekti so lahko koncepti na različni ravni hierarhije ali pa objekti, ki predstavljajo primerke teh konceptov.

4.4.3.2 Faza odkrivanja vzorcev

Uspešna uporaba področnega znanja zahteva nadgradnjo osnovnih algoritmov podatkovnega rudarjenja, da lahko le-ti prepoznavajo kompleksne semantične objekte in tako generirajo »semantične« vzorce uporabe, ki ponujajo večjo fleksibilnost (zaradi neodvisnosti od identitet predmetov), a ponujajo tudi nove izzive (razvoj nadgradljivih in učinkovitih algoritmov) (Berend, Hotho, Stumme, 2002, str. 271). Pri razvrščanju v skupine nam uporaba semantike

[80] Poleg količine dokumentov, ki jih je treba označiti (obstoječih in novih), se pojavlja težava tudi zaradi subjektivnosti ročnega označevanja (Doctrow, 2001).

poveča lahko učinkovitost iskanja skupin; tudi če v transakcijah dveh uporabnikov ni skupnih predmetov, ju imamo lahko vseeno za podobni, če so si posamezni predmeti v transakcijah med seboj semantično podobni. Seveda pa tak pristop zahteva spremembo metod za izračun podobnosti med (semantičnimi) vektorji. Dai in Mobasher (2005) predlagata pristop, pri katerem bi semantično podobnost med objektoma *i* in *j* definirali kot linearno kombinacijo podobnosti na ravni posameznih atributov:

$$SemSim(i,j) = \alpha_1 * Attrib_1 Sim(i,j) + \alpha_2 * Attrib_2 Sim(i,j) + \ldots$$ [81]

Drug možni pristop za uporabo semantičnih informacij v procesu odkrivanja vzorcev je uporaba verjetnostne prikrite semantične analize (ang. probabilistic latent semantic analysis, PLSA) (Jih, Zhou, Mobasher, 2004), saj lahko na osnovi sopojavljanja vzorcev ogledanih strani v uporabniških sejah odkrije prikrite semantične povezave med različnimi uporabniki in stranmi. Ker so povezave merjene v obliki verjetnosti, jih lahko z uporabo verjetnostnega sklepanja uporabljamo tudi v fazi analize vzorcev in poosebljanja.

4.4.3.3 Faza priporočanja

Izkoriščanje področnega znanja nam lahko v tej fazi pomaga pri podrobnejšem pojasnjevanju vzorcev uporabe in pri izločanju nepomembnih vzorcev. Semantično poosebljanje je lahko ločeno od semantičnega odkrivanja vzorcev (kar pomeni, da iz odkritih vzorcev samo izločamo semantično neprimerne predloge), učinkovitejši pa je pristop na osnovi integracije področnih profilov uporabe in semantično razširjenih uporabniških profilov in njihovih preslikav na realne spletne objekte (Dai, Mobasher, 2005); ta pristop je računsko dovolj učinkovit, da ga lahko uporabimo za izvajanje v stvarnem času, hkrati pa omogoča tudi izkoriščanje strukturnih povezav med posameznimi ontološkimi razredi. Za priporočanje se lahko na osnovi verjetnostnega sklepanja uporabljajo tudi koncepti, poznani iz verjetnostne prikrite semantične analize (Jin, Zhou, Mobasher, 2004).

4.5 ZAHTEVE ZA IZDELAVO INTEGRIRANE ARHITEKTURNE REŠITVE POOSEBLJANJA SPLETIŠČA Z UPORABO RUDARJENJA PO SPLETU

Na osnovi doslej zapisanega lahko opišemo shemo tipičnega sistema za spletno poosebljanje z uporabo rudarjenja po podatkih o uporabi spleta. Kot je prikazano na sliki 7, je sistem zasnovan ciklično, pri čemer se določene faze izvajajo nepovezano v zalednih sistemih, spet druge pa v stvarnem času, rezultat vsake faze procesa pa je hkrati tudi osnova za izvedbo naslednjega koraka v procesu.

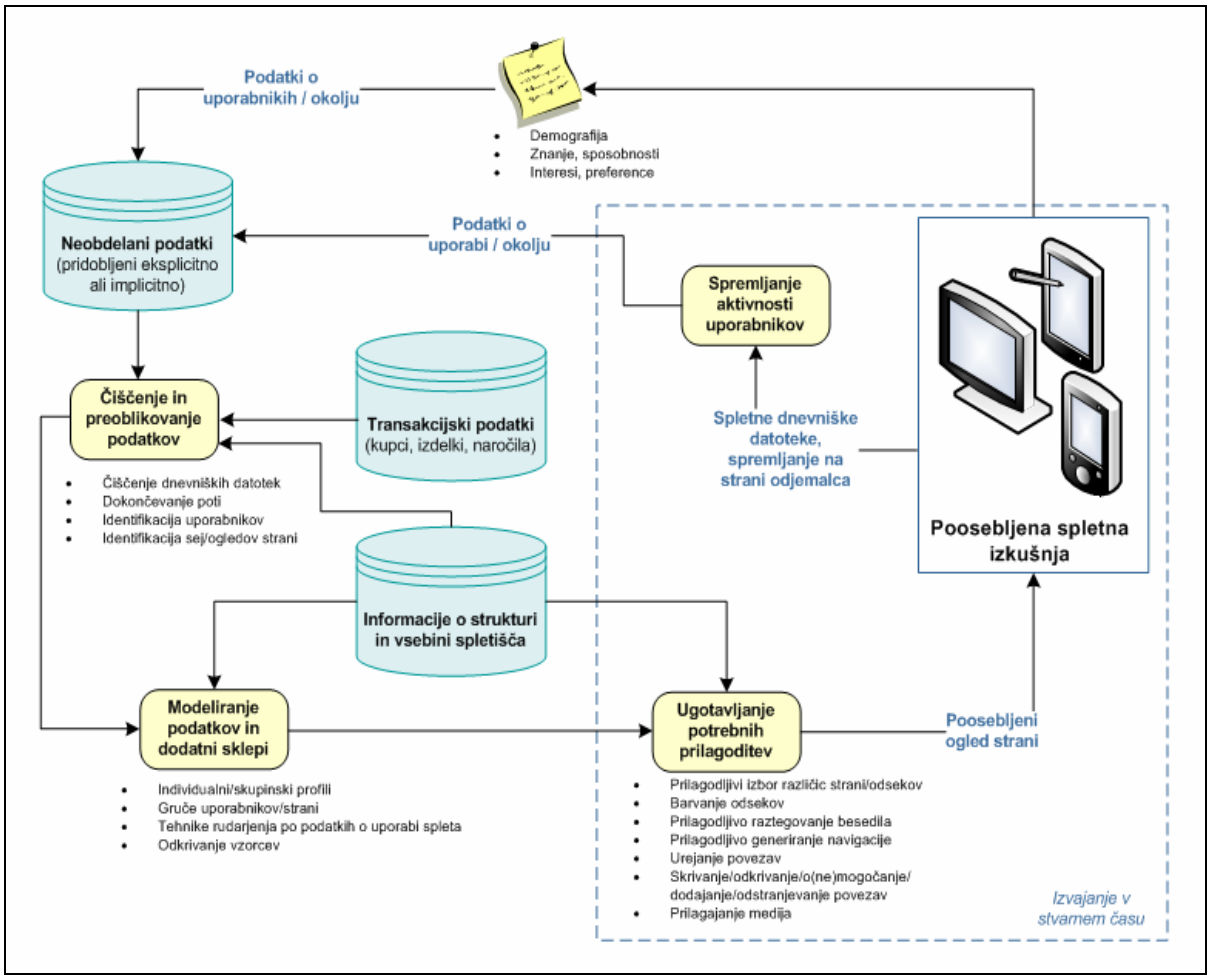
Celoten proces se začne z zbiranjem podatkov o uporabnikih in spremljanjem njihove

[81] α_i predstavlja vnaprej definirano utež za atribut *i*.

aktivnosti; če je sistem sposoben prepoznavati uporabnike, potem so lahko podatki odraz navigacijskega obnašanja skozi daljše časovno obdobje, v nasprotnem primeru pa gre samo za analizo obnašanja v okviru posameznega obiska. Podatki o uporabi in uporabnikih se nato prečistijo in preoblikujejo v obliko, ki je primerna za aplikacijo algoritmov podatkovnega rudarjenja, pri čemer je posebej pomembna čim natančnejša identifikacija uporabnikov in uporabniških sej. Rezultat faze modeliranja so skupinski in/ali individualni profili uporabnikov ter različni vzorci uporabe, ki jih v naslednjem koraku uporabimo za ugotavljanje potreb uporabnika in primernih prilagoditev spletišča ter nato tudi za poosebljanje vsebine. S spremljanjem odziva uporabnika na poosebljeno vsebino dobimo nove podatke o uporabniku, o njegovih željah in potrebah ter druge povratne informacije, ki jih nato uporabimo v naslednjem ciklu poosebljanja spletne izkušnje.

Poosebljanje z uporabo rudarjenja po podatkih o uporabi spleta lahko sicer temelji zgolj na podatkih o obiskih spletišča, bolj pogosto pa se za kakovostnejše izvajanje posameznih nalog v okviru procesa poosebljanja uporabljajo tudi drugi podatki – največkrat so to podatki o vsebini in strukturi spletišča ter različni transakcijski podatki. V fazi priprave podatkov nam ti dodatni podatki pomagajo pri lažji identifikaciji uporabnikov in sej, pri modeliranju in prilagajanju vsebine pa preprečujejo oblikovanje nesmiselnih ali neprimernih pravil in povečujejo kakovost spremembe vsebine.

Slika 7: Shema sistema za spletno poosebljanje z uporabo rudarjenja po spletu



Vir: prirejeno po Markellou, Rigou, Sirmakessis, 2005, str. 30, in Mobasher, 2007, str. 13

Celoten proces je kompleksen in obsežen, zato je potrebno za uspešno uporabo spletnega poosebljanja zagotoviti njegovo integracijo v obstoječo arhitekturo spletišča; pri tem se morajo posamezne zahtevane naloge izvajati brez obveznega posredovanja skrbnika spletišča. Hkrati je pomembno tudi, da se je sistem v predvidenih okvirih zmožen sam prilagajati spremembam na spletišču in spremembam vedenja obiskovalcev ter v skladu s tem posodabljati svoje znanje. Osnovne zahteve, ki jim mora v ta namen zadostiti, so:

- **Avtomatizirano zbiranje, čiščenje in obdelovanje podatkov.** Nove podatke o obisku je potrebno na podlagi ugotovitev iz faze načrtovanja procesa dodajati v učno množico zapisov. Najbolj smiselno je definirati procedure ETL[82], ki se izvajajo v vnaprej predvidenih intervalih ter vse zbrane podatke preoblikujejo v končno obliko.

- **Avtomatizirano učenje modelov.** Samodejno posodabljanje znanja, zbranega v

[82] Procedure ETL (Extract Trasform Load) so zaporedja optimiziranih ukazov za izločanje, preoblikovanje in shranjevanje podatkov, ki se izvajajo v posebnih programih ali programskih modulih.

modelih podatkovnega rudarjenja, je ključno za ohranjanje kakovosti spletnega poosebljanja. Osnovni predpogoj za tako integracijo procesa učenja v celoten proces odkrivanja znanja je obstoj programskega vmesnika v orodju za podatkovno rudarjenje (API), ki programom, ki jih pripravijo skrbniki in načrtovalci sistema, omogoča dostop do njegovih funkcij za učenje, prilagajanje in objavljanje modelov. Razvilo se je več standardiziranih programskih vmesnikov, ki sem jih natančneje predstavil v poglavju 3.4.

- **Avtomatizirano poosebljanje spletišča.** V zadnjem koraku v procesu spletnega poosebljanja mora orodje za podatkovno rudarjenje spletišču posredovati ustrezne napovedi glede uporabnikovega vedenja ali njegovih drugih lastnostih, na osnovi katerih lahko programsko okolje spletišča izvede ustrezne prilagoditve. Za izvajanje priporočanja v stvarnem času je potrebno slediti uporabnikovi poti čez spletišče, nato pa orodju za podatkovno rudarjenje v predvidenem formatu posredovati ustrezno oblikovano poizvedbo ter prebrati odgovor. V primeru poosebljanja, opisanem v poglavju 5, sem priporočanje v spletišče integriral tako, da sem pripravil spletno storitev, ki je od spletišča sprejemala podatke o straneh, ki jih je uporabnik že obiskal, nato pa dobljeni seznam priporočil v zapisu XML posredovala funkciji na spletišču, ki je storitev zahtevala.

- **Avtomatizirano vzdrževanje ontologij.** Ontologije kot oblika semantičnega znanja so za proces poosebljanja izjemno pomembne. Tudi če se semantično znanje ne uporablja v fazi modeliranja in izvajanja poosebitev, je le-to nujno potrebno za kakovostno pripravo podatkov. Za ohranjanje kakovosti vhodnih podatkov in s tem kakovosti samih poosebitev je potrebno ontologije samodejno vzdrževati v skladu s spremembami vsebine in topologije spletišča.

5 SPLETNO POOSEBLJANJE NA SPLETIŠČU SURS-A

5.1 O STATISTIČNEM URADU REPUBLIKE SLOVENIJE IN NJEGOVEM SPLETIŠČU

Statistični urad Republike Slovenije (SURS) je glavni izvajalec in povezovalec dela na področju državne statistike. Poleg povezovanja in usklajevanja statističnega sistema sodijo med njegove najpomembnejše naloge še mednarodno sodelovanje, določanje metodoloških in klasifikacijskih standardov, predvidevanje potreb uporabnikov, zbiranje, obdelava in izkazovanje podatkov ter skrb za njihovo zaupnost. Urad s pomočjo ostalih pooblaščenih izvajalcev statističnih raziskovanj zagotavlja organom in organizacijam javne uprave,

gospodarstvu in javnosti podatke o stanju in gibanjih na ekonomskem, demografskem in socialnem področju ter na področju okolja in naravnih virov (Statistični urad Republike Slovenije, 2007). V skladu z razvojem interneta se je leta 1996 tudi SURS odločil izkoristiti novo nastajajoči potencial svetovnega spleta ter svoje podatke ponuditi uporabnikom brezplačno tudi prek lastne spletne strani. Spletne strani so se redno osveževale z novo vsebino, vendar pa so postajale zaradi povečanega in zahtevnejšega povpraševanja po statističnih podatkih in tehnološkega razvoja računalniških komunikacij vsebinsko in tehnološko vedno manj primerne za kakovostno izkazovanje statističnih podatkov.

Te je bil razlog, da se je SURS leta 2003 odločili za temeljito prenovo spletne ponudbe uradnih statističnih podatkov domačim in tujim uporabnikom ter se osredotočil na uporabo sodobnih tehnologij, ki so prijazne do uporabnika, ne pa organizacijsko pogojene. Namen prenove je bil obenem tudi posodobitev predstavitve Statističnega urada, povečanje zaupanja uporabnikov v njegovo delo ter izboljšanje odnosov z medijsko javnostjo (Kožuh, 2004). Vzpostavitev popolnoma prenovljenega spletišča je močno vplivala tudi na notranji proces priprave in objavljanja statističnih podatkov, hkrati pa je povzročila tudi spremembe v uporabniških navadah naših uporabnikov. Internet, ki je skoraj čez noč postal glavni medij pri posredovanju statističnih podatkov in informacij uporabnikom, je krog uporabnikov precej povečal in tako postal osrednja točka sodelovanja z uporabniki (Stražišar, 2005, str. 4).

5.1.1 Vsebina spletišča SURS

Spletišče SURS omogoča uporabnikom enostaven in brezplačen dostop do vseh objavljenih podatkov in publikacij, posebnih orodij ter informacij o državni statistiki ter delu Statističnega urada. Uporabniki se lahko prijavijo na seznam prejemnikov obvestil o novih prvih objavah podatkov, lahko si pripravijo poosebljene RSS-kanale[83] ali pa do podatkov iz različnih aplikacij dostopajo prek spletnih poizvedb.

Podatki, ki jih objavlja SURS, so vsebinsko razvrščeni po načelu statističnih področij in so uporabnikom na voljo v različnih oblikah. Osnovna oblika posredovanja podatkov končnim uporabnikom je t. i. *Prva objava*, ki se dnevno objavlja izključno na spletnih straneh SURS-a in vsebuje najosnovnejše (prvič objavljene) podatke in grafične prikaze. Isti podatki so z zamikom v podrobni obliki dostopni prek *Si-Stat baze*, ki je interaktivno orodje za samostojno pripravo, analizo in izvoz statističnih podatkov. Podatki so dostopni tudi v številnih drugih periodičnih in priložnostnih publikacijah; med temi je najpomembnejši *Statistični letopis*. Objava nekaterih izmed teh publikacij je prilagojena značilnostim svetovnega spleta, večina pa je na voljo izključno v PDF obliki.

---

[83] RSS (Real Simple Sydication) je tehnologija na podlagi standarda XML, ki uporabnikom omogoča, da so samodejno obveščeni o temah, ki jih zanimajo, ne da bi jim bilo potrebno obiskati spletno stran.

Konceptu statističnih področij v osnovi sledi tudi struktura spletišča SURS[84]; za vsako področje je pripravljena posebna spletna stran, od koder so dostopni vsi podatki s tega področja in vse druge povezane vsebine, in sicer poleg že omenjenih objav in publikacij tudi metodološka pojasnila, podrobni pregledi in analize, leksikon statističnih izrazov, daljše časovne vrste ipd. Ker vsi uporabniki niso dovolj dobro seznanjeni s statističnimi področji, je večina vsebin dostopna tudi prek posebnih tematskih strani (npr. Publikacije, Metodološka pojasnila, Kazalniki idr.).

Posebno popestritev vsebine spletišča SURS predstavljajo različna interaktivna orodja, ki so bila pripravljena z namenom, da bi se širše promoviralo delo Statističnega urada in vsebino posameznih statističnih raziskovanj, za izobraževanje uporabnikov statističnih podatkov ali pa, da bi se omogočilo dostop do različnih statističnih vsebin, ki jih sicer ne objavljamo. Spletne strani teh orodij, med katerimi velja omeniti zlasti *Bazo rojstnih imen in priimkov*, *Bazo rojstnih dni*, *Prebivalstveno uro*, *Kalkulator inflacije* in *Dinamične preračune kazalnikov*, so od vsega začetka med najbolj obiskanimi stranmi na spletišču SURS.

V želji, da bi povečali zaupanje uporabnikov v statistiko in uradne statistične podatke, si je SURS zamislili tudi *Vodič po statistiki*; ta je leksikon statističnih izrazov in definicij, v katerem so opisi strokovnih izrazov pripravljeni tako za splošne kot tudi strokovne uporabnike in so povezani s posameznimi področji statistike. V skladu z *Zakonom o informacijah javnega značaja* SURS na svojem spletišču objavlja tudi vse potrebne informacije o delovanju sistema državne statistike ter o izvajalcih programa državne statistike, poročila in predloge programa ter informacije o rednem delovanju Statističnega urada.

5.2 NAMEN IN VRSTE POOSEBLJANJA SPLETIŠČA STATISTIČNEGA URADA REPUBLIKE SLOVENIJE

Načrtovanje tako velikega in vsebinsko raznolikega spletišča, kot je spletišče SURS, je zapleteno opravilo. Z namenom, da bi spletišče zadovoljevalo potrebe čim širšega kroga uporabnikov, je projekt prenove potekal ob tesnem sodelovanju z uporabniki statističnih podatkov in ob upoštevanju njihovih predlogov in mnenj. S spremljanjem odzivov uporabnikov in nenehnim izboljševanjem njihove izkušnje smo sistematično nadaljevali tudi po koncu projekta; z rednimi anketami o zadovoljstvu uporabnikov želi SURS prepoznati skupine uporabnikov, ugotoviti, kakšne so njihove potrebe in katere storitve pogrešajo, ter posledično ustrezno izboljšati kakovost spletišča.

V prvem letu po prenovi so uporabniki posebej poudarili opaženo povečanje uporabnosti

[84] Statističnih področij je 29 in so zaradi večje preglednosti na spletišču urejena v 4 širše tematske sklope (demografsko-socialno področje, ekonomsko področje, okolje in naravni viri, splošno).

strani in s tem povečano zadovoljstvo, pa tudi težave, ki so se pojavile zaradi določenih pomanjkljivosti novega spletišča; zaradi množice podatkov, ki jim je nenadoma postala dostopna na spletnih straneh, so opozorili predvsem na potrebo po večji preglednosti spletišča, ki bi omogočila lažje iskanje in pridobivanje statističnih podatkov in informacij (Stražišar, 2005, str. 4). Leto pozneje smo izvedli posebno anketo, ki se je nanašala zgolj na zadovoljstvo uporabnikov s spletnimi stranmi (Stražišar, 2005), s katero smo želeli pridobiti dodatne informacije o navadah in potrebah naših spletnih uporabnikov. Pri interpretaciji rezultatov smo skušali slediti tudi klasifikaciji po Grossenbacherju (2005), ki uporabnike statističnih spletnih strani deli na tri glavne interesne skupine:

- Za 1. skupino, ki jo poimenuje *turisti* (ang. tourists), je značilen splošen interes za uradno statistiko, zato sem poleg osnovnošolcev in dijakov uvršča še medije in splošno javnost.

- Za 2. skupino, ki jo imenuje *pridelovalci* (ang. farmers), je značilen poslovni interes za uradno statistiko, zato sem uvršča uporabnike, ki so zadolženi za sprejemanje odločitev (ang. decision makers).

- V 3. skupino, ki jo imenuje *rudarji* (ang. miners), uvršča uporabnike s fakultet, svetovalce, vladne institucije, saj je zanje značilen predvsem raziskovalni interes.

Na osnovi razlogov za obisk spletne strani, ki smo jih identificirali v anketi, ocenjujemo, da med uporabniki spletišča prevladuje druga in tretja skupina uporabnikov (»pridelovalci« in rudarji«). Spletišča večinoma ne obiščejo zgolj slučajno med brskanjem po internetu, temveč so njihovi razlogi za obisk precej bolj jasni; pri tem želene informacije večinoma najdejo, a se pri iskanju srečujejo s težavami. Predvsem tisti uporabniki, ki spletišče obiskujejo manj pogosto, mlajši uporabniki, ki šele pridobivajo izkušnje v povezavi z uradno statistiko, ter statistično manj izobraženi uporabniki so svoje predloge za izboljšanje spletne strani usmerili v orientacijo na spletni strani oz. na izboljšanje iskanja statističnih informacij. Po drugi strani naprednejši uporabniki (»rudarji«) pogrešajo več orodij za lastno analizo podatkov, bolj poglobljene analize in bolj specifične storitve; vse to tako nas kot tudi same uporabnike navaja k razmišljanju o uvedbi poosebljanja in prilagajanja vsebine osnovnim profilom uporabnikov in njihovim glavnim preferencam.

Na osnovi zgoraj omenjenih ugotovitev o pričakovanjih in težavah uporabnikov ter usmerjenosti SURS-a k izboljševanju njihove izkušnje je upravičena domneva, da ima uporaba rudarjenja po spletu na SURS-u velik potencial. Prva izmed aplikacij podatkovnega rudarjenja, ki bi vsekakor pripomogla k boljšemu poznavanju uporabnikov storitev, ki jih ponuja SURS, in posledično k lažjemu zadovoljevanju njihovih potreb, je uporaba algoritmov za iskanje značilnih skupin; na SURS-u je trenutno v veljavi segmentacija, ki uporabnike na podlagi njihovega interesa za uradno statistiko deli na osem skupin (SURS: Navodilo za

vodenje nove segmentacije podatkov, 2004); te so bile oblikovane na osnovi različnih raziskav, vendar pa njihova primernost v spletnem okolju še ni bila preverjena. Z uporabo algoritmov za odkrivanje skupin bi lahko oblikovali skupine uporabnikov, ki izkazujejo podoben interes za vsebine, ki jih SURS objavlja na svojem spletišču, in na tej osnovi bi (če bi prišlo do neujemanja) prilagodili uradno segmentacijo ter nabor storitev, ki jih tem uporabnikom ponujamo. Na drugi strani se ponuja možnost uporabe rudarjenja po spletu za spletno poosebljanje, saj bi lahko na ta način uporabnikom olajšali iskanje statističnih podatkov in jim pomagali pri doseganju njihovih (večinoma jasno) zastavljenih ciljev – ne samo da bi lahko nevešči uporabniki hitreje našli želene informacije in bi jih lahko lažje razumeli, ampak bi lahko tudi naprednejši uporabniki spletišče uporabljali bolj učinkovito in bi lahko imeli več koristi od informacij, ki jih objavljamo.

5.2.1 Predlagane oblike poosebljanja spletišča SURS-a

Kot sem napisal, bi lahko SURS s poosebljanjem svojega spletišča olajšal težave, ki jih imajo uporabniki pri iskanju in uporabi statističnih informacij, ter povečal njihovo zadovoljstvo in zaupanje v uradno statistiko, hkrati pa tudi izboljšal kakovost statističnih podatkov glede njihove ustreznosti, dostopnosti in jasnosti (SURS: Kakovost statistike, 2007). Ne nazadnje bi poosebljanje spletišča pripomoglo tudi k uspešnosti enega izmed strateških ciljev SURS-a – zagotavljanju enakega dostopa do statističnih rezultatov vsem uporabnikom na enak način in na isti datum (SURS: Politika diseminacije statističnih podatkov, 2007).

Ključno vprašanje, ki se zastavlja ob razmišljanju o poosebljanju spletišča SURS-a, je, katere oblike poosebljanja so primerne in učinkovite za izboljševanje njegove uporabnosti in uporabniške izkušnje ter so ne nazadnje tudi srednjeročno izvedljive. Na osnovi namena in strukture spletišča, rezultatov anket o zadovoljstvu uporabnikov, interesa v Statističnem uradu in razpoložljivosti kadrov v nadaljevanju na osnovi klasifikacije, prikazane v poglavju 4.1.1, predlagam šest oblik poosebljanja, ki bi lahko pomagale pri lajšanju dostopa do statističnih podatkov in izboljševanju izkušnje spletnih uporabnikov, hkrati pa so tudi izvedljive.

5.2.1.1 Pomnjenje

- Ustvarjanje zaznamkov. S samodejnim ustvarjanjem in prikazovanjem seznama strani, ki jih je uporabnik obiskal v preteklosti, bi lahko uporabnikom, ki stran obiščejo večkrat, hitro ponudili nabor vsebin, ki jih preverjeno zanimajo, in mu s tem skrajšali čas za ponovno iskanje informacij; kot se je pokazalo pri uporabnikih baze Si-Stat, v kateri si lahko registrirani uporabniki shranjujejo svoje poizvedbe po bazi, je taka storitev zelo dobrodošla in do uporabnika prijazna. Za spletno poosebljanje, kot ga pojmujem v tem delu, registracija uporabnikov in njihovo eksplicitno shranjevanje strani in poizvedb z namenom, da ustvarijo zbirko zaznamkov, ne zadostuje, čeprav seveda lahko dopolnjuje proces samodejnega sestavljanja seznama zaznamkov; ta je v

osnovi sestavljen iz treh korakov:

1. Unikatno identificiraj uporabnika (na prvi strani na osnovi prijave ali piškotkov, kasneje na osnovi identifikatorjev seje).
2. Če je uporabnik prepoznan in se s spremljanjem aktivnosti strinja, potem prikaži zaznamke (če je prikaz dovoljen) in shranjuj podatke o aktivnosti uporabnika.
3. Če uporabnik ni prepoznan, a je unikatna identifikacija možna, potem uporabnika vprašaj, ali želi uporabljati to obliko poosebljanja. Če je odgovor pozitiven, začni s spremljanjem aktivnosti.

Pri prikazu seznama zaznamkov obstaja več možnih pristopov; sam menim, da enostavno prikazovanje zgodovine v slogu najpriljubljenejših brskalnikov ne zadostuje, saj uporabnikom ne nudi nobene dodane vrednosti. Predlagam prikazovanje zaznamkov v posebni lebdeči plasti (ang. floating layer)[85], ki je prvič privzeto prikazana, kasneje pa je njen prikaz odvisen od tega, ali uporabnik želi, da je vidna ali skrita. Uporabniku lahko ponudimo dve vrsti zaznamkov: seznam vsebinskih strani[86], ki jih je obiskal, in seznam poizvedb, ki jih je izvajal v bazi Si-Stat. Pri tem posebno dodano vrednost ponuja možnost različnih prikazov zaznamkov:

→ *kronološko* (seznam zadnjih strani/poizvedb, seznam strani/poizvedb v določenem obdobju),
→ *po priljubljenosti* (seznam največkrat obiskanih strani/poizvedb),
→ *po kategorijah v drevesnem pregledu* (v skladu z ontologijo spletišča),
→ *po vsebini* (uporaba rudarjenja po vsebini za oblikovanje gruč zaznamkov),
→ *po navigacijski podobnosti* (uporaba rudarjenja po spletu za oblikovanje gruč zaznamkov glede na njihovo sopojavljanje v transakcijah uporabnikov).

Kot sem že omenil, bi lahko samodejno spremljanje zaznamkov dopolnili z možnostjo ročnega dodajanja strani med priljubljene, pa tudi z ocenjevanjem zaznamkov in možnostjo njihovega urejanja.

- Hranjenje vpisanih/izbranih vrednosti. Za izvajanje klasične oblike tega poosebljanja na spletišču SURS sicer ni veliko priložnosti, saj uporabnikom za doseganje njihovih ciljev na spletnih straneh ni treba izpolnjevati veliko obrazcev. Vseeno pa ne gre prezreti njenega potenciala pri pomoči uporabnikom pri izdelavi poizvedb v Si-Stat bazi in v obrazcu za napredno iskanje po spletišču (v pripravi). V bazi Si-Stat bi lahko uporabnikom prihranili čas s predizbiro možnosti glede načina prikaza na zaslonu in predizbiro statističnih spremenljivk, pri obrazcu za iskanje pa s predizpolnjevanjem možnosti glede območja in načina iskanja. Še bolj se uporabnost takega poosebljanja pokaže v primeru, da vanj vključimo tudi hranjenje izbir pri komunikaciji uporabnika z uporabniškim vmesnikom – npr. pri določanju vidnosti posameznih delov besedila

[85] Dodatno bi se lahko uporabniku ponudil poseben programski dodatek za njegov brskalnik, ki bi omogočil prikazovanje zaznamkov v posebnem podoknu (ang. pane).
[86] Gre za klasifikacijo predstavljeno v poglavju 3.2.3, pri kateri se strani delijo na vsebinske in navigacijske.

(glej poglavje 5.2.1.3).

Načinov za izvajanje tega načina pomnjenja je več; najenostavnejše, a vseeno dokaj učinkovito, je pomnjenje vrednosti v okviru posamezne seje, pri katerem si mora sistem v posameznem obrazcu zapomniti zgolj zadnje vpisane vrednosti in pri katerem uporabnika sploh ni treba posebej identificirati. Dolgoročnejše pomnjenje je bolj zapleteno, saj je pri njem potrebno identificiranje uporabnikov in dolgoročno hranjenje vzorcev izpolnjevanja obrazcev, vendar pa omogoča predizpolnjevanje obrazcev v okviru več obiskov in tudi ob prvem prikazu obrazca v okviru posamezne seje. Tipičen primer, ki bi ga lahko uporabili na spletišču SURS, je samodejni izbor možnosti, ki določajo obseg iskanja na podlagi vsebin, do katerih je uporabnik dostopal ob svojih predhodnih obiskih.

5.2.1.2 Vodenje

- Priporočanje vsebine. Vodenje uporabnika po spletišču s priporočanjem strani, ki naj jih obišče, bi morala biti po mojem mnenju ena izmed temeljnih oblik poosebljanja na spletišču SURS in je še posebej primerna za manj spretne uporabnike. Za oblikovanje seznama priporočil je najprimernejša uporaba enega izmed algoritmov odkrivanja zaporednih vzorcev in mora biti enako nevsiljivo kot že opisano prikazovanje zaznamkov.

  Celoten proces oblikovanja priporočil z rudarjenjem po podatkih o uporabi spleta in učinkovitost te oblike poosebljanja sta podrobno predstavljena v poglavju 5.3.

5.2.1.3 Prilagajanje

- Prilagajanje vsebine in strukture. S poosebljanjem v smislu samodejnega prilagajanja vsebine in strukture bi lahko na spletišču SURS rešili veliko težav, ki se pojavljajo zaradi razkoraka med razumevanjem pri različnih skupinah uporabnikov, in s tem bolj zadovoljili uporabnike; predvsem manj vešči uporabniki so namreč že večkrat opozorili na dostop do posebnih strani, ki bi bile namenjene javnosti, katere statistično znanje je majhno, in na katerih bi bile objavljene vsebine z več enostavnimi grafičnimi prikazi in bolj enostavnimi komentarji z manj zahtevno statistično terminologijo.

  Prilagajanje vsebine je zapleteno, saj zahteva vnaprejšnje oblikovanje skupin uporabnikov (z uporabo algoritmov za razvrščanje v skupine), njihovo podrobno analizo ter temeljito pripravo različnih tipov vsebin in označevanje sklopov obstoječih vsebin z vidika njihove primernosti za posamezne uporabniške skupine. Taka možnost prilagajanja bi omogočila, da bi uporabniku (potem ko bi ga s klasifikacijskimi metodami uvrstili v eno izmed skupin) prikazali njegovemu profilu prilagojene vsebine (posebne strani, raztegljivo besedilo) za posamezne dele besedil.

  Prilagajanje strukture je kratkoročno na SURS-u bolj izvedljivo, ker ne zahteva velikih

vsebinskih sprememb statističnih informacij, rezultatov in drugih vsebin. Na prilagajanje strukture lahko namreč gledamo tudi kot na priporočanje vsebine s pomočjo preurejanja povezav in prioritet posameznih delov strani, pri čemer je to priporočanje omejeno na vsebino prikazane spletne strani. Postopek za tako obliko poosebljanja je naslednji:

1. Na osnovi njegovega navigacijskega vedenja in drugih značilnosti uporabnika dodeli v eno izmed skupin (glej poglavje 3.3.2).
2. Ugotovi, kateri deli izbrane spletne strani bolj ustrezajo uporabnikovim preferencam in njegovim predvidenim željam v okviru obravnavanega obiska.
3. Izvedi ustrezne prilagoditve vsebine – uredi vrstni red povezav na druge strani in poudari zanj bolj smiselne povezave, izpostavi tiste dele vsebin, ki so za uporabnika bolj primerne, degradiraj (ang. demote) manj pomembne vsebine.

Na sliki 8 je prikazan primer prilagajanja strukture in vsebine osnovne spletne strani za pregled publikacij[87] SURS-a za uporabnika, ki je izkazal zanimanje za vsebine s področja *Prebivalstvo*; na levi strani je prikazana trenutna spletna stran s pregledom publikacij, na desni strani pa poosebljena stran. Na slednji so za uporabnika sprva »nesmiselno« urejene povezave na zbirke in posamezne izvode publikacij (levi meni) urejene tako, da izpostavimo publikacije, ki vsebujejo podatke o prebivalstvu. Dostop do ostalih (skritih) publikacij je uporabniku še vedno omogočen, vendar pa konkretni seznami teh publikacij privzeto niso vidni.

Slika 8: Primer prilagajanja strukture in vsebine na spletišču SURS-a

Vir: lastno delo

[87] URL: http://www.stat.si/publikacije

- Prilagajanje načina predstavitve. Z vedno večjim razmahom uporabe mobilnih naprav za brskanje po spletu je treba prikaz vsebin prilagoditi zaslonom teh naprav. Za ta namen je treba pripraviti različne predloge (ang. template) za postavitev vsebine in jih uporabiti, ko uporabnik spletišče obišče z nestandardnim brskalnikom.

5.2.1.4 Podpora učinkovitejšemu izvajanju nalog

- Poosebljeno iskanje. Uporaba iskalnika je ena izmed najpogostejših opravil na spletišču SURS-a. V letu 2006 je bilo v iskalnik vnesenih 312.188 izrazov, pri čemer iskalnik v dobrih 25 % primerov ni našel nobenega zadetka, glede na odzive uporabnikov pa tudi v ostalih primerih pogosto ni bilo relevantnih vsebin. Tudi zato je prenova iskalnika ena izmed prioritetnih nalog SURS-a, njegovo poosebljanje pa eden izmed vidikov prenove. V tem okviru bi bilo potrebno (Kožuh, 2006):
  - → že pri vpisovanju iskalnih izrazov ponuditi nabor njegovih predhodnih »uspešnih« iskanj in nabor najpogostejših iskalnih pojmov njemu podobnih uporabnikov (z uporabo tehnik AJAX);
  - → z uporabo rudarjenja po vsebini spleta (poleg osnovnih tehnik iskanja) poiskati vsebine, ki so najbolj sorodne vnesenemu iskalnemu nizu in uporabnikovim preferencam;
  - → v seznamu zadetkov posebej izpostaviti povezave do vsebin, ki so za uporabnika bolj relevantne (glej prilaganje strukture).

5.3 IZVEDBA POOSEBLJANJA SPLETIŠČA SURS-A Z UPORABO RUDARJENJA PO SPLETU

Med uporabniki, ki s spletiščem SURS-a niso zadovoljni in imajo pri iskanju informacij največ težav, so večinoma tisti, ki spletišča in njegove strukture ne poznajo, velikokrat pa tudi niso dovolj seznanjeni s statistično terminologijo in povezavami med posameznimi statističnimi koncepti. Da bi izboljšali njihovo uporabniško izkušnjo, sem se odločil preveriti, kakšna je učinkovitost na navigacijskem vedenju obiskovalcev temelječega priporočanja vsebine in kakšne so možnosti uvedbe te oblike poosebljenja na spletišče SURS-a; za modeliranje sem uporabil algoritem *Microsoft SequenceClustering* (v nadaljevanju MSSC), ki ga uvrščamo med algoritme za odkrivanje zaporednih vzorcev, ki so za tako poosebljanje najprimernejši (Spilipoulou, Pohle, 2001; Pierrakos et al., 2004; Mobasher, 2007), hkrati pa so njegova dostopnost ter uporaba in integracija v druge aplikacije relativno enostavne. Podrobnosti procesa predstavljam v nadaljevanju poglavja.

5.3.1 Zbiranje podatkov

Na spletnem strežniku Statističnega urada Republike Slovenije je vse od prenove spletišča v

začetku leta 2004 omogočeno spremljanje obiska na spletišču s spletnimi dnevniškimi datotekami, v katere se zapisujejo vsi podatki, določeni z razširjenim zapisom W3C. Kljub dolgi časovni vrsti podatkov, ki so mi bili tako na voljo, sem se odločil, da jih za potrebe rudarjenja po podatkih o uporabi spletišča ne bom uporabil, ampak bom poskušal zbrati podatke na strani odjemalca. Za tak korak sem se odločil, ker menim, da je tak način spremljanja statistike obiska zaneslivejši (glej poglavje 3.1.1).

Pred dokončno odločitvijo za metodo zbiranja podatkov sem preveril učinkovitost vsake izmed omenjenih metod z vidika zajetja uporabniških ogledov strani in ogledov, ki jih sprožajo roboti, ter z vidika uporabe trajnih in začasnih piškotkov, ki so zelo pomembni za proaktivno prepoznavanje strežniških sej. Za spremljanje obiska na strani odjemalca sem pripravil ustrezne ukazne datoteke v jeziku JavaScript (v nadaljevanju tudi CS), za spremljanje statistike na strani strežnika pa v tehnologiji ASP (v nadaljevanju tudi SS) – z obema metodama sem v obdobju od 28. 12. 2006 do 7. 1. 2007 spremljal identično statistiko na enakem naboru strani in rezultate tudi na enoten način zapisovali v bazo. Obe metodi dejansko ne omogočata spremljanja povsem enakega nabora podatkov – na strani odjemalca se npr. ne da zajeti podatka o IP-številki uporabnika, pri strežniških skriptah pa je težavno takojšnje prepoznavanje podpore piškotkom. Prvo težavo sem rešil z dopolnjevanjem podatkov ob vpisu v podatkovno bazo, težavo s piškotki pa z uporabo notranjih okvirov (Kožuh, 2007). Zahtevke robotov sem prepoznaval naknadno s pomočjo podatkov, ki se zbirajo v okviru projekta Browser Capabilities Project (Keith, 2006).

Tabela 3: Primerjava podatkov o ogledih strani, zbranih na odjemalcu (CS) in strežniku (SS)

	CS	SS
Skupno število zabeleženih ogledov	133.026	174.142
Število zabeleženih ogledov s strani »etičnih« robotov	2	49.215
Število zabeleženih ogledov običajnih uporabnikov	133.024	124.927
Delež običajnih uporabnikov, pri katerih je bila mogoča uporaba začasnih piškotkov	99,16 %	99,08 %
Delež običajnih uporabnikov, pri katerih je bila mogoča uporaba trajnih piškotkov	98,35 %	97,74 %

Vir: statistika uporabe spletišča SURS-a, lastni izračuni

Kot je prikazano v tabeli 3 je spremljanje obiska po CS-metodi učinkovitejše kot SS spremljanje. Čeprav 0,34 % uporabnikov ni omogočalo izvajanja jezika JavaScript[88], je bilo na strani odjemalca identificiranih za 6,5 % več zahtevkov običajnih uporabnikov kot na strežniški strani, kar gre pripisati dejstvu, da v spletnih dnevniških datotekah niso zabeleženi zahtevki za ogled predpomnjenih strani.

Največja slabost CS-spremljanja statistike obiska je v tem, da lahko na ta način spremljamo

[88] Podatek temelji na analizi spletnih dnevniških datotek spletišča SURS s programom za analizo statistike obiska SurfStats 8 v obdobju od 28. 12. 2006 do 7. 1. 2007.

samo zahtevke za oglede spletnih strani, ne pa tudi statistike zahtevkov za ogled različnih tipov dokumentov (npr. PDF, DOC, XLS in drugih datotek) in slik. Težavo lahko rešimo na dva načina: zahtevke za datoteke, ki nas zanimajo, lahko preusmerjamo prek posebne prehodne strani ali pa CS statistiko kombiniramo s SS statistko, v kateri so tovrstni zahtevki zabeleženi. Obe rešitvi dajeta podobne rezultate, a je prva dokaj nepraktična, saj zahteva celovito prenovo spletišča, hkrati pa po nepotrebnem zapleta objavljanje novih dokumentov in slik. V večini primerov je zato primernejše naknadno (a posteriori) dopolnjevanje CS statistike s podatki iz spletnih dnevniških datotek.

Na osnovi doslej prikazanih ugotovitev sem se odločil, da za potrebe rudarjenja po podatkih o uporabi spleta podatke o obisku na spletišču SURS spremljam na strani odjemalca z uporabo jezika JavaScript, ki podatke o obisku pošilja na posebno zaledno (ang. back-end) stran, kjer se le-ti ustrezno dopolnijo in shranijo v podatkovno bazo; za identifikacijo zahtevkov dokumentov sem podatke na osnovi identifikatorjev sej naknadno dopolnil z ustreznimi zahtevki iz spletnih dnevniških datotek. Za učno množico sem izbral podatke o obisku na osnovnih slovenskih straneh spletišča[89], zbrane med 23. 1. 2006 in 22. 2. 2007; v tem času je bilo zabeleženih 460.813 zahtevkov za ogled spletnih strani, v dnevniških datotekah pa je bilo 227.248 zahtevkov za ogled drugih dokumentov.

5.3.2 Predobdelava in priprava podatkov

Faza priprave in predobdelave podatkov je bistvena za uspešnost procesa iskanja vzorcev (glej poglavje 3.2), ki se skrivajo v njih, zato sem veliko truda namenil temu, da bi zbrani podatki predstavljali realno sliko navigacijskega vedenja uporabnikov. S tem namenom sem podatke najprej v dveh ciklih prečistil in razmejil v seje ter jih nato preoblikoval v obliko, ki je potrebna za učinkovito aplikacijo algoritma MSSC; tako zaporedje opravil sicer nekoliko odstopa od tradicionalnega pristopa, v katerem se za identifikacijo sej uporabljajo že povsem prečiščeni podatki, vendar pa je učinkovitejše, saj se določeni navigacijski vzorci, ki jih je treba izločiti iz podatkov, pokažejo šele v okviru uporabniških sej.

5.3.2.1 Prvi cikel predobdelave podatkov

Namen prvega čiščenja je iz podatkov izločiti vse podatke, za katere je vnaprej znano, da izkrivljajo sliko navigacijskega vedenja obiskovalcev spletišča. Skupaj sem v tem koraku izbrisal 45.162 zapisov, pri čemer sem upošteval več različnih predpostavk in informacije o načinu delovanja spletišča. Tako sem odstranil:

- vse zahtevke, pri katerih je bil uporabljen kateri koli drug protokol, razen HTTP – ti

[89] V spremljanje niso bile vključene angleške spletne strani ter spletne strani Statističnega letopisa, podatkovne baze Si-Stat, Popisa prebivalstva 2002 in Kataloga regionalnih delitev Slovenije.

zahtevki predstavljajo obiske zaposlenih na SURS-u v okviru procesa objavljanja novih vsebin (709 zapisov);

- vse zahtevke za strani, na katere je uporabnik preusmerjen v primeru kakršnekoli napake ali neobstoječe strani (15.770 zapisov);

- vse zahtevke za strani, na katerih se zgolj obdelujejo vrednosti iz obrazcev in ki ne ponujajo posebej določene vsebine (28.298 zapisov);

- zahtevke za strani, na katerih se statistika ni spremljala, a so bili zaradi različnih napak vseeno zabeleženi (385 zapisov).

Ko so bili podatki grobo prečiščeni, sem se lotil njihovega razmejevanja na uporabniške seje. Kot sem pokazal v poglavju 3.2.2, je za identifikacijo uporabniških sej najbolje uporabiti proaktivni pristop, ki ga je treba ustrezno in postopno dopolnjevati z izbranimi hevrističnimi algoritmi. V skladu s tem sem že v procesu zbiranja podatkov vsak zahtevek za ogled strani ali datoteke opremil z unikatnim identifikatorjem, kar mi je omogočilo, da sem seje identificiral z uporabo algoritma, prikazanega v sliki 8.

Algoritem zahtevke najprej uredi po časovnem zaporedju, nato pa v posamezne uporabniške seje preprosto združi vse zapise, ki imajo enako vrednost identifikatorja seje. Le-ta se uporabnikom, ki omogočajo zapisovanje začasnih piškotkov (99,16 % v učni množici), dodeli ob vstopu na spletišče in poteče 30 minut po njihovem zadnjem dejanju ter tako uporablja pristop časovnega razmejevanja sej, ki je za identificirane uporabnike najprimernejši (glej poglavje 2.2.2). Zapise brez identifikatorjev algoritem najprej razdeli na podmnožice zapisov, ki imajo enak par IP-naslov / uporabniški agent (*Najdi_IpAgent*), saj lahko predpostavljamo, da uporabniki v okviru enega obiska uporabljajo samo en brskalnik in imajo stabilen IP-naslov; pri tem algoritem upošteva tudi možnost, da je lahko uporabniški agent pri zahtevkih za ogled določenih tipov datotek zaradi uporabe t. i. upraviteljev prenosov (ang. download manager) tudi drugačen od uporabniškega agenta, ki ga uporabnik uporablja za brskanje po spletišču. Tako urejeni zahtevki se nato v nasprotju z zahtevki proaktivno identificiranih uporabnikov v seje razmejijo na podlagi podatkov o straneh, ki so bile v okviru seje že ogledane (*OblikujSeje_RefSessionizer*).

Slika 9: Algoritem za razmejevanje sej

Z – časovno urejena množica zapisov o zahtevkih za ogled spletnih strani in dokumentov
IPA – množica ogledov strani uporabnika z enakim parom IP-naslov/agent

1. for each $z_i \in Z$
2. for each z_i.*Session_id* <> NULL
3. DodajZapis_Seja(z_i)

```
4.          end
5.          for each z_j.Session_id = NULL
6.              IpAgent_id = Najdi_IpAgent(z_i.ip, z_i.agent)
7.              DodajZapis_IpAgent(z_i.ip, z_i.agent)
8.          end
9.          for each ipa_j ∈ IPA
10.             OblikujSeje_ RefSessionizer(z_i, ipa_j)
11.         end
12.     end
```

Vir: lastno delo

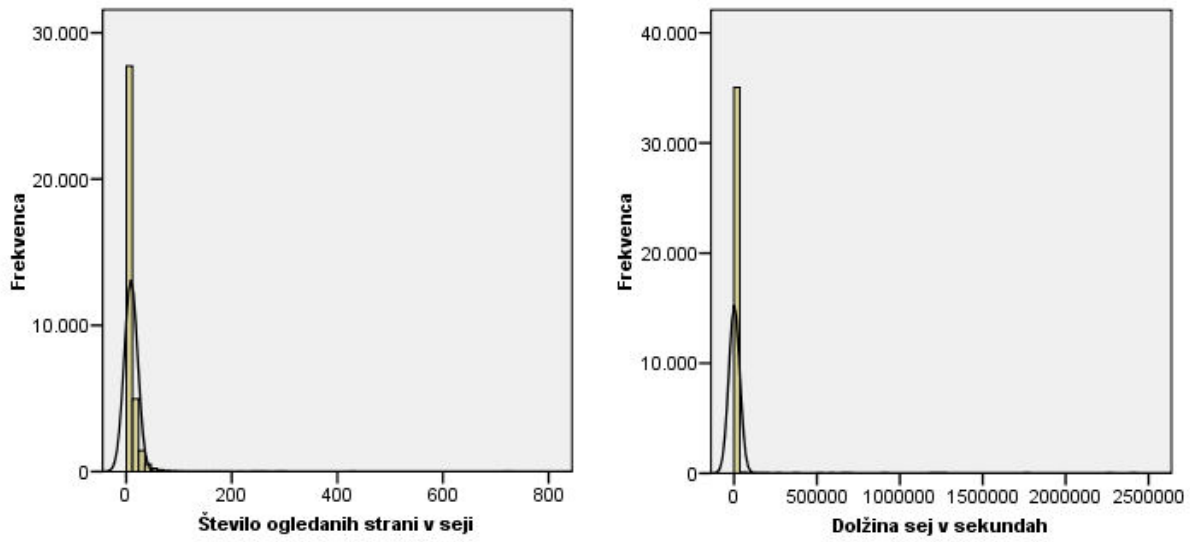
V prvem ciklu mi je uspelo identificirati 66.061 uporabniških sej; v okviru posameznega obiska, ki je v povprečju trajal dobrih 38 minut in 49 sekund, so si obiskovalci v poprečju ogledali 6,29 strani.

5.3.2.2 Drugi cikel predobdelave podatkov

V prvem koraku predobdelave podatkov sem očistil večino nepotrebnih in nezaželenih zahtevkov, vendar pa so podatki še vedno vsebovali skrite zahtevke robotov, zahtevke za oglede nepomembnih strani in druge zahtevke, ki ne predstavljajo prave slike obnašanja uporabnikov. Da b izboljšal kakovosti podatkov, sem zato iz množice učnih podatkov najprej izbrisal vse uporabniške zahtevke, za katere je bilo očitno, da so zabeleženi, ker je obiskovalec stran (zaradi težav) ponovno naložil, in vse uporabniške seje, v katerih je uporabnik odprl samo eno spletno stran. Nato sem posodobil podatke o identificiranih sejah in na tej osnovi zbrisal še tiste zahtevke, ki so preveč odstopali od ostalih.

Kot je prikazano na sliki 10 so podatki številu ogledanih strani v okviru posamezne uporabniške seje in podatki o dolžini sej v sekundah normalno porazdeljeni. Na osnovi statističnega sklepanja sem zato iz učne množice dodatno zbrisal še vse tiste seje, pri katerih je bil vsaj en podatek zunaj 95-odstotnega intervala zaupanja – tj. seje, v okvir katerih je bilo zahtevanih več kot 46 strani, in seje, katerih dolžina je presegala 11.817 sekund.

Slika 10: Histograma števila ogledanih strani v seji in dolžine sej v sekundah



Vir: statistika obiska spletišča SURS

Po končani predobdelavi podatkov je bilo v učni množici 283.543 zapisov, ki so bili razmejeni v 36.020 uporabniških sej; seje so bile v povprečju dolge 28 minut in 39 sekund, vsak uporabnik pa je v okviru svojega obiska v povprečju pregledal 7,87 strani.

5.3.2.3 Priprava podatkov

Namen priprave podatkov je dokončno preoblikovanje predhodno očiščenih zahtevkov v obliko, ki jo zahtevata obravnavani problem podatkovnega rudarjenja in izbrani algoritem. V prečiščene podatke sem v tej fazi najprej dodal še zahtevke za oglede dokumentov[90], nato pa sem imenom datotek dodal poizvedbene nize, poenotil poimenovanje ogledanih datotek, ki so bile semantično sicer enake[91], in na osnovi ontologije spletišča vsakemu zahtevku določil vsebinsko kategorijo. Sledilo je preoblikovanje zahtevkov, ki so bili dotlej shranjeni v eni sami podatkovni tabeli, v podatkovno strukturo, ki jo zahteva algoritem MSSC – v t. i. tabelo primerov (ang. case table) in t. i. gnezdeno tabelo (ang. case table) (glej sliko 10):

- **tabela primerov** je podatkovna tabela, ki vsebuje vse razpoložljive podatke o uporabnikih; ker na spletišču SURS-a ni mogoče spremljati podatkov, ki bi omogočili oblikovanje demografskega profila, je ta tabela vsebovala samo unikatni identifikator posameznih uporabnikov;

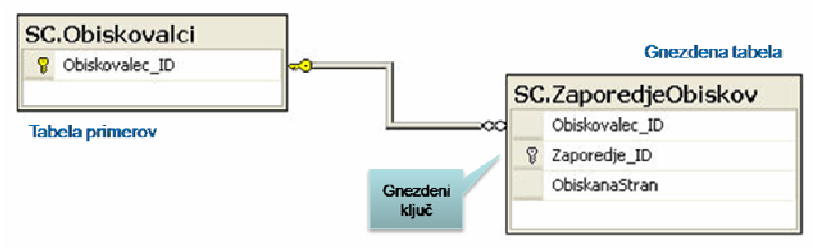
---

[90] Podatke o ogledih odkumentov iz dnevniških datotek (skupaj 227.248) sem najprej očistil neželenih zahtevkov, nato pa sem v prečiščene podatke o zahtevkih za oglede strani dodal tiste dokumente, ki so jih zahtevali uporabniki v končni različici testne množice (16.583).

[91] Tipičen primer je ogled strani z novicami, kjer je vsebina v celoti odvisna od vrednosti poizvedbenega niza. Poleg tega se v novicah statistično enake vsebine objavljajo v rednih časovnih intervalih, zato sem zahtevke za oglede novic poenotil na ravni statističnega koncepta (tipa podatkov).

- **gnezdena tabela** je podatkovna tabela, ki vsebuje zaporedje podatkov o obiskanih straneh; v tabeli so za vsakega uporabnika shranjeni podatki o straneh, ki jih je obiskal, pri čemer ima vsak zapis identifikator (t. i. gnezdeni ključ – ang. nested key), ki določa pozicijo strani v okviru posameznega zaporedja.

Slika 11: Podatkovni model za aplikacijo algoritma Microsoft Sequence Clustering



Vir: prirejeno po Tang, MacLennan, 2005, str. 217

Nekateri avtorji (glej poglavje 3.2.3) predlagajo za učinkovito odkrivanje vzorcev porabe razmejevanje sej v transakcije, vendar pa obiskovalci na spletišču SURS-a v okviru posameznega obiska večinoma ne zasledujejo različnih ciljev (Stražišar, 2005), zato identifikacija transakcij znotraj sej ni smiselna. Namesto tega sem se odločil, da učinkovitost algoritma za priporočanje preverim, če se v analizi uporabljajo zahtevki preslikani v različne semantične ravni. Zanimalo me je, kakšne so razlike v učinkovitosti, če uporabo spremljamo na ravni vsake posamezne strani oz. dokumenta, in kakšne, če jo spremljamo na ravni vsebinskih kategorij, ki sem jih oblikoval na osnovi ontologije spletišča. Priporočanje v primeru, ko algoritem učimo na ravni strani, je načeloma natančnejše, vendar pa je tudi računsko bolj potratno, zato se pojavlja vprašanje, kateri pristop je bolj smiseln in v spletnem okolju tudi realno izvedljiv.

5.3.3 Odkrivanje vzorcev uporabe za poosebljanje

V poglavju 3.3 sem opisal več pristopov, ki jih je mogoče uporabiti za odkrivanje vzorcev uporabniškega navigacijskega vedenja; vsak izmed pristopov je uporaben v določenih okoliščinah in ob določenih osnovnih pogojih glede strojne in programske opreme. Za odkrivanje pravil za priporočanje vsebine sem se odločil uporabiti že omenjeni algoritem MSSC, saj spada v kategorijo algoritmov za odkrivanje zaporednih vzorcev, ki so za tovrstno aplikacijo rudarjenja najprimernejši (glej poglavje 3.3.4). MSSC je hibrid med algoritmi za odkrivanje gruč in algoritmi za odkrivanje zaporednih vzorcev, namenjen pa je za analizo primerov, ki vsebujejo zaporedne podatke, in oblikovanje skupin teh primerov v bolj ali manj homogene segmente na osnovi podobnosti njihovih zaporedij (za podrobnosti glede delovanja algoritma glej poglavje 3.3.4.2).

Z orodjem *SQL Server Business Intelligence Development Studio* (v nadaljevanju tudi BIDS)

sem izdelal dve strukturi za rudarjenje (ang. mining structure): v prvi strukturi (Kategorije) sem algoritem učil na podatkih o ogledih strani na ravni kategorij, v drugi (Strani) pa na podatkih na ravni posameznih strani. Znotraj vsake strukture sem pripravil tri modele, pri katerih sem skušal ugotoviti, kakšen vpliv imajo na natančnost algoritma različne nastavitve glede števila gruč, v katere algoritem porazdeli uporabnike. Zaradi enostavnejše primerjave sem v obeh strukturah definiral model z desetimi gručami (privzeta vrednost algoritma), model, pri katerem algoritem s hevurističnimi pristopi sam poišče optimalno število gruč, in model s petnajstimi gručami[92]. Primer rezultatov je prikazan v prilogi 7.

5.3.3.1 Preverjanje učinkovitosti posameznih modelov

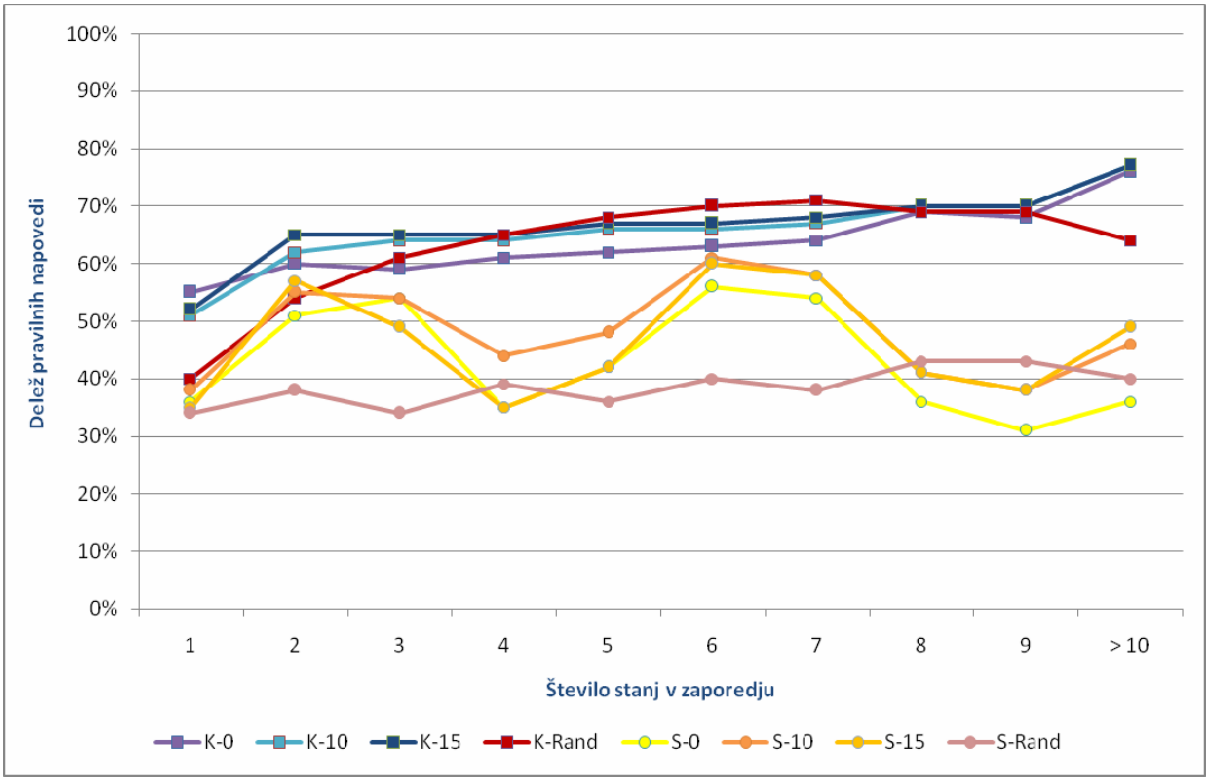
Ključna lastnost algoritma podatkovnega rudarjenja pri njegovi uporabi za priporočanje vsebine je njegova natančnost, tj. njegova zmožnost, da obiskovalcem priporoča vsebine, ki jih dejansko zanimajo. Natančnost algoritma v realnih okoliščinah pa ni odvisna zgolj od njegove interne strukture, ampak tudi od reprezentativnosti učnih podatkov v celotni populaciji, raznolikosti navigacijskega vedenja obiskovalcev spletišča in dolžine trajanja obiska, v okviru katerega izvajamo priporočanje (glej poglavji 3.4 in 4.3.2). Medtem ko ustreznost podatkov za učenje algoritma večinoma ni problematična, saj so podatki na voljo, znanje modelov pa lahko tudi dograjujemo z novimi podatki, na obstoječe navigacijsko vedenje obiskovalcev nimamo vpliva, zato se lahko v določenih primerih zgodi, da ne moremo izdelati modela, ki bi nam dal zadovoljive rezultate.

Primernost algoritma za priporočanje vsebine na spletišču SURS sem želel preveriti empirično, pri čemer so me zanimale tudi razlike v natančnosti priporočanja glede na to, koliko strani je obiskovalec v okviru uporabniške seje že pogledal. Za ta namen sem uporabil testno množico podatkov, ki je zajemala podatke o obisku na spletišču SURS od 24. 2. 2007 do vključno 28. 2. 2007, ki sem jih prečistil in preoblikoval na enak način kot podatke v učni množici; identificiral sem 34.699 zahtevkov in 4.714 uporabniških sej. Poskus je potekal tako, da sem za vsakega obiskovalca ob vsakem njegovem ogledu strani na osnovi pravila zadnjega podniza velikosti *n-1* (glej poglavje 3.3.4.1) izdelal seznam petih priporočenih strani za nadaljevanje obiska (glede na njihovo izračunano verjetnost), nato pa preverjal, ali je bila na tem seznamu tudi stran, ki si jo je obiskovalec nato tudi dejansko ogledal. Poleg tega sem, da bi ugotovil relativno učinkovitost in primernost posameznih modelov v realni situaciji tako na ravni kategorij kot tudi na ravni strani, izvedel tudi simulacijo naključnega priporočanja, pri kateri sem uporabnikom vedno priporočal pet najbolj obiskanih kategorij oz. strani; ti podatki dejansko odražajo delež, ki ga v posamezni fazi obiska predstavlja pet najbolj obiskanih kategorij oz. strani.

---

[92] Model, v katerem je vrednost parametra za želeno število gruč znašala 15, se je ob analizi gruč z vidika njihove raznolikosti in razumljivosti pokazal za najprimernejšega. Dejansko pa vrednost tega parametra pri algoritmu MSSC ne določa tudi končnega števila gruč, saj algoritem v procesu razstavljanja gruč (ang. cluster decomposition) tiste gruče, ki vsebujejo več značilnih zaporedij, še nadalje razdeli.

Slika 12: Primerjava učinkovitosti različnih modelov podatkovnega rudarjenja za priporočanje vsebine[93]



Vir: rezultati rudarjenje po podatki o uporabi spletišča SURS, lastni izračuni

Kot je bilo pričakovano, se je kakovost priporočil na osnovi modela povečevala s podaljševanjem obiska, saj je mogoče na osnovi daljših zaporedij natančneje ugotoviti uporabnikove preference. Prav tako pričakovano so se pri priporočanju strani bolje odrezali modeli priporočanja na osnovi kategorij; število različnih kategorij zahtevkov je bilo namreč manjše od števila različnih strani, zato lahko uporabljeni algoritem ob uporabi kategorij najde bolj čiste skupine (Tang, MacLennan, 2005). Nasprotno primerjava uporabe različnih nastavitev modelov v okviru posameznega pristopa ne pokaže večjih razlik v učinkovitosti; modela s petnajstimi gručami sta sicer najnatančnejša, vendar pa razlike niso bistvene. Zanimivejša je primerjava učinkovitosti modelov glede na učinkovitost naključnega poskusa, s katero lahko enostavno preverimo, ali lahko z uporabo modelov uporabniškega obnašanja dosežemo večjo napovedno natančnost kot z uporabo priporočil na osnovi splošne priljubljenosti strani. Priporočanje na osnovi modela je sicer v povprečju natančnejše, vendar pa je razlika relativno majhna in ni statistično značilna; poleg tega je naključno priporočanje pri zaporedjih od dolžine 4 do vključno dolžine 7 celo najnatančnejše.

[93] Modeli z začetno črko K so modeli, pri katerih se priporočanje izvaja na ravni kategorij, modeli z začetno črko S pa modeli za priporočanje na ravni strani. Številka v imenu modela označuje število gruč; beseda *Rand* označuje naključni model.

5.3.4 Predstavitev priporočil uporabnikom spletišča

Predstavitev priporočil končnim uporabnikom je tisti del predlaganega poosebljanja spletišča SURS, ki se izvaja v stvarnem času in mora poleg merila natančnosti po mojem mnenju zadostiti še naslednjim zahtevam:

- izvajati se mora hkrati s serviranjem spletnih vsebin, ki jih je zahteval uporabnik, in ne sme zmanjšati odzivnih časov;
- za uporabnika ne sme biti moteče in ga ne sme ovirati pri ogledu vsebin;
- način prikazovanja in obdelave priporočil mora biti funkcionalno in oblikovno ločen od samega spletišča.

Na teh zahtevah in zahtevah za integrirano arhitekturo spletnega poosebljanja (glej poglavje 4.5) sem zasnoval na sliki 13 prikazano rešitev za serviranje seznama priporočenih vsebin vsakemu uporabniku in priporočanje vsebin na spletišču SURS.

Slika 13: Prikaz priporočenih strani na spletišču SURS v lebdeči plasti

Vir: lastno delo

Rešitev je zasnovana tako, da se v stvarnem času spremlja pot obiskovalca prek spletišča, na vsaki strani pa se mu nato lahko ponudi nabor priporočenih strani za nadaljevanje obiska; pri

tem je možen nadzor nad tem, koliko strani mora uporabnik najprej obiskati in kolikšna je najmanjša potrebna verjetnost posameznega priporočila, da se mu priporočila dejansko pokažejo. Za prikaz ustreznih priporočil sem uporabil naslednji postopek:

1. Posebna strežniška skripta, ki se jo vključi v vsako stran (podobno kot kodo JavaScript za spremljanje obiska), prek AJAX-tehnike pokliče posebej pripravljeno spletno storitev in ji poda seznam strani, ki jih je uporabnik že obiskal (možno je tudi podajanje parametrov glede uporabljenega modela, števila želenih priporočil in najmanjše zahtevane verjetnosti priporočil). Uporaba AJAX-tehnik omogoča nemoteno prikazovanje in delovanje strani tudi pri nekoliko počasnejšem odzivanju strežnika za podatkovno rudarjenje.

2. Spletna storitev na osnovi prejetih podatkov oblikuje enkratno napovedno poizvedbo DMX (ang. singleton DMX prediction query; za primer poizvedbe glej prilogo 9) in jo prek OLEDB/DM (glej poglavje 3.4.1) posreduje strežniku za podatkovno rudarjenje (Microsoft Analysis Services).

3. Strežnik za podatkovno rudarjenje na osnovi prejete poizvedbe prek spletne storitve vrne ustrezni seznam priporočil, podatke o gruči, ki ji uporabnik pripada, in verjetnosti posameznih priporočil.

4. Strežniška komponenta prikaže priporočila v lebdeči plasti, ki jo uporabnik lahko premika ali skrije in je v celoti oblikovana z jezikom CSS[94].

5.4 PRIMERNOST UPORABE SPLETNEGA POOSEBLJANJA IN RUDARJENJA PO SPLETU NA SPLETIŠČU SURS

V prejšnjih poglavjih sem prikazal, da je uporaba rudarjenja po spletu splošno priznana kot zelo primeren način za analiziranje in izboljševanje kakovosti spletišč, prav tako pa se je pokazalo, da je uporaba rudarjenja po spletu eden izmed najučinkovitejših načinov za izvajanje spletnega poosebljanja. Medtem ko glede uporabe rudarjenja po spletu za analizo vzorcev uporabe ni večjih pomislekov, si pri spletnem poosebljanju strokovnjaki niso tako enotni. V poglavjih 4.1.2 in 4.1.3 predstavljena mnenja o primernosti spletnega poosebljanja jasno kažejo na to, da je uporaba spletnega poosebljanja učinkovita rešitev zgolj v primeru, da omogoča edinstveno dodano vrednost za uporabnike.

Na tej osnovi, na podlagi rezultatov Anket o zadovoljstvu uporabnikov (SURS: Poročilo o

[94] CSS (Cascading Style Sheet, tudi prekrivni slogi) je standardizirani jezik za oblikovanje stilne predloge na spletni strani, v kateri je zapisana oblika spletne strani ali njenih posameznih delov.

Anketi o zadovoljstvu uporabnikov v letu 2003, 2004; Stražišar, 2005) in na podlagi Akcijskega načrta za nadgradnjo spletišča SURS (Kožuh, 2005b) sem preizkusil tudi primernost priporočanja vsebine in v poglavju 5.2 predlagal nekatere druge oblike spletnega poosebljanja za izboljševanje uporabniške izkušnje na spletišču SURS. V nadaljevanju na osnovi rezultatov primera iz poglavja 5.3 podajam analizo primernosti uporabe predlaganih in predvidenih rešitev.

5.4.1 Analiza učinkovitosti priporočanja vsebine z rudarjenjem po podatkih o uporabi spleta

V rešitvi za priporočanje vsebine na spletišču SURS sem se na podlagi že predstavljenih razlogov odločil za uporabo algoritma MSSC. Algoritem se je sprva pokazal za dokaj učinkovitega, saj bi lahko z njim obiskovalcem pravilno priporočili do 80 odstotkov vsebin, če bi se omejili na modeliranje na ravni kategorij, v primeru modeliranja na ravni strani pa bi lahko pravilno napovedali do 60 odstotkov vsebin. Natančnost je bila zelo spodbudna predvsem zaradi velikega številka kategorij oz. strani na spletišču.

V drugačni luči se ugotovljena natančnost pokaže, če jo primerjamo z natančnostjo naključnega priporočanja; v primeru, da bi obiskovalcu vedno priporočili pet najbolj obiskanih strani na spletišču, bi lahko dosegli zelo podobno ali celo večjo natančnost kot v primeru, da bi za priporočanje uporabili model. Ta primerjava kaže, da uporaba algoritma MSSC za poosebljano priporočanje na spletišču SURS v obliki, kot sem jo opisal v poglavju 5.3, ni utemeljena, saj ne daje rezultatov, ki jih ne bi mogli doseči tudi z manj truda in manj sredstvi. Drugi, po mojem mnenju še pomembnejši, dejavnik, ki kaže na neprimernost modela, je dejstvo, da je relativna natančnost modela predvsem posledica pojavljanja najpriljubljenejših strani v priporočenih vsebinah; natančnost priporočanja na straneh, ki niso med najbolj priljubljenimi, je glede na izračune namreč v povprečju za več kot 30 odstotkov nižja od skupne natančnosti.

Razlogov za relativno napovedno neučinkovitost modela je več, med njimi pa se zdijo najpomembnejši naslednji:

- Algoritem MSSC je optimiziran za priporočanje v situacijah, kjer je število različnih stanj manjše od 64 (Tang, MacLennan, 2005); število stanj pri modeliranju navigacijskega vedenja je na ravni kategorij znašalo 122, na ravni strani pa je bilo stanj več kot 2.000, kar predstavlja precejšnje odstopanje od priporočenih vrednosti.

- Algoritem MSSC je sicer namenjen analizi zaporedij, vendar pa za modeliranje uporablja markovske verige prvega reda, kar pomeni, da pri napovedovanju upošteva samo zadnje stanje v zaporedju. Napovedna natančnost je zaradi hkratne uporabe razvrščanja v skupine večja, kot je napovedna natančnost tradicionalnih markovskih modelov prvega reda, vendar pa pri stanjih od 1 do vključno *p-1* ni upoštevano

zaporedje ogledov[95], zato se izgubi pomemben del informacij o vedenju obiskovalcev.

- Algoritem MSSC ne upošteva semantičnih povezav med posameznimi stranmi (glej poglavje 4.4), ampak posamezne strani oz. kategorije priporoča izključno na podlagi njihovega sopojavljanja v uporabniških sejah.

- Na spletišču SURS je delež ogledov petih najbolj priljubljenih strani glede na delež ogledov ostalih strani relativno velik, kar močno poveča natančnost naključnega priporočanja. K tej situaciji najbolj pripomore priljubljenost strani v sklopu *Baze rojstnih imen in priimkov* (kategorija *Imena*), ki so vse od njihove objave najbolj obiskane strani na spletišču.

Neuspešnost opisanega poskusa priporočanja vsebin na spletišču SURS in njegova neprimernost za redno uporabo ne kažeta nujno tudi na neuspešnost drugih predlaganih oblik poosebljanja ali na neustreznost spletnega poosebljanja samega. Situacija, pri kateri izdelani model ni dovolj dober, je pri podatkovnem rudarjenju relativno pogosta in spodbuja k testiranju novih pristopov in rešitev v posameznih fazah procesa, v primeru spletnega poosebljanja pa nas lahko napoti tudi k iskanju alternativnih načinov za izboljševanje kakovosti spletišča in uporabniške izkušnje.

Ideja o uporabi spletnega poosebljanja na spletišču SURS temelji na analizi odzivov uporabnikov, ki se med seboj zelo razlikujejo in odražajo velike razlike v njihovih ciljih, znanju in pristopih ter veliko raznolikost vsebin na spletišču. Alternativna možnost, ki bi jo lahko uporabili za izboljšanje kakovosti spletišča, je načrtno prilagajanje spletišča posameznim značilnim uporabniškim skupinam; podprli bi ga lahko s testi uporabnosti in analizami potreb uporabnikov, ki jih izvajajo v Informacijskem središču SURS-a. S tem pristopom bi lahko načeloma dosegli dobre rezultate, vendar pa je njegova dolgoročno kakovostna izvedba izredno težavna – najprej namreč temelji na pravilni identifikaciji značilnih skupin spletnih uporabnikov[96], nato pa na zmožnosti pripraviti kakovostne vsebine za vsako izmed teh skupin ter ustrezno strukturo tistega dela spletišča, ki je namenjen posamezni skupini. Ob upoštevanju raznolikosti vsebin in dinamike sprememb na spletišču, razpoložljivosti ustreznih kadrov in finančnih sredstev ter izkušenj z neuspelimi podobnimi rešitvami ocenjujem, da ta rešitev v primeru spletišča SURS ni izvedljiva; pri tem težave ne predstavlja zgolj začetna zasnova strukture in vsebine posameznih delov spletišča, ampak predvsem potreba po nenehnem spremljanju potreb vsake uporabniške skupine in

[95] *p* označuje zadnjo obiskano stran v okviru posameznega obiska.

[96] Identifikacija uporabniških skupin mora temeljiti na analizi značilnosti uporabnikov in njihovih ciljev, zato SURS-ova segmentacija uporabnikov (SURS: Navodilo za vodenje nove segmentacije uporabnikov, 2004), ki v osnovi temelji na tem, v kateri organizaciji je uporabnik zaposlen, ni primerna. Tipične skupine uporabnikov lahko najdemo tudi z uporabo tehnike razvrščanja v skupine (glej poglavje 3.3.2).

posledičnem posodabljanju vsebine in strukture.

Nielsen (1998) trdi, da je mogoče večino funkcij spletnega poosebljanja nadomestiti z dobro premišljeno in intuitivno zasnovano navigacijsko strukturo spletišča. Logična navigacija je vsekakor pomemben faktor pri ocenjevanju kakovosti spletišča in tudi na spletišču SURS-a bo potrebno vložiti kar precej truda v zasnovo drugačnega načina krmarjenja po spletišču (Stražišar, 2005), vendar pa mnoge težave uporabnikov, ki na spletišču ne morejo najti iskanih informacij, izhajajo tudi iz njihovega nepoznavanja vsebine; te težave zgolj z boljšo statično navigacijo ni mogoče rešiti.

Zgoraj opisane težave so med glavnimi dejavniki za interes, ki ga za spletno poosebljanje izkazujejo tako lastniki spletišč kot tudi sami obiskovalci (Global outlook for personalization applications, 2001; ChoiceStream, 2006). V primeru neučinkovitosti poosebljenih storitev je torej smiselno raziskati, ali obstajajo alternativne rešitve pri izvajanju posameznih nalog v procesu, ki bi lahko pomagale povečati primernost posamezne rešitve; v primeru, ki sem ga obravnaval v magistrskem delu, bi lahko na učinkovitost poosebljanja vplivali tudi ukrepi, ki jih povzemam v tabeli 4 in se lahko izvajajo posamično ali kombinirano.

Tabela 4: Nabor možnih ukrepov za povečanje učinkovitosti priporočanja vsebine na spletišču SURS

Ukrep	Utemeljitev
Izdelava dodatnih modelov z uporabo drugih vrednosti algoritma MSSC	Natančnost algoritmov je zelo odvisna tudi od parametrov, s katerimi usmerjamo njihovo delovanje – obstaja možnost, da bi s spreminjanjem vrednosti parametrov podpore, števila upoštevanih stanj in števila zaporednih stanj dosegli večjo napovedno natančnost.
Izdelava modelov na podatkih o obiskih, preslikanih v dodatne semantične ravni	Učinkovitost modela sem v okviru magistrskega dela testiral na primeru uporabe kategorij vsebin in posameznih spletnih strani kot osnovne enote pri definiranju spletnega pogleda. Z uporabo modelov, ki bi temeljili na združevanju strani v manj splošne skupine, kot je vsebinska kategorija, bi lahko kakovost modelov potencialno izboljšali.
Uporaba semantičnega znanja v procesu priporočanja vsebine	Pri opisanem primeru poosebljanja se vsebina priporoča zgolj na osnovi tega, katere vsebine se pogosto pojavljajo skupaj; če bi iz priporočil, ki jih najdemo z uporabo modela, izločili tiste, ki trenutni strani na semantični ravni niso sorodne, bi se natančnost priporočanja najverjetneje povečala. Primerno bi bilo tudi vključevanje semantičnega znanja v druge faze procesa.
Uporaba drugih algoritmov podatkovnega rudarjenja	Različni algoritmi so v posameznih primerih različno uspešni. Preizkusili bi lahko druge algoritme za odkrivanje zaporednih vzorcev (npr. algoritme družb SAS ali SPSS), pa tudi druge vrste algoritmov, ki se lahko pojavljajo v sistemih za poosebljanje (asociacijska pravila, razvrščanje v skupine, klasificiranje). Mogoče bi bilo tudi večstopenjsko modeliranje[97] z uporabo več algoritmov.
Razširitev nabora podatkov o uporabnikih	Na spletišču SURS za posameznega uporabnika nimamo drugih podatkov, kot so podatki o njegovem navigacijskem vedenju. Če bi lahko poleg teh podatkov zbirali tudi različne demografske podatke, bi se natančnost modela povečala, saj bi lahko pri priporočanju upoštevali tudi druge dimenzije, ki vplivajo na potrebe in želje uporabnikov (glej poglavje 3.1.2).
Izdelava različnih modelov za posamezne dele spletišča ali posamezne	Natančnost modela zmanjšuje tudi potreba po njegovi uporabi na celotnem spletišču in vseh uporabnikih. Z večstopenjskim modeliranjem bi lahko izdelali modele, ki bi bili zaradi manjšega dosega bolj natančni.

[97] Z izrazom večstopenjsko modeliranje označujem podatkovno rudarjenje, pri katerem v fazi modeliranja zaporedoma uporabimo več algoritmov; primer večstopenjskega modeliranja je uporaba algoritma za klasificiranje, s katerim uporabnike razvrstimo v skupine, ki jih nato posebej analiziramo.

skupine uporabnikov	

Vir: lastno delo

5.4.2 Druge možnosti uporabe spletnega poosebljanja in rudarjenja po spletu

Večkrat sem že omenil, da uporaba rudarjenja po spletu ni omejena zgolj na spletno poosebljanje. Kljub temu da lahko z njim lahko podpremo ali celo omogočimo izvajanje različnih nalog na spletišču, je v organizacijah rudarjenje po spletu še vedno v veliki meri prezrto (Kožuh, 2005b). Poleg dokajšnje zapletenosti tehnologije in neuniformiranosti metodologije lahko med razloge za tako stanje pripišemo tudi dejstvo, da večina avtorjev (Dai, Mobasher, 2004; Mobasher, Jin, Zhou, Mobasher, 2006; Markellou, Rigou, Sirmakessis, 2005; Chen, 2003; Spilipoulou et al., 2003, idr.) kot osnovni cilj rudarjenja po spletu enostavno navaja avtomatično zajemanje in modeliranje vedenjskih vzorcev spletnih uporabnikov. Tako pojmovanje cilja je za odgovorne osebe v organizacijah preozko, saj ne omogoča vpogleda v razloge, zakaj se rudarjenja in poosebljanja sploh lotiti. Cilje je potrebno zastaviti širše in se vprašati, kako lahko z uporabo rudarjenja po spletu in spletnim poosebljanjem podpremo izvajanje strategije, ki si jo je organizacija postavila s tem, ko se je odločila, da bo del svojega poslovanja prestavila na svetovni splet.

V vsakem primeru je jasno, da si morajo organizacije prizadevati, da spletišče uporabnikom omogoča čim enostavnejše izpolnjevanje njihovih nalog. Da bi bilo spletišče uporabnikom v podporo in ne v nadlogo, je potrebno ugotoviti, s kakšnimi težavami se obiskovalci srečujejo, kako jih rešujejo in kakšna so njihova naknadna dejanja; nekaj tovrstnih ciljev in način njihovega doseganja z uporabo rudarjenja in spletnega poosebljanja povzemam v tabeli 5.

Tabela 5: Splošni cilji, ki jih lahko podpremo z rudarjenjem po spletu in spletnim poosebljanjem

Cilj	Uporaba rudarjenja in poosebljanja za dosego cilja
Ugotoviti, ali uporabniki spletišče uporabljajo, tako kot je načrtovano	• Analiza poti prečenja spletišča in smeri, ki jih uporabniki ubirajo iz navigacijskih strani.
Ugotoviti, zakaj uporabniki zapuščajo spletišče, ne da bi izvedli želena dejanja	• Analiza razlik med vzorci obnašanja uporabnikov, ki taka dejanja izvajajo, in tistimi, ki jih ne. • Identifikacija značilnih vzorcev obnašanja, ki vodijo do predčasnega odhoda s spletišča. • Identifikacija skupin in lastnosti uporabnikov, ki ne izkazujejo želenega navigacijskega obnašanja.
Povečati relevantnost predstavljenih vsebin	• Poosebiti vsebino na osnovi uporabnikovega navigacijskega obnašanja in njegovih drugih lastnosti.
Zvišati delež izvajanja želenih dejanj	• Izboljšati navigacijsko strukturo in vsebino spletišča na osnovi identifikacije vzrokov za neizvajanje želenih dejanj. • Poosebiti vsebino in strukturo spletišča ter način iskanja informacij.
Ugotoviti, katere so najbolj cenjene storitve med uporabniki	• Analizirati navigacijske vzorce pri uporabnikih, ki so izvedli želena dejanja.

Cilj	Uporaba rudarjenja in poosebljanja za dosego cilja
Ugotoviti, kakšne so tipične skupine uporabnikov	• Identificirati gruče uporabnikov na osnovi njihovega navigacijskega vedenja in drugih razpoložljivih podatkov.

Vir: lastno delo

Alternativne možnosti uporabe so se pokazale tudi pri rudarjenju po podatkih o uporabi spletišča SURS. Poleg pravil, ki jih lahko uporabimo neposredno za napovedovanje, modeli namreč vsebujejo tudi vrsto drugih informacij, ki bodo SURS-u pomagale pri razumevanju vedenja obiskovalcev (glej tabelo 6) – te informacije so analitiku na voljo v pregledovalnikih orodja BIDS (glej prilogi 7 in 8), naprednejša analiza odkritih vzorcev pa je mogoča tudi z uporabo že predstavljenega jezika DMX; ta poleg napovedovanja omogoča tudi pregledovanje vsebine celotnega modela. Rudarjenje po podatkih o uporabi spleta se je ne nazadnje pokazalo za koristno tudi zato, ker je zahtevalo temeljito in premišljeno pripravo podatkov o obisku na spletišča – prečiščeni podatki so namreč pokazali na nekatere nepravilnosti v načinu analize statistike obiska in bodo služili za večjo natančnost poročil v prihodnje.

Tabela 6: Primeri koristnih informacij, odkritih pri izdelavi modelov za spletno priporočanje

Vir informacij	Opis
Oblikovanje gruč obiskovalcev	Za večino gruč je značilen obisk vsaj ene strani v kategoriji *Imena* – obstaja gruča obiskovalcev, ki obiskuje zgolj te strani, pri skoraj polovici gruč pa so te strani najbolj značilne. Identificiramo lahko tudi zelo homogene gruče obiskovalcev, ki jih zanimajo samo strani v okviru kategorij *Vodič po statistiki* in *Indikatorji*. Medtem ko je vedenje obiskovalcev, ki jih zanimajo zgolj rojstna imena, dokaj logično (velika promocija te storitve po številnih forumih, mlajša populacija itd.), je potrebno raziskati razloge za obstoj drugih dveh omenjenih gruč.
Značilnosti populacije in gruč	Z analizo značilnosti celotne populacije in posameznih gruč ter vsebin, ki jih zanimajo, lahko identificiramo najzanimivejše vsebine in najznačilnejše povezave med njimi. S temi informacijami lahko analiziramo in revidiramo naše predpostavke o tem, katere vsebine se običajno pregledujejo skupaj in kakšen je vrstni red njihovega pregledovanja.
Značilne povezave med stranmi	Analiza povezav med stranmi, ki se oblikujejo na osnovi značilnih poti obiskovalcev, je koristna za preverjanje pravilnosti in koristnosti obstoječe navigacijske strukture. Podrobno je potrebno preučiti zlasti značilne povezave med tistimi stranmi, ki prek navigacije niso neposredno povezane.

Vir: rezultati rudarjenja po podatkih o uporabi spletišča SURS

6 ZAKLJUČEK

Spletno poosebljanje z uporabo rudarjenja po spletu je eden izmed najučinkovitejših načinov za izboljševanje uporabniške izkušnje na spletišču in zbliževanje ciljev organizacije in njenih strank. V magistrskem delu sem prikazal celosten pogled na tovrstno poosebljanje spletišč in

predstavil njegov potencial za izvrševanje in nadgrajevanje strategij, ki jih organizacije izvajajo s pomočjo svetovnega spleta; pri tem sem spletno poosebljanje obravnaval v smislu samodejnega procesa, ki temelji na operativnem znanju v obliki modelov, izdelanih z rudarjenjem po spletu. Ponudil sem ogrodje, ki na eni strani organizacijam omogoča seznanjanje z možnostmi, ki jih ponujata rudarjenje po spletu in spletno poosebljanje, na drugi strani pa v luči opisanih težav, rezultatov in odprtih vprašanj tudi ocenjevanje uporabnosti pristopa v vsakem posameznem primeru.

Poosebljanje vsebine in strukture v smislu prilagajanja zahtevam in ciljem uporabnikov postaja ena izmed najpomembnejših zahtev pri razvoju spletnih rešitev. Potreba po poosebljanju se poraja iz različnih vidikov interakcije med uporabniki in spletišči: (1) uporabniki so zaradi različnih interesov in ciljev, globalne razpršenosti informacij in storitev ter drugih dejavnikov vse bolj raznoliki, (2) spletne rešitve morajo biti prirejene za različne tipe odjemalcev, ki se ne razlikujejo samo na nivoju programske opreme, ampak tudi po različnih ergonomijah vmesnikov, načinu povezovanja v internet ipd., (3) spletišča postajajo zaradi hitrega večanja razpoložljivih storitev vedno bolj zapletena in begajoča za običajnega uporabnika. Na drugi stani je poosebljanje pomembno, ker uporabniki velikokrat informacije iščejo zato, ker želijo na ta način razrešiti svojo težavo ali doseči zastavljeni cilj, a jim njihovo obstoječe znanje v danem trenutku ne zadošča. Posledično to pomeni, da pogosto ne vedo natančno, katere vsebine ali storitve bi bile za njih koristne, in tako niti ne morejo podrobno opisati glavnih značilnosti potencialno uporabnih informacijskih objektov. Poleg tega uporabniki ponavadi niso dovolj dobro seznanjeni z načinom delovanja in strukturo spletišča niti z naborom izrazov, ki jih spletna rešitev uporablja za dostop do posameznih informacij. Zaradi navedenega je zaželjeno, da je spletišče (v določeni meri) zmožno uporabniku predlagati različna dejanja ali vsebine, ki jim kratkoročno pomagajo rešiti zastavljene naloge, dolgoročno pa jim pomagajo boljše razumeti spletišče in njegovo vsebino ter omogočajo večjo učinkovitost storitev, ki jih spletišče ponuja.

Kljub obstoju različnih sistemov za poosebljanje je omogočanje tovrstnih storitev na spletišču vse prej kot enostavno. Spletno poosebljanje ni programska rešitev, ki bi jo lahko kupili in jo preprosto namestili na spletni strežnik ter nato čakali, da bo čudežno povečala zadovoljstvo naših uporabnikov; spletno poosebljanje je proces, v okviru katerega mora organizacija temeljito analizirati svoje storitve, svoje uporabnike in dodano vrednost, ki jim jo ponuja, nato pa z analizo medsebojnih odnosov identificirati težave in najti načine za njihovo razrešitev. Uporaba metod rudarjenja po podatkih o uporabi spleta in drugih vrst rudarjenja po spletu v procesu spletnega poosebljanja postaja vse bolj nujna in potrebna, saj je podatkov preveč, da bi jih lahko učinkovito analizirali s tradicionalnimi pristopi in metodami ter da bi v njih sami našli (skrite) vzorce in pravila.

Izvajanje posameznih nalog v procesu rudarjenja po podatkih ni preprosto in od vseh vpletenih zahteva temeljito poznavanje rudarjenja po podatkih, spletnih tehnologij in samega

vsebinskega področja; težave še povečujejo mladost tehnologije rudarjenja po spletu, relativna nepreizkušenost in nedovršenost algoritmov za odkrivanje vzorcev ter pomanjkljivosti trenutnih spletnih protokolov, ki otežujejo kakovostno zbiranje in pripravo podatkov. Uporaba rudarjenja po spletu za poosebljanje velja za eno najperspektivnejših aplikacij rudarjenja po podatkih, vendar pa je lahko naraščajoča priljubljenost tega pristopa tudi past, saj ta oblika izboljševanja kakovosti in uporabnosti spletišča tako z vidika vloženega truda za razvoj rešitev kot tudi z vidika dodane vrednosti za uporabnike ni vedno primerna.

Samo tehnologijo in primernost pristopa sem preizkusil na primeru spletišča Statističnega urada RS. Izkazalo se je, da natančnost algoritma MSSC, ki sem ga uporabil za modeliranje navigacijskega vedenja obiskovalcev, ni tako visoka, da bi upravičila njegovo uporabo. Spletnega priporočanja v predstavljeni obliki tako še ni mogoče kakovostno uporabiti na spletišču, obstaja pa več pristopov, s katerimi bi lahko natančnost modelov izboljšali ali pa bi rezultate uporabili pri drugih nalogah izboljševanja kakovosti spletišča. V celoti lahko tako ugotovimo, da potencial spletnega poosebljanja z rudarjenjem po spletu na spletišču SURS ostaja in bi ga z rešitvijo nekaterih tehničnih težav koristno uporabili za izboljševanje uporabniške izkušnje.

V praksi se je pokazalo, da sta tako spletno poosebljanje kot tudi rudarjenje po spletu sami po sebi še vedno vse prej kot zreli tehnologiji. Posledično se pri njuni uporabi pojavljajo številne tehnične težave in odprta vprašanja. V fazi zbiranja in priprave podatkov je potrebno najti celostne rešitve in modele za objektiven zajem podatkov, v fazi modeliranja pa je potrebno najti pristope in algoritme, ki so učinkoviti za zajem vedenjskih vzorcev uporabnikov, hkrati pa tudi uporabni v stvarnem času ter dosegljivi za širšo razvijalsko skupnost. Dodatna težava je povezana z vprašanjem zasebnosti uporabnikov, ki pa ga v magistrskem delu nisem podrobneje obravnaval; spletno poosebljanje mora biti transparentno, pri čemer moramo uporabniku omogočiti nadzor nad zbiranjem podatkov in izvajanjem poosebljanja, predvsem pa jim je potrebno predstaviti način uporabe zbranih podatkov in pojasniti koristi; te jim mora sistem za poosebljanje prinesti, če želimo, da je njegova uvedba smiselna.

Iz vsega napisanega lahko sklepamo, da je odločitev za uvedbo spletnega poosebljanja na osnovi rudarjenja po spletu pravilna, če so izpolnjeni ustrezni pogoji, nista pa rudarjenje in poosebljanje čarobni paličici, s pomočjo katerih bi lahko podjetje enostavno povečalo svoje prihodke. Če upoštevamo, da naj bi 79 odstotkov uporabnikov zanimala poosebljena vsebina in da naj bi se bila več kot polovica uporabnikov pripravljena odreči delu zasebnosti v zameno za poosebljeno izkušnjo (ChoiceStream, 2006), lahko trdimo, da je opisani pristop prava rešitev, če vemo, kaj želimo z njim doseči, če je osnovan na želji, da bi uporabnikom pomagali pri zadovoljevanju njihovih potreb, in ga uvedemo zgolj v primeru, da uporabnikom preverjeno prinaša dodano vrednost pri njihovem obisku spletišča.

7 LITERATURA IN VIRI

7.1 LITERATURA

[1] Agrawal Rakesh, Imielinski Tomasz, Swami Arun N.: Mining Association Rules between Sets of Items in Large Databases. Buneman Peter, Jajodia Sushil, eds.: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, 1993, str. 207-216.

[2] Agrawal Rakesh, Srikant Ramakrishnan: Fast Algorithms for Mining Association Rules. Proceedings of the 20th International Conference on Very Large Databases, Santiago, 1994, str. 487-499.

[3] Agrawal Rakesh, Srikant Ramakrishnan: Mining sequential patterns: Generalizations and performance improvements. Proceedings of the Eleventh International Conference on Data Engineering, Washington, 1995, str. 3-24.

[4] Almeida Virgilio et al.: Analyzing robot behavior in e-business sites. ACM SIGMETRICS Performance Evaluation Review, ACM Press, 29(2001), 1, str. 338-339.

[5] Alpar Paul: Satisfaction with a web site: Its measurements, factors, and correlates. Scheer Wilhelm, Nüttgens Markus, eds., Electronic Business Engineering, Heidelberg : Springer-Verlag, 1999, str. 271–287.

[6] Anand Sarabjot Singh, Mobasher Bamshad: Intelligent Techniques for Web Personalization. Anand Sarabjot Singh, Mobasher Bamshad, eds., Intelligent Techniques for Web Personalization, 2005, str. 1–36.

[7] Anderson Corin R., Domingos Pedro, Weld Daniel S.: Relational markov models and their application to adaptive web navigation. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, 2002, str. 143–152.

[8] Baglioni Miriam et al.: Preprocessing and Mining Web Log Data for Web Personalization. Proceedings of 8th National Conference of the Italian Association for Artificial Intelligence, 2003.

[9] Berendt Bettina et al.: Measuring the accuracy of sessionizers for web usage analysis. Workshop on Web Mining at the First SIAM International Conference on Data Mining, Chicago, USA, 2001, str. 7-14.

[10] Berendt Bettina et al.: The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. Proceedings of the WebKDD 2002 Workshop, Edmonton, 2002, str. 159-179.

[11] Berendt Bettina, Hotho Andreas, Stumme Gerd: Towards Semantic Web Mining. First International Semantic Web Conference on the Semantic Web, Sardinia, 2002, str. 264–278.

[12] Berendt Bettina, Mobasher Bamshad, Spilipoulou Myra: Web usage mining for e-business applications. ECML/PKDD-2002 Tutorial, Helsinki, 2002.

[13] Berendt Bettina, Spiliopoulou Myra: Analysis of Navigation Behaviour in Web Sites Integrating Multiple Information Systems. VLDB Journal, 9(2000), 1, str. 56-75.

[14] Berendt Bettina: Using Site Semantics to Analyze, Visualize, and Support Navigation. Data Mining and Knowledge Discovery, 6(2002), 1, str. 37–59.

[15] Berry Michael J.A., Linoff Gordon S.: Mastering Data Mining. New York : J. Wiley, 2000. 494 str.

[16] Berry Michael J.A., Linoff Gordon: Data Mining Techniques. New York : J. Wiley, 1997. 454 str.

[17] Berthon Pierre, Pitt, Leyland F., Watson Richart T.: The World Wide Web as an advertising medium: Toward an understanding of conversion efficiency. Journal of Advertising Research, 36(1996), 1, str. 43-55

[18] Blom Jan: Personalization – A Taxonomy. Conference on Human Factors in Computing Systems, The Hague, 2000, str. 313–314.

[19] Boiko Bob: Content Management Bible. Indianapolis : Wiley Publishing, 2005. 1122 str.

[20] Byrne et al.: The Tangled Web We Wove: A Taskonomy of WWW Use. Proceedings of the International Conference on Human Factors in Computing Systems, Pittsburgh, 1999, str. 544-551.

[21] Cabena Peter et al.: Discovering data mining: from concept to implementation. Upper Saddle River : Prentice Hall PTR, 1997. 195 str.

[22] Cadez Igor et al.: Model-Based Clustering and Visualization of Navigation Patterns on a Web Site. Data Mining and Knowledge Discovery, 7(2003), 4, str. 399–424.

[23] Catledge Lara, Pitkow James: Characterizing Browsing Strategies in the World-Wide Web. Proceedings of the Third International World-Wide Web conference on Technology, tools and applications table of contents, New York : Elsevier North-Holland. 1995, str. 1065-1073.

[24] Chan Philip K.: Constructing Web User Profiles: A Non-invasive Learning Approach. Masand Brij, Spiliopoulou Myra, eds., Web Usage Analysis and User Profiling. Berlin : Springer-Verlag, 2000, str. 39–55.

[25] Chapman Pete et al.: CRISP-DM 1.0 – Step-by-step data mining guide. B.k. : 1999, 78 str.

[26] Chen Ming-Syan, Park Jong Soo, Yu Philip S.: Data Mining for Path Traversal Patterns in a Web. Sixteenth International Conference on Distributed Computing Systems, Hong Kong, 1996, str. 385-393.

[27] Chen Zhixiang, Fu Ada Wai-Chee, Tong Frank Chi-Hung: Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs. Internet and Web

Information Systems, 6(2003), 3, str. 259–279.

[28] Chi Ed H.: Improving web usability through visualization. IEEE Internet Computing, 6(2002), 2, str. 64–71.

[29] ChoiceStream: 2006 ChoiceStream Personalization Survey. ChoiceStream, [URL: http://www.choicestream.com/pdf/ChoiceStream_PersonalizationSurveyResults2006.pdf], 14. 8. 2006.

[30] Cooley Robert, Mobasher Bamshad, Srivastava Jaideep: Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1(1999), 1, str. 5-32.

[31] Craven Mark: Learning to Construct Knowledge Bases from the World Wide Web. Artificial Intelligence, 118(2000), 1-2, str. 69-113.

[32] Cunha Carlos, Bestavros Azer, Crovella Mark: Characteristics of WWW Client-based Traces. Boston University, Department of Computer Science. [URL: http://www.cs.bu.edu/techreports/pdf/1995-010-www-client-traces.pdf], 1995.

[33] Dai Honghua, Mobasher Bamshad: Integrating Semantic Knowledge with Web Usage Mining for Personalization. Scime Anthony, ed., Web Mining: Applications and Techniques, Hershey : Idea Group Publishing, 2005, str. 276-306.

[34] Doctrow Cory: Metacrap: Putting the torch to seven straw-men of meta-utopia. [URL: http://www.well.com/~doctrow/metacrap.htm], 26.8.2001.

[35] Eirinaki Magdalini, Vazirgiannis Michalis: Web Mining for Web Personalization. ACM Transactions on Internet Technology, 3(2003), 1, str. 1-27.

[36] Fayyad Usama, Piatetsky-Shapiro Gregory, Smyth Padhraic: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 3(1996), str. 37-54.

[37] Feldman Robert: Text Mining. Klösgen Willi, Żytkow Jan M., eds., Handbook of Data Mining and Knowledge Discovery. Oxford : Oxford University Press, 2002, str. 749–757.

[38] Fensel Dieter: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag : Berlin, 2004, 162. str.

[39] Fleming Jennifer: Web navigation: Designing the user experience. O'Reilly Media, Inc. : Sebastopol, 1998, 264 str.

[40] Fogg B.J.: Persuasive Technology: Using Computers to Change What We Think and Do. San Francisco : Morgan Kaufmann Publishers, 2003, 312 str.

[41] Friedman Jerome H.: Data Mining and Statistics: What's the Connection? Stanford University. [URL: http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps], november 1997.

[42] Galeas Patricio: Web Mining. [URL: http://www.galeas.de/webmining.html], september 2005.

[43] Getoor Lise: Learning probabilistic models of relational structure. Proceedings of the 18th International Conference on Machine Learning, Williamstown, 2001, str. 1300-

1309.

[44] Grinstead Charles M., Snell Laurie J.: Introduction to Probability. American Mathematical Society : Rhode Island, 1997. 510 str.

[45] Grossenbacher Armin: The shift from print to electronic publishing. International Marketing and Output Databases Conference, The Hague, 2005. 23 str.

[46] Han Jiawei, Kamber Micheline: Data Mining: Concepts and Techniques. San Francisco : Morgan Kaufmann Publishers, 2000. 500 str.

[47] Hand David, Mannila Heikki, Smyth Padhraic: Principles of data mining. Cambridge : MIT Press, 2001. 546 str.

[48] Herder Eelco, Weinreich Harald: Interactive web usage mining with the navigation visualizer. Conference on Human Factors in Computing Systems, Oregon, 2005, str. 1451–1454.

[49] Huang Xiangji: Dynamic Web Log Session Identification With Statistical Language Models. Journal of the American Society for Information Science and Technology, 55(2004), 14, str. 1290-1303.

[50] Jin Xin, Zhou Yanzan, Mobasher Bamshad: Web Usage Mining Based on Probabilistic Latent Semantic Analysis. Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'04), Seattle, 2004. [URL: http://maya.cs.depaul.edu/~mobasher/papers/DM04-WM-Book.pdf], 2004.

[51] Johnson Grant A.: Personalization to Boost Response Rates. Direct Magazine. [URL: http://directmag.com/disciplines/creative/marketing_article_8/], 15.10.2006.

[52] Joinson Adam: Who's watching you? Power, personalization and on-line compliance. Institute of Educational Technology, The Open University. [URL: http://iet.open.ac.uk/pp/a.n.joinson/Joinson_CSI_2005.ppt], 4.2.2007.

[53] Jorge Alipio et al.: Web Site Access Analysis for a National Statistical Agency. Mladenić Dunja et al., eds.: Data Mining and Decision Support - Integration and Collaboration. Boston : Klower Academic Publishers, 2003, str. 167-176.

[54] Jupitermedia: Jupiter Research reports that web site "personalization" does not always provide positive results. Jupitermedia Corporation. [URL: www.jupitermedia.com /corporate/releases/03.10.14-newjupresearch.html], 14.10.2003.

[55] Kardaun Jan W.P.F., Alanko Timo: Exploratory Data Analysis and Data Mining in the setting of National Statistical Institutes. New Techniques and Technologies for Statistics, Sorrento. [URL: http://europa.eu.int/en/comm/eurostat/research/conferences /ntts-98/papers/cp/042c.pdf], 1998.

[56] Klösgen Willi, Lauer Stephan R. W.: Visualization and Data Mining Results. Klösgen Willi, Żytkow Jan M., eds., Handbook of Data Mining and Knowledge Discovery. Oxford : Oxford University Press, 2002, str. 509–515.

[57] Klösgen Willi, Żytkow Jan M.: Knowledge Discovery in Databases: The Purpose,

Necessity, and Challenges. Klösgen Willi, Żytkow Jan M., eds., Handbook of Data Mining and Knowledge Discovery. Oxford : Oxford University Press, 2002, str. 1-9.

[58] Klösgen Willi, Żytkow Jan M.: Multidisciplinary Contributions to Knowledge Discovery. Klösgen Willi, Żytkow Jan M., eds., Handbook of Data Mining and Knowledge Discovery. Oxford : Oxford University Press, 2002b, str. 22-32.

[59] Klösgen Willi, Żytkow Jan M.: The Knowledge Discovery Process. Klösgen Willi, Żytkow Jan M., eds., Handbook of Data Mining and Knowledge Discovery. Oxford : Oxford University Press, 2002a, str. 10-21.

[60] Kobsa Alfred, Koeneman Jürgen, Pohl Wolfgang: Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. The Knowledge Engineering Review 16(2001), 2, str. 111–155.

[61] Kolar Charles P., Leavitt John R., Mauldin Michael: Robot exclusion standard revisited. [URL: http://www.kollar.com/robots.html], 2.6.1996.

[62] Kosala Raymond, Blockeel Hendrik; Web Mining Research: A Survey. ACM SIGKDD Explorations Newsletter, 2(2000), 1, str. 1–15.

[63] Koster Martijn: The Web Robots Pages. [URL: http://www.robotstxt.org]. 1994.

[64] Kožuh Boštjan: Enhancing the experience of web users. International Marketing and Output Databases Conference, Avila, 2006. 15 str.

[65] Kožuh Boštjan: Kako do statističnih podatkov in informacij. 10. strokovno posvetovanje specialnih knjižnic in 3. strokovno posvetovanje visokošolskih knjižnic z mednarodno udeležbo, Ljubljana, 2004. 3 str.

[66] Kožuh Boštjan: Web usage mining for effective personalization and adaptation of statistical websites. International Marketing and Output Databases Conference, The Hague, 2005a. 10 str.

[67] Kramer Joseph, Noronha Sunil, Vergo John: A user-cetered design approach to personalization. Communications of the ACM, 43(2000), 8, str. 45–48.

[68] Leskovec Jure: Web Projections: Learning from Contextual Subgraphs of the Web. 16th International World Wide Web Conference, Banff, 2007.

[69] Lin Weiyang, Alvarez Sergio A., Ruiz Carolina: Efficient adaptive-support association rule mining for recommender systems. Data Mining and Knowledge Discovery, 6(2002), 1, str. 83-105.

[70] Linden Greg, Smith Brent, York Jeremy: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(2003), 1, str. 76-80.

[71] Manber Udi, Patel Ash, Robinson John: Experience with personalization on Yahoo!, Communications of the ACM, 43(2000), 8, str. 35–39.

[72] Markellou Penelope, Rigou Maria, Sirmakessis Spiros: Mining for Web Personalization. Scime Anthony, ed., Web Mining: Applications and Techniques, Hershey : Idea Group Publishing, 2005, str. 27-48.

[73] Meteren Robin, Someren Maarten: Using Content-Based Filtering for Recommendation. [URL: http://www.ics.forth.gr/~potamias/mlnia/paper_6.pdf], 3.2.2007.

[74] Mladenić Dunja, Grobelnik Marko: Text and Web Mining. Mladenić Dunja et al., eds.: Data Mining and Decision Support - Integration and Collaboration. Boston : Kluwer Academic Publishers, 2003, str. 15-22.

[75] Mobasher Bamshad, Cooley Robert, Srivastava Jaideep: Automatic personalization based on Web usage mining. Communications of the ACM, 43(2000), 8, str. 142-151.

[76] Mobasher Bamshad: Data Mining for Web Personalization. Brusilovsky Peter, Kobsa Alfren, Nejdl Wolfgang, eds., The Adaptive Web: Methods and Strategies of Web Personalization. Heidelberg : Springer, 2007.

[77] Mobasher Bamshad: Web Usage Mining and Personalization. Munindar P. Singh, ed., Practical Handbook of Internet Computing, Boca Raton : CRC Press, 2005.

[78] Mobasher et al.: Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery, 6(2002), 1, str. 61–82.

[79] Mulvenna Maurice D. et al., Personalization on the Net using Web Mining: introduction. Communications of the ACM, 43(2000), 8, str. 122–125.

[80] Murray Dan, Durrell Kevan: Inferring Demographic Attributes of Anonymous Internet Users. Masand Brij, Spiliopoulou Myra, eds., Web Usage Analysis and User Profiling. Berlin : Springer, 2000, str. 7–20.

[81] Nakagawa Miki, Mobasher Bamshad: A Hybrid Web Personalization Model Based on Site Connectivity. Proceedings of the WebKDD Workshop at the ACM SIGKKDD International Conference on Knowledge Discovery and Data Mining, Washington, August 2003.

[82] Nielsen Jacob: Personalization is Over-Rated. Jakob Nielsen's Alertbox. [URL: http://www.useit.com/alertbox/981004.html], 4.10.1998.

[83] Nielsen Jacob: Usability 101: Introduction to Usability. Jacob Nielsen's Alertbox. [URL: http://www.useit.com/alertbox/20030825.html], 25.8.2003.

[84] Paliouras Georgios et al.: Clustering the Users of Large Web Sites into Communities. Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco : Morgan Kaufmann Publishers Inc., 2000, str. 719–726.

[85] Perkowitz Mike, Etzioni Oren: Towards adaptive Web sites: Conceptual framework and case study. [URL: http://www8.org/w8-papers/2b-customizing/towards/towards.html], 2000.

[86] Peterle Polona: Vključitev teorije markovskih verig v model planiranja materialnih potreb, magistrsko delo. Ekonomska fakulteta : Ljubljana, 2002. 84 str.

[87] Pierrakos Dimitrios et al.: Koinotites: A Web Usage Mining Tool for Personalization. Avouris Nikolaos, Fakotakis Nikos, eds., Proceedings of 1st Panhellenic Conference with International participation, Patras, 2001, str. 231-236.

[88] Pierrakos Dimitrios et al.: Web Usage Mining as a Tool for Personalization: A Survey. User Modeling and User-Adapted Interaction, 13(2003), 4, str. 311-372.

[89] Pirolli Peter, Card Stuart: Information foraging in information access environments. Conference on Human Factors in Computing Systems, Denver, 1995, str. 51–58.

[90] Pirolli Peter, Pitkow James, Rao Ramana: Silk from a sow's ear: extracting usable structures from the Web. Proceedings of the SIGCHI conference on Human factors in computing systems, Vancouver, 1996, str. 118-125.

[91] Pitkow James E., Bharat Krishna A.: WebViz: A Tool for WWW Access Log Analysis. First International WWW Conference, Geneva, 1994, str. 271–277.

[92] Pitkow James, Pirolli Peter: Mining Longest Repeating Subsequences To Predict World Wide Web Surfing. Proceedings of the 1999 USENIX Annual Technical Conference, Monterey, 1999.

[93] Pitkow James: In search of reliable usage data on the WWW. Selected papers from the sixth international conference on World Wide Web. Essex : Elsevier Science Publishers, 1997, str. 1343–1355.

[94] Sarner Adam: GSI Commerce Uses CRM to Drive Online Sales. Garner Research, 19.3.2004.

[95] Sarukkai Ramesh R.: Link Prediction and Path Analysis Using Markov Chains. [URL: http://www9.org/w9cdrom/68/68.html], maj 2000.

[96] Sen Rituparna, Hansen Mark H.: Predicting Web User's Next Access Based on Log Data. [URL: http://www.stat.ucla.edu/~cocteau/papers/pdf/asa.pdf], 2000.

[97] Smyth Barry, Cotter Paul: MP^3 – Mobile Portals, Profiles and Personalization. Levene Mark, Poulovassilis, eds., Web Dynamics: Adpoting to Change in Content, Size, Topology and Use. Berlin : Springer, 2004, str. 411-433.

[98] Spiliopoulou Myra, Carsten Pohle: Data Mining for Measuring and Improving the Success of Web Sites, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, 5(2001), str. 85–114.

[99] Spilipoulou Myra et al.: A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. INFORMS Journal on Computing, 15(2003), 2, str. 171-190.

[100] Srikant Ramakrishnan, Yang Yinghui: Mining web log to improve website organization. 10th International World Wide Web Conference, Hong Kong, 2001, str. 430–437.

[101] Srivastava Jaideep et al.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 1(2000), 2, str. 12-23.

[102] Stražišar Nina: Predstavitev rezultatov analize Ankete o mnenju uporabnikov o spletni strani Statističnega urada RS v letu 2005. Statistični urad Republike Slovenije. [URL: http://www.stat.si/doc/drzstat/ank/rezultati_2005.pdf], 2005, 56 str.

[103] Su Zhong et al.: WhatNext: A Prediction System for Web Requests using N-gram Sequence Models. Proceedings of the First International Conference on Web Information System and Engineering Conference. Washington : IEEE Computer Society, 2000, str. 200–207.

[104] Sullivan Terry: Reading Reader Reaction: A Proposal for Inferential Analysis of Web Server Log Files. [http://www.pantos.org/ts/papers/rrr.html], 1. 6. 2007.

[105] Sundgren Bo et al.: Using Text Mining in Official Statistics. Knowledge Mining - Proceedings of the NEMIS Final Conference, Athens, 2004.

[106] Tam Siu-Ming: On-line communication of statistics: Back to basics. International Marketing and Output Databases Conference, The Hague, 2005.

[107] Tan Pang-Ning, Kumar Vipin: Discovery of web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery, 6(2002), 1, str. 9-35.

[108] Tang ZhaoHui, MacLennan Jamie: Data Mining with SQL Server 2005. Indianapolis : Wiley Publishing, 2005. 460 str.

[109] Valeyathan Ganesan, Yamada Seiji: Behaviour Based Web Page Evaluation. 16th International World Wide Web Conference, Banff, 2007.

[110] White Ryen W. et al.: Supporting exploratory search. Communications of the ACM, 49(2006), 4, str. 37–74.

[111] Woon Yew-Kwong, Ng Wee-Keong, Lim Ee-Peng: Web Usage Mining: Algorithms and Results. Scime Anthony, ed., Web Mining: Applications and Techniques, Hershey : Idea Group Publishing, 2005, str. 373-391.

[112] Wu Yi-Hung, Chen Arbee L. P.: Prediction of Web Page Accesses by Proxy Server Log. World Wide Web: Internet and Web Information Systems, 5(2002), 1, 67–88.

[113] Xing Dongshan, Shen Junyi: Efficient data mining for web navigation patterns. Information and Software Technology, 46(2004), str. 55–63.

[114] Yang Qiang, Li Tianyi, Wang Ke: Building Association-Rule Based Sequential Classifiers for Web-Document Prediction. Data Mining and Knowledge Discovery, 8(2004), 3, str. 253-273.

[115] Yang Qiang, Zhang Henry Hanning: Integrating Web Prefetching and Caching Using Prediction Models. World Wide Web, 4(2001), 4, str. 299–321.

[116] Zukerman I., Albrecht D. W., Nicholson A. E.: Predicting users' requests on the WWW. Proceedings of the seventh international conference on User modeling. New York : Springer-Verlag, 1999, str. 275–284.

7.2 VIRI

[1] AJAX : The Official Microsoft AP.NET AJAX Site. Microsoft. [URL: http://ajax.asp.net], 4.5.2006.

[2] Bayes' theorem. Wikipedia. [URL: http://en.wikipedia.org/wiki/Bayes%27_theorem], 13.1.2007.

[3] Browser Statistics. Refsnes Data. [URL: http://www.w3schools.com/browsers/browsers_stats.asp], 13.4.2007.

[4] Data Mining Group. [URL: http://www.dmg.org], 27.1.2007.

[5] Data Mining Methodology. KDNuggets. [URL: http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm], april 2004.

[6] Direktiva o zasebnosti in elektronskih komunikacijah. Uradni list Evropske skupnosti, [URL: http://europa.eu.int/eurlex/lex/LexUriServ/LexUriServ.do?uri=CELEX: 32002L0058: SL:HTML],

[7] Europe's Information Society Thematic Portal – Privacy Protection. European Comission. [URL: http://europa.eu.int/information_society/policy/ecomm/todays_ framework/privacy_protection/index_en.htm], 21.2.2007.

[8] Foresight 2020: Economic, industry and corporate trends. The Economist Intelligence Unit, 2006.

[9] Hypertext Transfer Protocol -- HTTP/1.1, RFC 2616, Internet Society. [http://www.ietf.org/rfc/rfc2616.txt], 1999.

[10] Identification des robots. [URL: http://www.actulab.com/identification-des-robots.php], 26.1.2007.

[11] Identification Protocol, RFC 1413. [http://www.ietf.org/rfc/rfc1413.txt], 1993.

[12] Inaccessibility of Visually-Oriented Anti-Robot Tests. World Wide Web Consortium. [URL: http://www.w3.org/TR/2003/WD-turingtest-20031105], 5.11.2003.

[13] Iprom: Vse več uporabnikov slovenskih spletnih medijev briše ali blokira piškotke. [URL: http://www.centraliprom.com/index.shtml?press;75], 10.2.2006.

[14] ISlovar. Slovensko društvo informatika. [URL: http://www.islovar.org].

[15] ISO/IEC 13249-6. International Organization for Standardization, 2006.

[16] JSR-73: Data Mining API. The Java Community Process Program. [URL: http://jcp.org/en/jsr/detail?id=73], 27.1.2007.

[17] Keith Gary Joel: Browser Capabilities Project. [URL: http://browsers.garykeith.com], 13.12.2006.

[18] Kožuh Boštjan: Simple cookie checking with ASP. CodingForums.com [URL: http://www.codingforums.com/showthread.php?t=104205], 3.1.2007.

[19] Mixture model. Wikipedia. [URL: http://en.wikipedia.org/wiki/Mixture_model], 18.1.2007.

[20] Netgenesis. SPSS, Inc. [URL: http://www.spss.com/netgenesis], 26.2.2007.

[21] OWL Web Ontology Language Overview. World Wide Web Consortium, [URL: http://www.w3.org/TR/owl-features/], 10.2.2004.

[22] Pahor David et al.: Leksikon računalništva in informatike. Ljubljana : Pasadena, 2002. 800 str.

[23] Paul Seth et al.: Microsoft SQL Server Data Mining Tutorial. Microsoft Corporation. 2005.

[24] Platform for Privacy Preferences (P3P) Project. World Wide Web Consortium. [URL: http://www.w3.org/P3P], 12.10.2006.

[25] Resource Description Framework. Wikipedia. [URL: http://en.wikipedia.org/wiki/Resource_Description_Framework], 18.2.2007.

[26] Sarka Dejan: Data mining with SQL Server 2005 – seminarsko gradivo. Solid Quality Learning, 2006.

[27] SAS: Getting Started with SAS Enterprise Miner 4.3, Cary : SAS Publishing, 2006.

[28] Statistični urad Republike Slovenije. [URL: http://www.stat.si/stat_urad.asp], 15.3.2007.

[29] Statistični urad RS: Kakovost statistike. [URL: http://www.stat.si/kakovost], 19.3.2007.

[30] Statistični urad RS: Politika diseminacije statističnih podatkov. [URL: http://www.stat.si/drz_stat_diseminacija.asp], 19.3.2007.

[31] Statistični urad RS: Navodilo za vodenje nove segmentacije uporabnikov. [URL: http://www.stat.si/doc/drzstat/ank/segmentacija_2003.pdf], 23.3.2004.

[32] Statistični urad RS: Poročilo o Anketi o zadovoljstvu uporabnikov v letu 2003. [URL: http://www.stat.si/doc/drzstat/ank/rezultati_2003.pdf], 31.5.2004.

[33] Surf Maps Visualising Web Browsing, An Atlas of Cyberspaces. [URL: http://www.cybergeography.org/atlas/surf.html], 27.1.2007.

[34] SurfStats – Web Statistics Software. Surfstats. [URL: http://www.surfstats.com].

[35] The Web Robots Database. [URL: http://www.robotstxt.org/wc/active.html], 26.1.2007.

[36] W3C Extended Log File Format (IIS 6.0). Microsoft Corporation. [URL: http://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/bea506fd-38bc-4850-a4fb-e3a0379d321f.mspx], 22. 3. 2007.

[37] W3C Semantic Web Activity. World Wide Web Consortium, [URL: http://www.w3.org/2001/sw/], 16.2.2007.

[38] Web Characterization Terminology & Definitions Sheet. World Wide Web Consortium. [URL: http://www.w3.org/1999/05/WCA-terms], 24.5.1999.

[39] WebTrends Web Analytics and Web Statistics. WebTrends Inc.. [URL: http://www.webtrends.com].

[40] Webtrends: The Essential Guide to Best Practices in e-commerce. WebTrends, 2006.

[41] Kožuh Boštjan: Akcijski načrt za izvedbo sprememb na spletni strani Statističnega urada in povezanih aktivnosti. Statistični urad RS, interno gradivo, 2005b.

[42] The Global Outlook for Personalization Applications. DMReview.com. [URL: http://www.dmreview.com/article_sub.cfm?articleId=4005], 21.9.2001.

Slovarček slovenskih prevodov tujih izrazov

Angleški izraz	Slovenski izraz
Accuracy chart	Grafikon natančnosti
Adaptable system	Adaptibilni sistem
Adaptive system	Prilagodljivi sistem
Antecedent window	Predhodno okno
Artificial intelligence	Umetna inteligenca
Association rule	Asociacijsko pravilo
Authority	Uveljavljeno spletišče
Auxiliary page	Pomožna stran
Back-end system	Zaledni sistem
Backward reference	Vzvratna referenca
Bottom-up approach	Pristop od spodaj navzgor
Branch	Veja
Brute-force method	Metoda grobe sile
Cache-busting	Razbijanje predpoomnjenja
Caching	Predpomnjenje
Case	Primer
Case table	Tabela primerov
Citation analysis	Analiza navedkov
Class	Razred
Class attribute	Razredni atribut
Classiffication matrix	Klasifikacijska matrika
Classifier	Klasifikator
Click-stream	Potek povezav
Cluster	Gruča, skupina
Cluster decomposition	Razstavljanje gruč
Clustering	Razvrščanje v skupine
Collaborative filtering	Filtriranje na osnovi sodelovanja
Conceptual hierarchy	Konceptualna hierarhija
Conceptual hierarchy	Konceptualna hierarhija
Confidence	Zaupanje
Confusion matrix	Klasifikacijska matrika
Consequent window	Sledeče okno
Constructed session	Konstruirana seja
Contact efficiency	Učinkovitost stika
Content filtering	Filtriranje na osnovi vsebine
Content page	Vsebinska stran
Contiguous sequence	Stično zaporedje
Conversion efficiency	Učinkovitost pretvorbe
Cookie	Piškotek
Cosine similarity	Kosinusna podobnost
Coverage	Pokritje
Customization	Prikrojevanje
Data mining	Podatkovno rudarjenje, rudarjenje po podatkih
Data pattern processing	Procesiranje podatkovnih vzorcev
Data visualization	Vizualizacija podatkov
Data warehouse	Skladišče podatkov
Database	Zbirka podatkov, podatkovna baza
Decision rule	Odločitveno pravilo
Description logic	Opisna logika
Distinctiveness	Raznolikost
Domain	Vsebinsko področje
Domain event model	Področni model dogodkov
Domain knowledge	Znanje o področju

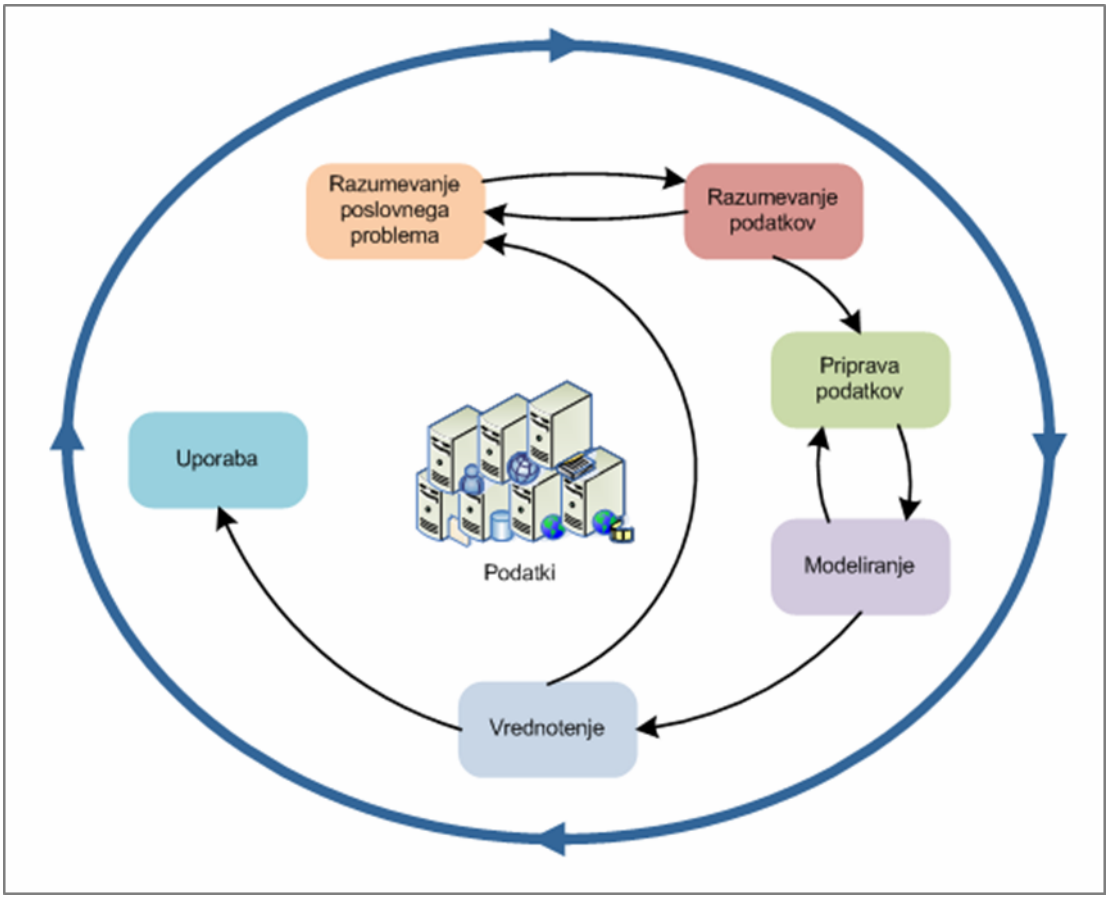
Angleški izraz	Slovenski izraz
Domain ontology	Področna ontologija
Download manager	Upravitelj prenosov
Downward closure	Zapiranje navzdol
Episode	Epizoda
Exploratory search	Raziskovalno iskanje
Exploratory statistics	Raziskovalna statistika
First order logic	Logika prvega reda
Floating layer	Lebdeča plast
Flow-chart	Grafikon poteka
Forward reference	Napredujoča referenca
Fraud detection	Odkrivanje prevar
Front-end system	Čelni sistem
Group usage profile	Skupinski profil uporabe
Guidance	Vodenje
Hard clustering	Trdo razvrščanje v skupine
Hierarchical method	Hierarhična metoda
Hub	Zvezdišče
Hyperlink	Hiperpovezava, spletna povezava
If-then rule	Pravilo če-potem
Information discovery	Odkrivanje informacij
Information foraging theory	Teorija iskanja informacij
Information harvesting	Žetje informacij
Information retrieval	Pridobivanje informacij
Information scent	Informacijska sled
Interaction economy	Interakcijsko gospodarstvo
Internet	Internet
Item	Predmet
Java applet	Javanski programček
Keyword	Ključna beseda
Killer application	Najperspektivnejša aplikacija
Knowledge discovery in databases	Odkrivanje znanja v zbirkah podatkov
Knowledge extraction	Izkopavanje znanja
Knowledge mining	Rudarjenje znanja
Latest-subsequence rule	Pravilo zadnjega podniza
Latest-subsequence rule	Pravilo zadnjega zaporedja
Leaf	List
Lift chart	Grafikon dviga
Link prediction	Napovedovanje povezav
Log table	Dnevniška tabela
Machine learning	Strojno učenje
Market basket analysis	Analiza nakupovalne košarice
Markov chain	Markovska veriga
Markov model	Markovski model
Memorization	Pomnjenje
Mining model	Model podatkovnega rudarjenja
Mining structure	Struktura podatkovnega rudarjenja
Mixture model	Mešani model
Model order	Red modela
Model-based method	Metoda, ki temelji na modelu
Moving window	Premikajoče okno
Multimedia	Večpredstavnostne vsebine
Naive Bayesian classifier	Naivni Bayesov klasifikator
Natural language processing	Procesiranje naravnega jezika
Navigation page	Navigacijska stran
Navigation pattern	Navigacijski vzorec
Nested key	Gnezdeni ključ
Nested table	Gnezdena tabela

Angleški izraz	Slovenski izraz
New item problem	Problem novega predmeta
Node	Vozlišče
Nondestructive change	Neškodljiva sprememba
Non-text file	Neznakovna datoteka
Object	Objekt
Object	Predmet
Observational personalization	Opazovalno poosebljanje
Offline	Nepovezan
Online	Povezan
Ontology	Ontologija
Ontology engineering	Grajenje ontologije
Ontology language	Ontološki jezik
Ontology learning	Učenje ontologije
Overfitting	Predobro prileganje
Page hit	Zadetek
Page interest estimator	Cenilka zanimivosti strani
Page layout	Postavitev strani
Page view	Ogled strani
Partitioning method	Delitvena metoda
Path completion	Zaključevanje poti
Pattern recognition	Razpoznavanje vzorcev
Persistent cookie	Trajni piškotek
Personalization	Poosebljanje
Personalization policy	Politika poosebljanja
Pessimistic confidence	Pesimistično zaupanje
Platform for privacy preferences, P3P	Platforma za zasebnostne nastavitve
Plug-in	Vtičnik
Predicate	Povedek
Prediction	Napoved
Proactive	Proaktiven
Probabilistic latent semantic analysis	Verjetnostna prikrita semantična analiza
Process instance	Procesni primer
Profit chart	Grafikon dobička
Query	Poizvedba
Query string	Poizvedovalni niz
Reactive	Reaktiven
Real session	Prava seja
Real-time	Stvarni čas
Reference length	Dolžina reference
Referrer	Napotitelj
Robot	Robot
Rule body	Telo pravila
Rule head	Glava pravila
Scatter accuracy chart	Raztreseni grafikon natančnosti
Semantic web	Semantični splet
Sequence	Zaporedje
Sequence pattern	Zaporedni vzorec
Server session	Strežniška seja
Session cookie	Piškotek seje, začasni piškotek
Singleton query	Enkratna poizvedba
Social network	Socialna mreža
Sparse data	Raztreseni podatki
Spider	Pajek
State space	Prostor stanj
Stretchtext	Raztegljivo besedilo
Subject	Osebek
Subsequent rule	Pravilo zaporedja

Angleški izraz	Slovenski izraz
Subset rule	Pravilo podmnožice
Substring rule	Pravilo podniza
Support	Podpora
Task performance support	Podpora za učinkovitejše izvajanje nalog
Template	Predloga
Text mining	Rudarjenje po besedilih
Threshold	Mejna vrednost
Top-bottom approach	Pristop od zgoraj navzdol
Transition probability	Prehodna verjetnost
Transition probability matrix	Prehodna matrika
Traversal path	Pot prečenja
Usability testing	Testiranje uporabnosti
User	Uporabnik
User activity log	Dnevnik uporabniške aktivnosti
User session	Uporabniška seja
Vector space model	Model vektorskega prostora
Visit	Obisk
Web content mining	Rudarjenje po vsebini spleta
Web log file	Spletna dnevniška datoteka
Web log mining	Rudarjenje po spletnih dnevniških datotekah
Web mining	Rudarjenje po spletu
Web page	Spletna stran
Web site graph	Topologija spletnega grafa
Web structure mining	Rudarjenje po strukturi spleta
Web usability	Spletna uporabnost
Web usability	Uporabnost spletišča, spletna uporabnost
Web usage mining	Rudarjenje po podatkih o uporabi spleta
Website	Spletišče
Word-vector	Vektor besed
World wide web	Svetovni splet

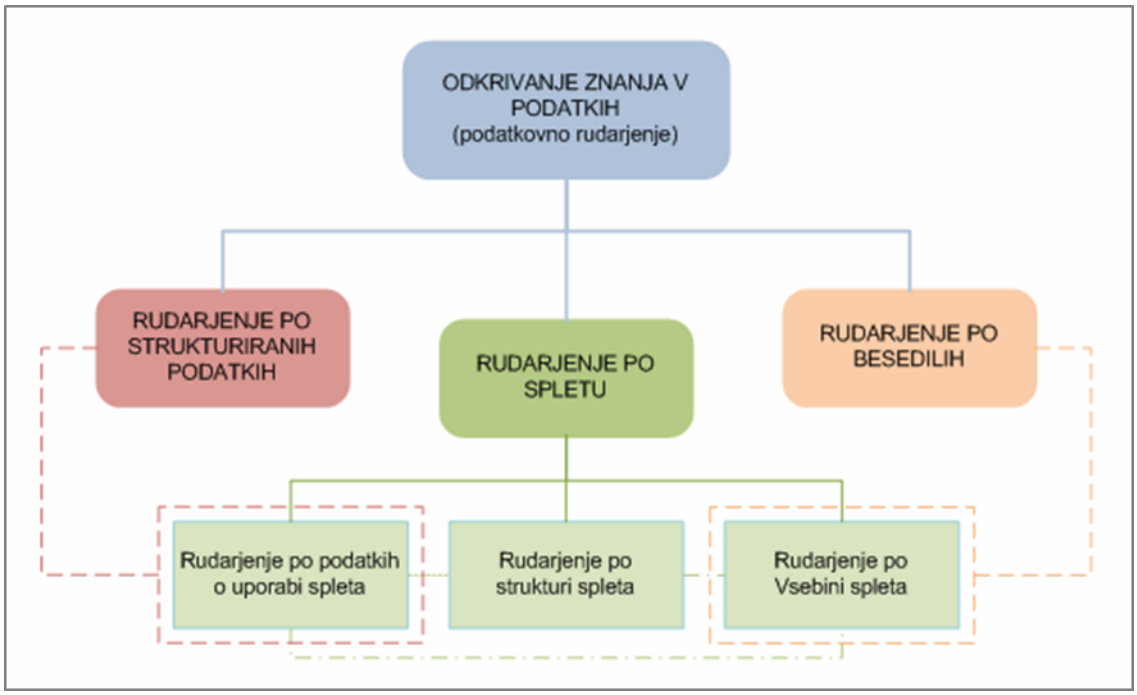
Priloge

Priloga 1: Faze v procesu podatkovnega rudarjenja kot ga definira metodologija CRISP-DM 1.0



Vir: prirejeno po Chapman, 1999, str. 13

Priloga 2: Taksonomija odkrivanja znanja v podatkih



Vir: lastno delo

Priloga 3: Seznam podatkov v spletni dnevniški datoteki v skladu z razširjenim W3C zapisom

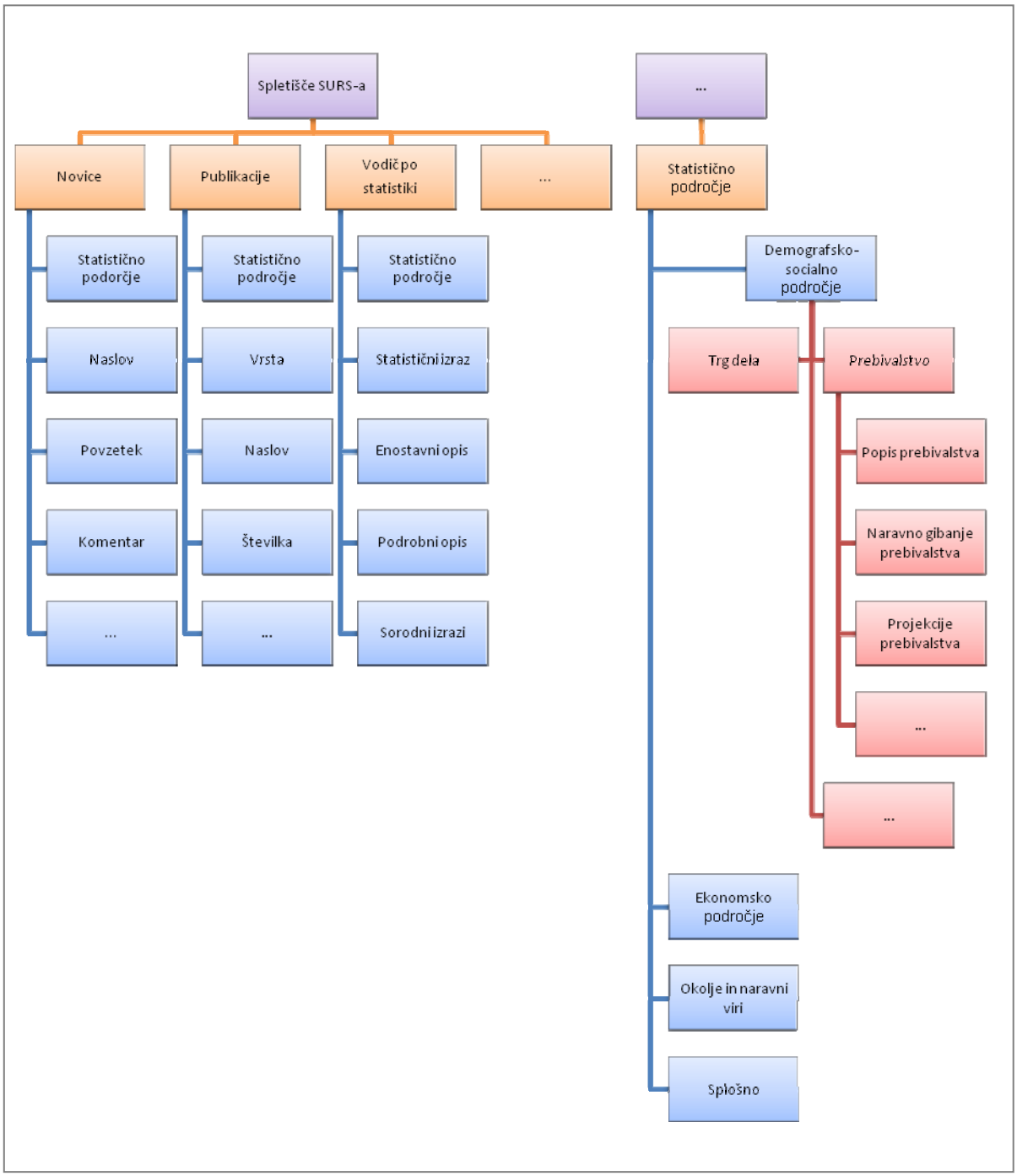
Polje	Oznaka	Opis
Datum	date	Datum izvedbe zahtevka.
Ura	time	Ura izvedbe zahtevka.
IP naslov odjemalca	c-ip	IP naslov odjemalca, ki je sprožil zahtevek.
Uporabniško ime	cs-username	Ime avtenticiranega uporabnika, ki se dostopal do strežnika. Anonimne uporabnike označuje pomišljaj.
Ime servisa in številka instance	s-sitename	Ime internetnega servisa, ki se je izvajal na odjemalcu, in njegova instančna številka.
Ime strežnika	s-computername	Ime strežnika, na katerem se je generiral zapis v dnevniški datoteki.
IP naslov strežnika	s-ip	IP naslov strežnika, na katerem se je generiral zapis v dnevniški datoteki.
Strežniška vrata	s-port	Številka uporabljenih strežniških vrat.
Metoda	cs-method	Zahtevano dejanje (npr. metoda GET)
URI naslov	cs-uri-stem	URI naslov prenesene datoteke.
URI poizvedba	cs-uri-query	Poizvedba (če obstaja), ki jo je poskušal izvesti uporabnik.
HTTP status	sc-status	Kodna številka HTTP statusa.
Win32 status	sc-win32-status	Kodna številka statusa operacijskega sistema.
Količina poslanih podatkov v bajtih	sc-bytes	Količina podatkov v bajtih, ki jih je poslal strežnik.
Količina prejetih podatkov v bajtih	cs-bytes	Količina podatkov v bajtih, ki jih je prejel strežnik.
Čas, potreben za izvedbo	time-taken	Trajanje izvajanja zahtevka v milisekundah.
Verzija protokola	cs-version	Verzija protokola (HTTP ali FTP), ki jo je uporabljal odjemalec.
Gostitelj	cs-host	Ime gostiteljske glave (če obstaja).
Uporabniški agent	cs(User-Agent)	Tip brskalnika, ki jo je uporabljal odjemalec.
Piškotek	cs(Cookie)	Vsebina poslanega piškotka (če obstaja).
Napotitelj	cs(Referrer)	URI naslov strani, prek katere je uporabnik prišel na trenutno stran. Zahtevke brez napotitelja označuje pomišljaj.
Podstatus protokola	sc-substatus	Kodna številka podstatusa protokola.

Vir: W3C Extended Log File Format (IIS 6.0), 2007.

Priloga 4: Primer P3P politike zasebnosti v obliki XML za spletišče SURS

```
<?xml version="1.0"?>
<POLICIES xmlns="http://www.w3.org/2002/01/P3Pv1">
    <EXPIRY date="Mon, 31 Dec 2007 12:00:00 GMT"/>

    <POLICY
        name="Politika varovanja zasebnosti"
        discuri="http://www.stat.si/zasebnost"
        opturi="http://www.stat.si/narocniki_stran.asp"
        xml:lang="sl"

        <ENTITY>
            <DATA-GROUP>
                <DATA ref="#business.contact-info.telecom.telephone.number">01/241-51-00</DATA>
                <DATA ref="#business.contact-info.online.email">info@stat.si</DATA>
                <DATA ref="#business.contact-info.online.uri">http://www.stat.si</DATA>
                <DATA ref="#business.contact-info.postal.organization">Statistični urad RS</DATA>
                <DATA ref="#business.contact-info.postal.street">Vožarski pot 12</DATA>
                <DATA ref="#business.contact-info.postal.city">Ljubljana</DATA>
                <DATA ref="#business.contact-info.postal.postalcode">1000</DATA>
                <DATA ref="#business.contact-info.postal.country">Slovenija</DATA>
                <DATA ref="#business.name">Statistični urad RS</DATA>
            </DATA-GROUP>
        </ENTITY>

        <ACCESS><nonident/></ACCESS>

        <DISPUTES-GROUP>
            <DISPUTES resolution-type="service" service="http://www.stat.si/pritozba"
                short-description="Pritožba">
                    <LONG-DESCRIPTION>
                        V primeru kakršnekoli nejasnosti glede spoštovanja vaše zasebnosti ali
                        dvoma glede njenega spoštovanja se lahko obrnete na naše Informacijsko središče.
                    </LONG-DESCRIPTION>
                    <REMEDIES><correct/></REMEDIES>
            </DISPUTES>
        </DISPUTES-GROUP>

        <STATEMENT>
            <EXTENSION optional="yes">
                <GROUP-INFO xmlns="http://www.software.ibm.com/P3P/editor/extension-1.0.html"
                    name="Basic information"/>
            </EXTENSION>
            <CONSEQUENCE>
                Pri vseh spletnih obiskovalcih se spremljajo naslednji podatko: podatki o dostopu
                do spletnih strani in morebitne poivedbene nize.
            </CONSEQUENCE>
            <PURPOSE><admin/><current/><develop/></PURPOSE>
            <RECIPIENT><ours/></RECIPIENT>
            <RETENTION><indefinitely/></RETENTION>
            <DATA-GROUP>
                <DATA ref="#dynamic.clickstream"/>
                <DATA ref="#dynamic.http"/>
                <DATA ref="#dynamic.searchtext"/>
            </DATA-GROUP>
        </STATEMENT>

        <STATEMENT>
            <EXTENSION optional="yes">
                <GROUP-INFO xmlns=http://www.software.ibm.com/P3P/editor/extension-1.0.html
                    name="Cookies"/>
            </EXTENSION>
            <CONSEQUENCE>
                Piškotke uporabljamo za spremljanje obiskovalcev na našem spletišču, da lahko
                bolje razumemo, kateri deli spletišča najbolj ustrezajo vašim potrebam.
            </CONSEQUENCE>
            <PURPOSE><develop/><tailoring/></PURPOSE>
            <RECIPIENT><ours/></RECIPIENT>
            <RETENTION><business-practices/></RETENTION>
            <DATA-GROUP>
                <DATA ref="#dynamic.cookies" optional="yes">
                    <CATEGORIES><uniqueid/></CATEGORIES>
                </DATA>
            </DATA-GROUP>
        </STATEMENT>

    </POLICY>

</POLICIES>
```

Priloga 5: Delna ontologija statističnih podatkov Statističnega urada RS



Vir: spletišče SURS.

Priloga 6: Primer RDF izjave za stavek »Okrajšava za Statistični urad Republike Slovenije je SURS« v zapisih RDF/XML in N-Triples ter Dublin Core semantiko za opis povedka

RDF/XML

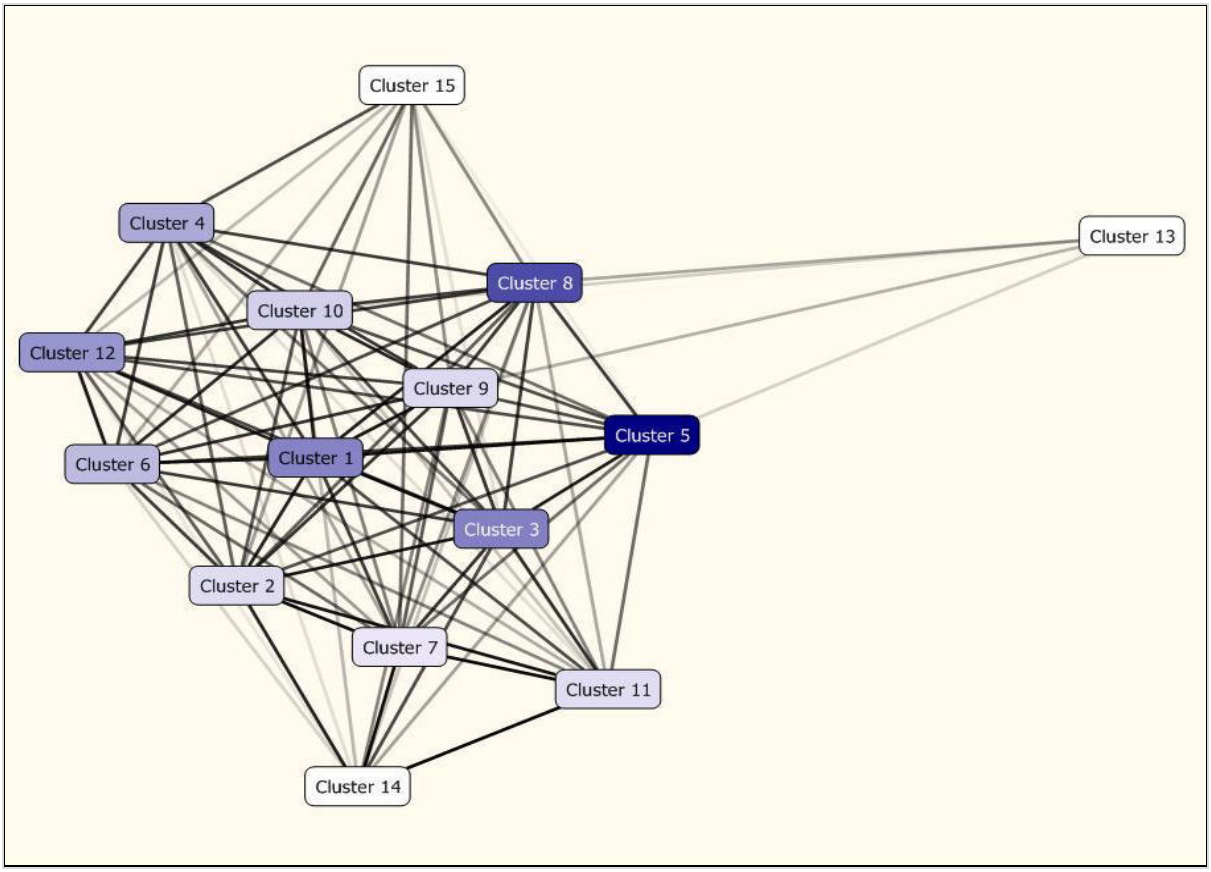
```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:terms="http://purl.org/dc/terms/">
 <rdf:Description rdf:about="urn:uradi:Statisti%E8ni%20urad%20Republike%20Slovenije">
 <terms:alternative>SURS</terms:alternative>
 </rdf:Description>
</rdf:RDF>
```

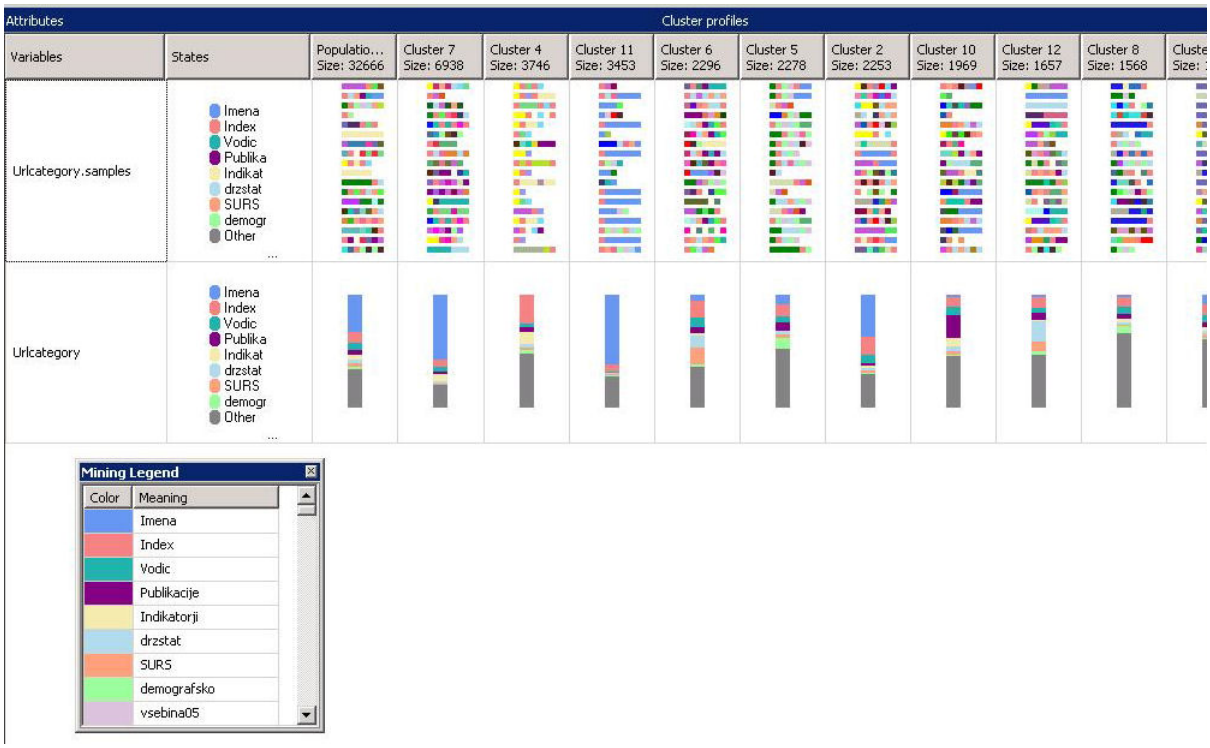
N-Triples

```
<urn:uradi:Statisti%E8ni%20urad%20Republike%20Slovenije>
<http://purl.org/dc/terms/alternative>
"SURS"
```

Vir: Resource Description Framework, 2007.

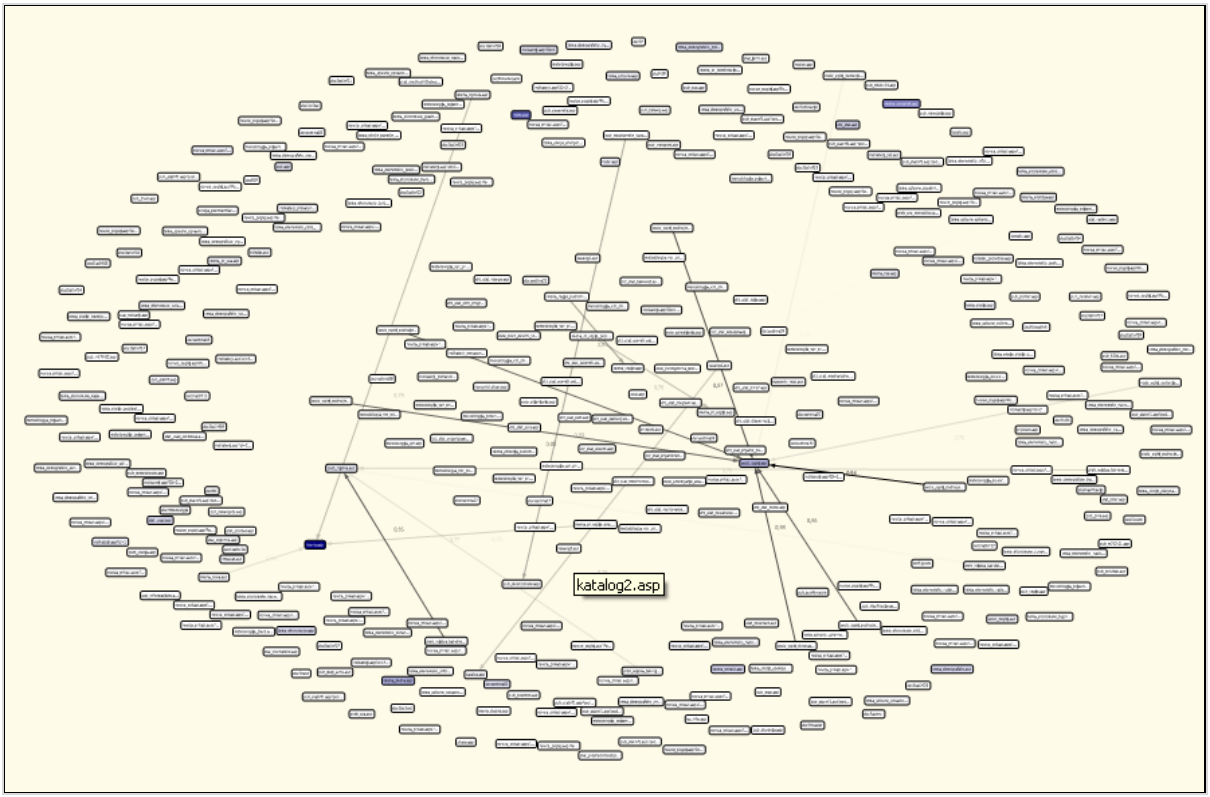
Priloga 7: Prikaz gruč, odkritih z algoritmom SequenceClustering, in njihovih profilov v pregledovalnikih orodja BIDS (primer za model *Kategorije10*)

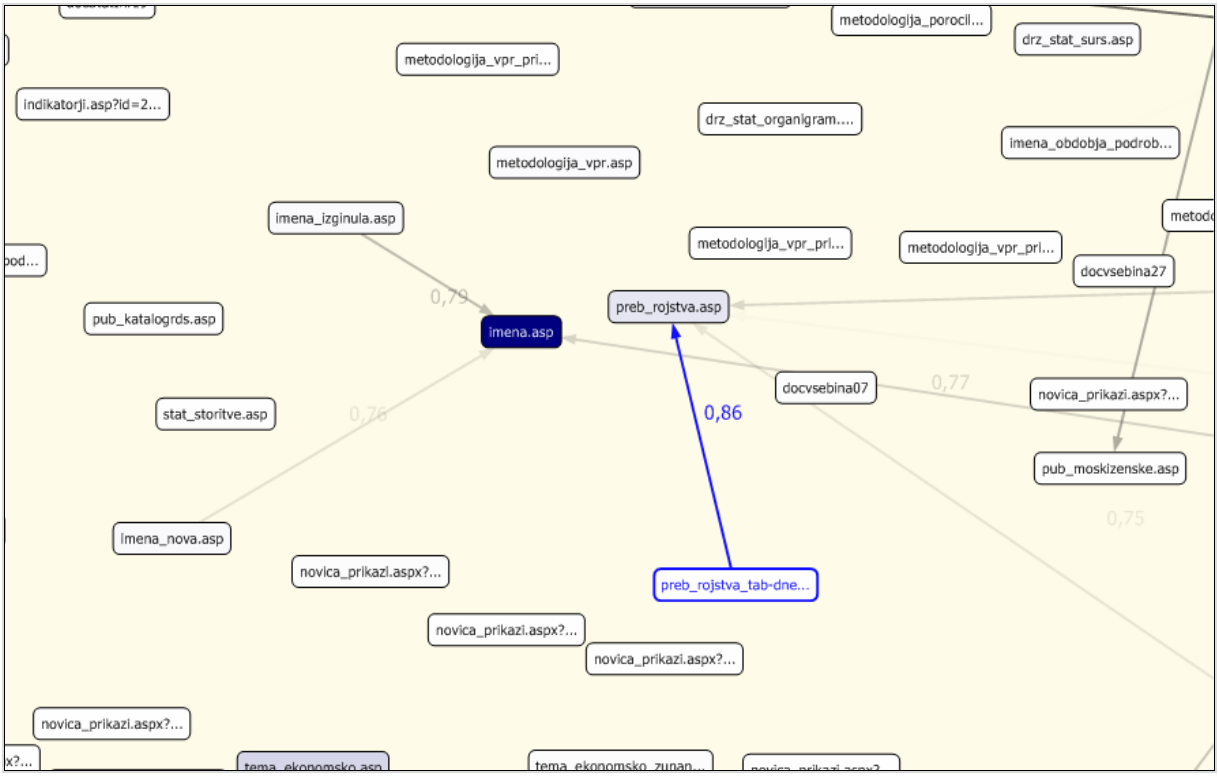




Vir: rezultati rudarjenje po podatki o uporabi spletišča SURS.

Priloga 8: Prikaz najznačilnejših prehodnih povezav med posameznimi strani, najdenih v okviru modela *Strani10* (povečana slika prikazuje podrobnejši pogled na povezave najbolj obiskane strani na spletišču)





Vir: rezultati rudarjenje po podatki o uporabi spletišča SURS.

Priloga 9: Struktura enkratne napovedne DMX poizvedbe, uporabljene pri rešitvi za priporočanje vsebine na spletišču SURS

```
SELECT FLATTENED
    Cluster(),
    (SELECT $Sequence,
            ObiskanaStran,
            PredictProbability(ObiskanaStran) As Verjetnost
     FROM PredictSequence([SC.ZaporedjeObiskov],p)
    ) As Napoved
FROM [model]
NATURAL PREDICTION JOIN
    (SELECT
        (SELECT 1 As Zaporedje_ID, 'stran1' As ObiskanaStran
         UNION
         SELECT 2 As Zaporedje_ID, 'stran2' As ObiskanaStran
         ...
         UNION
         SELECT n as Zaporedje_ID. 'stranN' As ObiskanaStran
        ) As [SC.ZaporedjeObiskov]
     ) As t
```

p *- število želenih priporočil*
model *– ime uporabljenega modela podatkovnega rudarjenja*
n *- število že obiskanih strani v okviru uporabniške seje*