UNIVERZA V LJUBLJANI

EKONOMSKA FAKULTETA

DIPLOMSKO DELO

UPORABA SPLETNEGA RUDARJENJA

Ljubljana, junij 2005 DAMJAN HARISCH

KAZALO

1 UVOD

Svet je postal vse bolj povezan. Skoraj vsaka pomembna poslovna aktivnost se sedaj zabeleži in prenese v bazo podatkov. Ta baza podatkov nam lahko predstavlja ključ do uspeha našega poslovanja. Rudarjenje podatkov je iskanje vzorcev v množici podatkov, na osnovi česar lahko omogočimo uspešnejše poslovanje.

Vsekakor se mi zdi pomembno, da se že na začetku spoznamo s pojmoma rudarjenje podatkov (lahko tudi podatkovno rudarjenje) in spletno rudarjenje. Že iz samih imen je razvidno, da sta si pojma zelo sorodna. Najprej se bom posvetil podatkovnemu rudarjenju, ki je predhodnik spletnega rudarjenja.

Rudarjenje podatkov se uporablja na različnih področjih in ima pomemben prispevek pri odkrivanju znanja iz podatkov (npr. odkritje presenetljivih značilnosti neke skupine predmetov) na različnih področjih, kot so medicina, genetika ali elektronska trgovina. Spletno rudarjenje pa se je razvilo šele v zadnjih nekaj letih in bo po pozornosti in zrelosti kmalu prevzelo položaj rudarjenju podatkov. Danes je predvsem pomembno, da je model poslovanja čim bolj usmerjen k uporabniku. Takšen model bo na svetovnem spletu igral vse pomembnejšo, če ne že glavno vlogo. Da bi učinkovito zadovoljili strankine potrebe, je za ponudnike spletnih storitev zelo pomembno, da čim prej zvedo kaj o samih strankah, kar pa je dostikrat vse prej kot enostavno. Tu se pokaže pomembna vloga spletnega rudarjenja. Spletno rudarjenje je potomec rudarjenja podatkov. Če je za rudarjenje podatkov značilno iskanje skritih vzorcev in odnosov v bazah in skladiščih podatkov, lahko rečemo, da spletno rudarjenje vključuje iskanje podobnih povezav iz podatkov, ki jih dobimo na podlagi klikanja z miško na spletni strani. S spletnim rudarjenjem bodo ponudniki spletnih storitev sposobni sestaviti profil ljudi, ki obiskujejo njihovo spletno stran, prav tako pa bodo vedeli, kakšne informacije obiskovalci strani običajno iščejo in tudi na kakšen način jih iščejo. Tako lahko napravijo takšno spletno stran, ki bo uporabnikom omogočala, da najdejo želeno informacijo s čim manj kliki.

Končni cilj združitve spletnega rudarjenja in dinamične tehnologije spletnih strani (angl. dynamic web page technology) je personalizacija izkušenj. Tako bo lahko vsak posameznik videl spletno stran, ki je oblikovana zgolj za njegove potrebe, takoj ko se nanjo prijavi (logira).

2 RUDARJENJE PODATKOV

Danes že vsaka organizacija, ki se ukvarja z informacijskimi sistemi, verjame, da je rudarjenje podatkov pomemben del njene prihodnosti in da je povezano z obstoječimi investicijami na področju skladiščenja podatkov. Vendar pa se za vsem tem navdušenjem skriva tudi precej nejasnosti. Kaj v bistvu je rudarjenje podatkov? Je to le neko splošno ime za analizo podatkov in ali za to potrebujemo kakšna posebna orodja ter znanja? Ali je skladiščenje podatkov neko razumljivo področje, ki temelji na znanju, ali pa je zgolj nek zbir nezdružljivih tehnik? In ko enkrat poznamo področje rudarjenja podatkov, ali lahko avtomatično uporabimo spletno skladišče podatkov (angl. webhouse), da pričnemo z rudarjenjem, ali pa je potrebno te podatke še dodatno pripraviti (Kimball, Merz, 2001, str. 251)?

Uporabo rudarjenja podatkov upravičijo velike dimenzije podatkov in njihova kompleksnost. Razvilo se je predvsem iz strojnega in statističnega učenja. Omogoča nam odkrivanje značilnih vzorcev iz velike množice podatkov, ki jih običajna statistična analiza spregleda. Pri odkrivanju znanja ne gre samo za večdimenzionalno predstavitev podatkov, kar je osnovni namen sistemov OLAP, ampak predvsem, da nam znanje tudi čim bolje »postreže« (Horvat, 2003, str. 1).

Rudarjenje podatkov je zbirka učinkovitih analitičnih tehnik, s pomočjo katerih lahko iz zelo obsežnih zbirov podatkov, izvlečemo smiselne in uporabne ugotovitve. Če jih znamo pravilno uporabiti, zna biti rudarjenje podatkov nekaj izredno dragocenega. Ne obstaja nek edinstven pristop rudarjenja podatkov, ampak je to bolj zbir različnih tehnik, ki se pogosto uporabljajo v medsebojni povezavi, da bi lahko iz podatkov, ki so nam na voljo, izvlekli čim več uporabnih informacij. Če nekdo veliko vlaga na področje rudarjenja podatkov, potem je moč pričakovati, da se bo srečal z mnogimi različnimi orodji, ki mu jih bo nudilo veliko različnih ponudnikov. Pri samem procesu rudarjenja podatkov, je zelo pomemben postopek skladiščenja podatkov (angl. warehousing), oziroma v primeru spletnega rudarjenja podatkov tudi spletnega skladiščenja podatkov (angl. webhousing). Namen skladiščenja podatkov je preskrba s pripravljenim naborom podatkov za rudarjenje podatkov. Ta nabor je običajno v obliki poročil, oziroma nekakšen razčlenjen skupek, skozi katerega lahko »vrtamo« (angl. drill-across), da pridemo do želenih informacij (Kimball, Merz, 2001, str. 251).

2.1 ZGODOVINA RUDARJENJA PODATKOV

Čeprav igra danes na tržišču za rudarjenje podatkov glavno vlogo mnogo novih proizvodov in podjetij, je še vedno osnova v bogati tradiciji raziskav in uporab, ki segajo zagotovo vsaj 30 let nazaj. Prvič se ime rudarjenje podatkov pojavi v

začetku šestdesetih let (1960), kot statistična analiza. Pionirji na področju statistične analize so bila podjetja SAS, SPSS in IBM. Vsa tri podjetja so tudi danes zelo aktivna na področju rudarjenja podatkov in ponujajo kar nekaj proizvodov, katerih uporaba temelji na dolgoletnih izkušnjah. V začetku so te statistične analize vsebovale klasične statistične postopke, kot so korelacija, regresija, hi-kvadrat (angl. chi-square) in tehniko kontingenčnih tabel (angl. cross-tabulation). Pri SAS in SPSS še vedno lahko zaznamo prisotnost tega klasičnega pristopa, vendar pa se je postopek rudarjenja podatkov oddaljil zgolj od uporabe statističnih postopkov. Približal se je bolj poglobljenim pristopom, kako iz zbranih podatkov nekaj razložiti oziroma napovedati (Kimball, Merz, 2001, str. 252, 253).

V poznih osemdesetih se je postopek klasične statistične analize nadgradil z bolj izbranimi tehnikami, kot so mehka logika (angl. fuzzy logic), hevristična metoda reševanja problemov (angl. heuristic reasoning) in nevronske mreže (angl. neural networks). To naj bi bil nekako višek umetne inteligence.

Konec devetdesetih smo spoznali, kako vzeti le najboljše iz različnih tehnik, kot so statistična analiza, nevronske mreže, drevesa odločanja, analiza nakupovalne košarice (angl. market basket analysis) in ostalih pomembnejših tehnik, ter kako jih združiti in uporabljati na čim bolj nadzorovan in učinkovit način. Verjamem, da je prihod resnih sistemov za skladiščenje podatkov potrebna sestavina, ki bo naredila rudarjenje podatkov bolj praktično in uporabno.

2.2 UPORABA RUDARJENJA PODATKOV

Rudarjenje podatkov je uspešno, če smo si sposobni neko informacijo razložiti tako, da je koristna in uporabna za sam management podjetja. S tem omogočamo dragocene odgovore na njegove negotove zahteve (Brandel, 2001, str. 67).

Lahko bi rekli, da je v današnjem svetu veliko povpraševanje in nizka ponudba znanja s področja rudarjenja podatkov. Veliko podjetij se je šele začelo prebujati pri izkoriščanju potencialov, ki jim jih ponujajo ogromne zaloge podatkov. Tako postaja rudarjenje podatkov vse pogosteje eden vodilnih dejavnikov pri tržnem uspehu podjetij. Podjetja, ki ga še ne uporabljajo, bodo prav gotovo v to prisiljena, že vsaj zaradi zagotovitve lastnega obstoja.

Za projekte, kjer so poudarki na rudarjenju podatkov, potrebujemo predvsem ljudi, ki imajo znanja s področij, kot so: poznavanje statističnih konceptov, temeljito razumevanje poslovnih znanj, znanja iz področja projektnega managementa (ta so posebej potrebna pri zelo hitrem razvoju dogodkov na področju raziskav in razvoja), izkušnje na področjih baz podatkov, skladiščenja podatkov, OLAP-a (angl. online analytical processing) in pa inteligentnih poslovnih sistemov. Včasih

iščemo vse te lastnosti kar v eni osebi. Ostala potrebna znanja bi lahko bila tudi izurjenost z orodji za dostop do baze podatkov (na primer SQL) in izkušnje z orodji za rudarjenje podatkov.

Vendar pa sta od vseh področij, ki naj bi jih obvladali tisti, ki se ukvarjajo z rudarjenjem podatkov, najbolj pomembni predvsem analiza podatkov in obvladanje poslovnih znanj. Slep je vsak, če počne nekaj, v čemer ne vidi nobenega poslovnega smisla. Seveda moraš poznati tehnike rudarjenja podatkov in uporabo za to namenjenih orodij, vendar pa je dosti bolj pomembno, da iz te informacije izluščimo tisto, kar lahko management podjetja nato koristno uporabi (Brandel, 2001, str. 67).

Zelo pomemben dejavnik je, kako prenesti splošno zahtevo managerjev, v neko koristno informacijo. Pojavi se namreč vprašanje, kakšno je njihovo znanje oziroma razumevanje samih podatkov in tehnik za rudarjenje podatkov. Mnogim ljudem, ki sestavljajo višji management podjetja, se niti ne sanja, kako celotna zadeva funkcionira. Včasih jih je potrebno strezniti z izjavami, kot so: seveda je to mogoče, vendar pa bi za to potrebovali najmanj 10 let. To pa potem prav gotovo ne bi imelo nobenega smisla.

Pri številnih aplikacijah, ki se danes pojavljajo na trgu, niso potrebna tako obsežna znanja s področij statistike in programiranja. Orodja za rudarjenje podatkov postajajo uporabniku vse prijaznejša. Podrobnosti dejanskih algoritmov so vse bolj zakrite, tako da so vidni zgolj poslovni parametri, ki nas zanimajo. Aplikacije so za uporabnike tudi dovolj enostavne, vendar pa še vedno potrebujemo nekoga, da nam pripravi podatke in zagotovi njihovo točnost (Brandel, 2001, str. 67).

2.3 AKTIVNOSTI PRI RUDARJENJU PODATKOV

Rudarjenje podatkov lahko razčlenimo na štiri glavne aktivnosti (Kimball, Merz, 2001, str. 253-255):

- zbiranje podatkov,
- klasificiranje,
- ocenjevanje in predvidevanje ter
- razvrščanje po sorodnosti.

2.3.1 ZBIRANJE PODATKOV

Zbiranju podatkov bi lahko rekli tudi priprava podatkov za spletno rudarjenje. Primer zbiranja je pregled seznama, na katerem se na začetku nahaja veliko število strank, ki se med seboj ne razlikujejo po nobenem kriteriju. S tem skušamo ugotoviti ali spadajo v naravne skupine. To je čisti primer »posrednega rudarjenja

podatkov«, kjer uporabnik nima vnaprej določenega postopka in upa, da bo orodje za rudarjenje podatkov samo od sebe odkrilo kakšne pomembne strukture. Podatki, ki jih dobimo pri takem zbiranju, bi morali biti nekakšen dolgovezen opis vsake stranke, tako z demografskimi kot tudi vedenjskimi kazalci, ki jih nato »pritrdimo« vsakemu vnosu. Priprava teh dolgoveznih opisov je glavna naloga skladišča podatkov. Običajno je zelo dolgotrajno in drago izdelovati takšne opise, ker moramo »vrtati« (angl. drill across) po mnogih tabelah dejstev v verigi vrednosti. Ko pa so ti opisi zaključeni, so lahko na voljo kot nekakšno zaključeno poročilo za zelo hitro uporabo orodij za rudarjenje podatkov.

2.3.2 KLASIFICIRANJE

Primer klasificiranja je preučevanje potencialnih strank, ki jim določimo že definirano gručo (angl. cluster) oziroma klasifikacijsko skupino. Primer klasifikacije je tudi medicinska diagnoza. V obeh primerih moramo dolgovezen opis stranke ali pacienta, ki ga dobimo, vstaviti v klasifikacijski algoritem. Oseba, ki je zadolžena za kvalificiranje določi, v katero gručo bi posamezna stranka ali pacient spadala, oziroma ji bila najbližje. Tako vidimo, da predhodni aktivnosti zbiranja, sledi aktivnost klasificiranja. Lahko rečemo, da je kvalificiranje, v nekem splošnem smislu, zelo uporabna aktivnost v okolju skladiščenja podatkov. Klasificiranje je sprejemanje odločitev. Stranke lahko klasificiramo glede na to ali so ali niso kreditno sposobne, paciente lahko klasificiramo po kriteriju ali potrebujejo zdravljenje ali ne ipd.

2.3.3 OCENJEVANJE IN PREDVIDEVANJE

Ocenjevanje in predvidevanje sta dve podobni aktivnosti, za kateri je značilno, da ponavadi opuščata numerične rešitve kot rezultat, npr. da imamo skupino obstoječih strank, ki ima enak opis kot potencialne stranke. Na podlagi te skupine lahko potem ocenimo celotno zadolženost potencialnih strank. Predvidevanje je zelo podobno ocenjevanju, le da skušamo tukaj določiti rezultat, ki naj bi se pojavil v nekem prihodnjem obdobju. Ocenjevanje in predvidevanje lahko usmerjata tudi sam proces klasifikacije, npr. pri odločanju ali so stranke, ki so zadolžene za več kot 100.000 USD, klasificirane kot nizko kreditno rizične. Numerično ocenjevanje ima tudi to prednost, da so lahko kandidati razvrščeni po rangih. Če imamo omejeno količino denarja v oglaševalskem proračunu, potem oglaševalsko ponudbo pošljemo 10.000 strankam, za katere smo ocenili, da so najvišje rangirane po bodoči vrednosti za naše podjetje. V tem primeru je samo ocenjevanje bolj koristno kot pa zgolj enostavna klasifikacija po zadolženosti (angl. binary classification).

2.3.4 RAZVRŠČANJE PO SORODNOSTI

Razvrščanje po sorodnosti je posebna vrsta zbiranja, ki nam da vedeti, kateri dogodki oziroma transakcije so se zgodili sočasno. Primer razvrščanja po sorodnosti je analiza nakupovalne košarice (angl. market basket analaysis). Analiza nakupovalne košarice nam skuša ugotoviti, kateri izdelki so prodani ob istem času. To zna biti, z vidika obdelave podatkov, zelo velik problem, saj vemo, da je v tipičnem okolju prodaje na drobno na voljo na tisoče različnih proizvodov. Nesmiselno bi bilo, da bi naštevali vse kombinacije izdelkov, ki so bili prodani skupaj, ker bi seznam dokaj hitro dosegel astronomske vrednosti. Umetnost analize nakupovalne košarice je v tem, da poskušamo najti kakšne pomembne kombinacije na različnih nivojih hierarhije izdelkov, ki so bili prodani skupaj. Npr. zelo pomembna bi bila ugotovitev, da se izdelek A (recimo neka priljubljena brezalkoholna pijača) zelo pogosto prodaja hkrati z izdelki, ki so kategorizirani kot hitro odmrznjena hrana.

3 SPLETNO RUDARJENJE

Spletno rudarjenje je interdisciplinarno in zelo dinamično znanstveno področje, ki vključuje področja, kot so: podatkovne baze, pridobivanje informacij, statistika, umetna inteligenca (strojno učenje in obdelava naravnega jezika) in druga. Kljub temu, da spada spletno rudarjenje v širše področje podatkovnega rudarjenja, ima več svojih značilnosti: integracija različnih podatkovnih virov, kot npr. zbirke dostopov, vpisov uporabnikov nekega spletnega mesta ali uporabniških profilov; razreševanje težav v zvezi z razpoznavo uporabnikov; razpoznava uporabniških sej ali epizod (transakcij) iz podatkov o uporabi, topologije mesta in modelov vedenja uporabnikov (Horvat, 2003, str. 45).

Zagotoviti želimo prilagodljivo delovanje spletnega mesta, ki bi lahko zadovoljevalo različne profile uporabnikov. Uporabniki imajo različne cilje, zahteve in želje, zato moramo do njih pristopati individualno (personalizacija) ali pa vsaj po več predvidenih scenarijih (angl. mass castomization). Pri tem jim ponudimo informacije različnih tipov, vsebin, struktur in/ali predstavitev.

3.1 RAZLIKA MED RUDARJENJEM PODATKOV IN SPLETNIM RUDARJENJEM

Medtem, ko je rudarjenje podatkov ugotavljanje zanimivih struktur v podatkih, je spletno rudarjenje odkrivanje in analiza uporabnih informacij s svetovnega spleta. Definicija vključuje samodejno iskanje in pridobivanje informacij ter virov, ki so na voljo preko spletnih mest in online podatkovnih baz, tj. rudarjenje po spletnih

vsebinah (angl. web content mining), kakor tudi odkrivanje in analizo vzorcev dostopa uporabnikov na osnovi podatkov o spletni uporabi, tj. spletno rudarjenje vzorcev uporabe (angl. web usage mining) (Horvat, 2003, str. 2).

Rudarjenje podatkov je odkrivanje znanja iz množice podatkov na podlagi pristopov strojnega učenja in statistike. Spletno rudarjenje je v širšem smislu odkrivanje in razlaga uporabnih informacij in znanja s svetovnega spleta. Definiramo ga lahko tudi kot (prilagojeno) uporabo metod in tehnik rudarjenja podatkov za svetovni splet. Obsega napovedovanje, razvrščanje po podobnosti, personalizacijo in druge tehnike, ki jih omogočajo navigacijski vzorci na podlagi zbirke dostopa spletnega strežnika, dokumenti besedil spletnega mesta in spletne storitve.

Tehnike za rudarjenje podatkov in spletno rudarjenje so enake, le da gre pri spletnem rudarjenju za podatke zbrane preko spleta. Rečemo lahko, da je internet postal nov vir podatkov, katere oblikujejo različna obnašanja uporabnikov. Tako imamo sposobnost opazovati razne vzorce obnašanja v določenih situacijah in jih nato razvrstiti (angl. classify) (Robinson, 2000, str. 34-35).

Bistvena razlika med rudarjenjem podatkov in spletnim rudarjenjem je v časovnem okvirju, v katerem se zgodi sama analiza. Pri rudarjenju podatkov traja analiza ponavadi tudi več tednov. Spletno rudarjenje je povezano s podatki, ki jih dobimo na podlagi tisočih klikov miške v eni uri, za analizo pa imamo na voljo samo nekaj minut ali pa celo sekund.

Na organizacijskem nivoju lahko nato izmerimo, kaj ljudje naredijo, da pridejo na našo stran in kakšna je uspešnost obiska naše strani. Če ljudje zapustijo našo stran, še preden dobijo kritično informacijo, potem moraš kot manager spletne strani vedeti, da imaš problem (Robinson, 2000, str. 34-35).

Prav tako je tu pomembna vloga skladiščenja podatkov. Spletno skladiščenje (angl. web warehousing) je pristop k izgradnji informacijskih sistemov. Primarne funkcije le-teh so identifikacija, zapisovanje, pridobivanje, shranjevanje in analiziranje informacij v obliki podatkov, besedil, grafike, slik, zvokov, filmov in ostalih večpredstavnostnih objektov s pomočjo uporabe spletne tehnologije. Njihov namen je pomagati posameznikom poiskati informacije, ki jih iščejo, in te informacije tudi učinkovito analizirati (Mattison, 1999, str. 8).

Spletno skladiščenje je ogrodje, ki določa zbirko orodij in procesov. Le-ti so namenjeni izgradnji uporabnih sistemov skladišč podatkov, katerih osnova je spletna tehnologija. Podatki, s katerimi operira spletno podatkovno skladišče, ne vsebujejo le besedil in števil, ampak tudi grafiko, zvok, video in ostale oblike.

Spletna podatkovna skladišča organizirajo in upravljajo shranjene podatke, vendar jih ne zbirajo, ravno tako ne kreirajo informacij, ampak jih zgolj pasivno obdelujejo. V tem pogledu se spletno skladiščenje le rahlo razlikuje od podatkovnega skladiščenja. O spletnem podatkovnem skladišču lahko govorimo, če v podatkovno skladišče vključimo podatke iz spleta, poleg tega pa podatke iz podatkovnega skladišča vključimo v splet (Hrastar, Krnc, Škoberne, 2003, str. 10,11).

3.2 OSNOVNE ZNAČILNOSTI SPLETNEGA RUDARJENJA

Orodja za spletno rudarjenje se od ponudnika do ponudnika razlikujejo, vendar pa je tehnika v osnovi pri vseh več ali manj enaka (Robinson, 2000, str. 34-36):

- zabeleženi podatki so v bazi podatkov segmentirani v skupine, v katere jih razvrstimo po določenih lastnostih. Te skupine so običajno opredeljene kot polja v bazi podatkov, ki vsebujejo podobne vrednosti, ali pa kot obsežnejši atributi, kot so podobni značaji kupcev oziroma obnašanj;
- vsako združevanje med skupinami, kot je npr. skupna storitev ali zahteva po informaciji, je določeno z vnaprej predpisanimi pravili. Vsi vzorci, ki se pojavijo v zvezi s tem združevanjem, so tako izolirani;
- model je zgrajen na podlagi prejšnjih opazovanj podatkov in nam lahko pove, kako bo npr. določen tip uporabnikov uporabljal bazo podatkov. Ta model se lahko nato uporabi pri prihodnjih uporabnikih, da ugotovimo, ali ustrezajo določenim vzorcem. Če jih večina ustreza, potem lahko ta model uporabimo, da predvidimo prihodnje vzorce;
- odstopanja od pričakovane norme in vzroki zanje se ugotavljajo z uporabo statistične analize in vizualnih tehnik (angl. visualization techniques).

Hitrost je pri spletnem rudarjenju zelo pomembna. Če je dovolj indikatorjev za spremenjeno uporabo vzorcev v neki bazi podatkov, potem moramo čim hitreje spremeniti model tipičnega uporabnika. Pri tradicionalnem rudarjenju podatkov lahko taka odločitev traja več tednov ali mesecev. Vendar pa si na svetovnem spletu tega enostavno ne moremo privoščiti, saj se uporabnikovi profili lahko menjajo iz dneva v dan.

Vsekakor je potrebno omeniti, da je eden glavnih ciljev spletnega rudarjenja, izboljšanje oblike spletne strani. Vendar pa se tu pojavi problem. Ravno proces spletnega rudarjenja zahteva za uspešno delovanje predhodno dobro oblikovano spletno stran, katera naj bi zagotavljala, da so podatki zajeti na dovolj kakovostni ravni. Tako kot je treba dovolj časa posvetiti temu, da zgradimo ustrezno skladišče podatkov, je že na samem začetku potrebno zagotoviti primerno oblikovano spletno stran. Drugače se nam lahko zgodi, da imamo polno nepotrebnih in neuporabnih podatkov (Robinson, 2000, str. 34-35).

3.3 VRSTE SPLETNEGA RUDARJENJA

Kaj predstavlja pojem spletnega rudarjenja? To prav gotovo ni enostavno razložiti, kajti ta pojem se ponavadi uporablja za poimenovanje treh dokaj različnih aktivnosti. Vse te aktivnosti se lahko kvalificirajo kot podatkovno rudarjenje in prav tako vse vsebujejo delovanje na svetovnem spletu. Vendar pa se razlikujejo glede na to, po kakšnih oziroma katerih podatkih bomo rudarili, kot tudi po motivu za rudarjenje. Te tri različne aktivnosti so naslednje (Linoff, Berry, 2001, str. 21, 22):

- rudarjenje po strukturi (angl. mining structure),
- rudarjenje vzorcev uporabe (angl. mining usage) in
- rudarjenje po vsebini (angl. mining content).

Rudarjenje po strukturi je proces iskanja informacij iz topologije svetovnega spleta – povezav med spletnimi stranmi. Katere strani so »tarče« (angl. targets) povezav iz drugih strani? Katere strani te usmerjajo k nekim drugim? Katere zbirke strani tvorijo otoke?

Rudarjenje vzorcev uporabe je proces iskanja informacij glede na to, kako ljudje »prehodijo« (angl. traverse) te povezave z brskalniki in kako jih uporabijo. Katere strani obiščejo? Kako dolgo ostanejo na določeni strani? Kaj potem klikajo na naslednji? Katere poti vodijo do odjave iz določene strani, katere pa do direktnega odhoda iz te strani?

Rudarjenje po vsebini je proces iskanja koristnih informacij iz samega teksta spletnih strani, kot tudi iz slik in ostalih oblik vsebine, ki se lahko nahajajo na strani. Katera stran nam najbolje predstavi recept za potico? Katere strani so napisane v nemščini? Katere strani so v sorodu z vzrejo psov? Iskalniki (angl. search engines), inteligentni agentje (angl. inteligent agents) in nekateri priporočilniki (angl. recommendation engines) uporabljajo rudarjenje po vsebini za pomoč uporabnikom pri iskanju igle v kupu sena, ki ga predstavlja svetovni splet.

Čeprav se razlike med temi tremi različnimi vrstami spletnega rudarjenja na prvi pogled ne zdijo tako velike, pa temeljijo na pomembnem praktičnem razmisleku, da viri in razpoložljivost podatkov niso povsod enaki. Npr. dejstvo, da je določena zbirka med seboj različnih strani povezana s tako imenovanimi hiperpovezavami (angl. hyperlinks), pomeni, da je to vsem javno dostopna informacija, ki se lahko takoj uporabi za aktivnost rudarjenja po strukturi. Po drugi strani pa so podatki o tem, kako pogosto so bile te povezave uporabljene in kdo jih je uporabljal, zabeleženi na številnih strežnikih, ki so v lasti različnih podjetij. Le-ta pa verjetno niso pripravljena takšnih podatkov deliti ravno z vsakomur.

Rudarjenje po strukturi se tiče tako globalne strukture celotnega svetovnega spleta, kot tudi lokalnih struktur posameznih spletnih strani. Rudarjenje vzorcev uporabe uporablja potek povezav (angl. clickstream) kot primarni vir podatkov za rudarjenje uporabniškega obnašanja. Ukvarja se s težavnim preoblikovanjem spletnih prijav (angl. web logs) v zbirko podatkov, ki jih nato lahko uporabimo za rudarjenje. Rudarjenje po vsebini pa je povezano z iskanjem informacij s pomočjo iskalnikov.

3.3.1 SVETOVNI SPLET Z VIDIKA SPLETNEGA RUDARJENJA

Osnovni material za rudarjenje po strukturi je zbirka hiperpovezav, ki združuje dokumente. Osnovni material za rudarjenje po vsebini pa vsebuje tekst, shranjen v ogromnem številu datotek, ki so na voljo vsakomur, ki ima s svojim brskalnikom (angl. browser) dostop do svetovnega spleta. Tako rudarjenje po strukturi kot po vsebini, delujeta na podlagi statične predstavitve svetovnega spleta; to pomeni, da strani in povezave obstajajo v določenem trenutku. Seveda se vsebina in povezave nenehno spreminjajo, zato so lahko ti vpogledi vsakič ažurirani in na ta način bolj aktualni za uporabnike (Linoff, Berry, 2001, str. 22-24).

Najboljša predstavitev uporabe rudarjenja po strukturi je graf (angl. directed graph). Idealni graf bi nam narisal vse povezave, ki povezujejo dokumente na celotnem svetovnem spletu. Najboljša predstavitev uporabe rudarjenja po vsebini pa je indeks. Popolni indeks bi povezal vsako besedo, melodijo, frazo, sliko ipd. na svetovnem spletu z vsemi stranmi, ki jih vsebujejo. V svoji izvorni obliki, rudarjenje po vsebini ne potrebuje znanja o povezavi med samimi dokumenti, kot tudi proces rudarjenja po strukturi ne rabi vedeti, kaj dokumenti vsebujejo.

Tako vidimo, da ne rudarjenje po strukturi in ne rudarjenje po vsebini, ne zahtevata ali nas ne preskrbujeta z znanjem o uporabnikovem obnašanju. Rudarjenje po strukturi nam npr. pokaže, katere strani so z določene strani dostopne v dveh korakih, ne pa koliko ljudi je dejansko to tudi naredilo. Rudarjenje po vsebini nam sicer prikaže vsebino neke strani, ne pa tudi, kdo vse jo je tudi pregledal. Lahko nam samo pove, koliko strani nam najde povezavo z določeno temo. Rudarjenje po strukturi pa nam lahko te strani združi v gruče (angl. cluster). Če dejansko hočemo izvedeti, kdo obiskuje oziroma bere te strani in kako ti obiski vplivajo na kasnejši nakup, oziroma kako se kupci posameznih izdelkov med seboj razlikujejo, se moramo poslužiti postopka rudarjenja vzorcev uporabe. Ta je osredotočen predvsem na obnašanje uporabnikov, še posebej skozi določeno časovno obdobje. Včasih je časovni okvir uporabnikovega interesa dokaj kratek, recimo v obdobju ene seje (angl. session). Po drugi strani pa lahko uporabnike opazujemo skozi daljše časovno obdobje, recimo če analiziramo ponavljajoče se nakupe registriranih uporabnikov na določeni strani neke e-trgovine.

Kaj bi torej predstavljalo idealno predstavitev rudarjenja vzorcev uporabe? To bi lahko bila zbirka (knjižnica) uporabniških profilov z nenehno ažuriranimi profili za vsakega uporabnika spleta. Vsak profil bi povzel zgodovino posameznikovega udejstvovanja na spletu, med drugim tudi strani, ki jih je obiskal, poti in poizvedb, ki jih je napravil, dokumente, ki jih je prebral in pa izdelke, ki jih je kupil.

Na žalost pa je to zelo zapleten postopek. Precej lažje je določiti idealen indeks za rudarjenje po vsebini ali pa izdelati idealen graf za rudarjenje po strukturi. Informacije za izgradnjo indeksa in grafa so na voljo vsakomur, ki ima dostop do spleta. Informacije, ki jih potrebujemo za izgradnjo idealnega uporabniškega profila, pa so raztresene po raznih spletnih prijavah, prijavah na uporabniških strežnikih (angl. application server logs), oglasnih (angl. ad server logs) in trgovinskih strežnikih (angl. commerce server logs), bazah podatkov o izdelkih ter strankah. Te pa so v lasti različnih organizacij in mnoge od njih nimajo ne možnosti in ne interesa, da bi takšne informacije s komerkoli delile. Rezultat tega je, da je rudarjenje vzorcev uporabe ponavadi omejeno na modeliranje obnašanja uporabnikov določene spletne strani oziroma malo večjega omrežja strani. Lahko pa izberemo tudi manjši vzorec uporabnikov, ki so se javili in beležimo njihovo gibanje med večimi nepovezanimi stranmi.

Glede na to, da je rudarjenje po strukturi z vidika primera uporabe iz zadnjega poglavja manj pomembno, bo v nadaljevanju večja pozornost namenjena predvsem rudarjenju vzorcev uporabe.

3.3.2 RUDARJENJE PO STRUKTURI

S spletnim rudarjenjem po strukturi lahko ugotavljamo tako priljubljenost neke spletne strani, kot tudi »razdaljo« do ostalih strani. Več kot je povezav, ki vodijo do določene strani, bolj pomembna je ta stran. V matematičnem jeziku je svetovni splet usmerjen graf. Vsaka stran je vozlišče (angl. node) na tem grafu in vsaka povezava je rob (angl. edge) (Linoff, Berry, 2001, str. 24-34).

Dejstvo, da so nekatere strani bolj pomembne kot ostale, četudi so posvečene istim temam, je nam dokaj enostavno razumeti, vendar pa je to precej težje razložiti računalniku. Kadar uporabimo rudarjenje po vsebini za iskanje tem, o katerih je veliko napisanega, je v tako ogromni zbirki težko izločiti nam najbolj zanimive dokumente.

John Kleinberg iz Cornell University je odkril tehniko za rešitev tega problema. Njegov pristop se nanaša na to, da z ustvarjanjem povezav med stranmi, spletni skrbnik (angl. human webmaster) nato določi vrednost strani v povezavi; vsaka povezava do naslednje strani je neke vrste priporočilo za to stran. Ko vse to

seštejemo, ta neodvisna ovrednotenja različnih oblikovalcev spletnih strani, ki so večkrat določili povezave na neko določeno tarčo (target), potem lahko to tarčo označimo kot avtoriteto (angl. authority), oziroma kot bolj pomembno stran. Priporočila strani, ki že imajo mnogo dobrih priporočil, imajo mnogo večjo težo v določanju avtoritet, kot pa ostala priporočila.

3.3.3 RUDARJENJE VZORCEV UPORABE

Rudarjenje vzorcev uporabe je proces uporabe tehnik rudarjenja podatkov za odkrivanje vzorcev uporabe iz spletnih podatkov. V splošnem se odkrivanje znanja iz uporabe spleta sestoji iz treh korakov: priprave podatkov (čiščenje, obdelava manjkajočih vrednosti, pretvorba), odkrivanja vzorcev in analize dobljenih vzorcev (Srivastava, 2000, str. 14).

Dosti bolj kot sama struktura in vsebina spletnih strani, je za nas pomembno obnašanje samih uporabnikov. Čeprav se vsebina in struktura spletnih strani nenehno spreminjata, predstavljata za potrebe spletnega rudarjenja bolj statičen vidik. Po drugi strani pa se uporabnikovo vedenje opazuje skozi čas. Časovni okvir je lahko zelo različen; od trajanja ene seje pa do opazovanja, ki traja celo več let. Vedno pa je prisotna ideja, da se vzorci oblikujejo skozi določeno časovno obdobje (Linoff, Berry, 2001, str. 34-42).

Uporabniške vzorce lahko zaznamo in rudarimo na večih nivojih; lahko kot posledico klikov posamezne seje enega uporabnika ali tudi kot iskanje vzorcev celotnega razreda strank v obdobju več mesecev ali celo let. Ponavadi so podatki, ki jih zajamemo v nekem obdobju, urejeni v profile, ki določajo tekoči moment stranke. Ti profili se lahko uporabljajo za storitve priporočanja in personalizacije.

Rudarjenje vzorcev uporabe je lahko tako uporabno za izboljšanje oblike spletne strani, kot tudi za izboljšanje CRM – ravnanja odnosov s strankami (angl. customer relationship management).

ZBIRANJE PODATKOV

Najnižji nivo podatkov, ki jih lahko uporabimo za spletno rudarjenje vzorcev uporabe (angl. web usage mining), je potek povezav (angl. clickstream), tj. serije zahtev za strani, ki so jih prejeli različni spletni strežniki, ki gostijo te strani. Tako je na najnižjem nivoju zabeležena vsaka zahteva za gif, jpeg in html datoteke, ki jih zahteva uporabnikov brskalnik. Ti zadetki (angl. hits) se nato agregirajo v pregledih strani (angl. page wievs), ki se združujejo v serije, še preden se lotimo kakršnihkoli resnih analiz (tako se npr. ne moremo lotiti resnejših analiz obnašanja

posameznih strank v neki veleblagovnici, dokler ne najdemo načina, kako bomo posamezne seje povezali z obiskovalci, ki jih lahko identificiramo).

S premikanjem navzgor po hierarhiji, od spletnih pogledov do posameznih obiskovalcev, najdemo sicer vedno manj podatkov, vendar pa so ti vse bolj zanimivi. Najbolj zanimive analize najdemo na nivoju individualnih obiskovalcev, ko zasledujemo obnašanje uporabnikov skozi daljše časovno obdobje, ali pa zasledujemo obnašane uporabnikov skozi nivo sej na primeru posameznega obiska. Vsak posamezni uporabnik lahko napravi več sej. Vsaka seja lahko vsebuje več spletnih pogledov in vsak spletni pogled je lahko zabeležen tolikokrat, kolikor je zadetkov v spletnih prijavah (angl. web logs).

PRIPRAVA PODATKOV

Analiza poteka povezav se prične v spletnih prijavah. Potek povezav je definiran kot serija datotek, ki jih eksplicitno zahteva obiskovalec spletne strani s klikom na povezave. Čeprav se zdi, da je spletna stran sestavljena iz večih strani, so tisto, kar nam beležijo spletne prijave, posamezne zahteve spletnega brskalnika.

Preden se spletne prijave uporabijo za kakršnokoli učenje na podlagi obnašanja uporabnikov, moramo napraviti tudi še samo čiščenje in pregled le-teh. To lahko napravimo z aktivnostmi, kot so:

- filtriranje,
- odstranjevanje pajkovih sledi,
- prepoznavanje uporabnika,
- oblikovanje sej in
- zaključevanje poti.

FILTRIRANJE

Količina podatkov, ki jih dobimo s prijavami na spletni server (angl. web server logs), je ponavadi ogromna. Ko se enkrat takšni »surovi« podatki zbrani, je eden prvih korakov, ki ga moramo narediti za pripravo analize, da »prefiltriramo« nezaželene zabeležke (angl. records). Mnogo zadetkov, ki so zabeleženi (angl. recorded) v prijavah, predstavljajo zahteve za grafične prikaze, ki pa so, iz našega vidika, samo del izvorne HTML strani. Tako je potrebno odstraniti vse vhodne prijave (angl. log entries), ki predstavljajo zahteve po datotekah, katere se končajo z končnicami kot so gif, jpeg, png in ostale, in katere se uporabljajo za identifikacijo slikovnih datotek. Takšno filtriranje bi nam pripravilo podatke, ki bi bili na nivoju posameznih spletnih ogledov. Količina tako prefiltriranih podatkov, bi običajno obsegala desetino vseh surovih podatkov.

ODSTRANJEVANJE PAJKOVIH SLEDI

Iskalniki in drugi uporabniki vsebinskega rudarjenja so močno odvisni od tega, da imajo ažurirane indekse za vsebino na svetovnem spletu. Ti indeksi so narejeni s pajki, ki se plazijo (angl. crawl) po spletu. Pajki težijo k drugačnemu obnašanju kot »človeški« obiskovalci. Najprej opravijo obsežno iskanje na neki strani, preiščejo vsako povezavo od domače strani in nato še vsako povezavo od teh strani. Z uporabo pajka tako »onesnažimo« analizo »tipične« poti, saj je njegovo obnašanje prisotno v samih prijavah (angl. logs).

Pajek je program, ki samodejno išče in prenaša spletne strani. Uporabljajo jih predvsem iskalniki, ki z njihovo pomočjo preiskujejo spletne strani in dodajajo nove v svoje podatkovne zbirke, po katerih lahko uporabniki iščejo želene podatke. Pajki v spletnih straneh samodejno poiščejo povezave na druge strani in jim sledijo, zaradi česar lahko poiščejo veliko število spletnih strani (Islovar, 2005).

Odstranjevanje pajkovih sledi lahko naredimo s prepoznavo imena pajka v polju Agent v zbirki dostopov (angl. server log). Če pa so pregledi strani že združeni v seje, lahko odstranjevanje pajkovih sledi opravimo s prepoznavanjem vzorcev dostopov, ki so značilni za pajke.

IDENTIFIKACIJA UPORABNIKA

Preden lahko pričnemo s spletnim rudarjenjem vzorcev uporabe, moramo priti do podatkov obnašanja na stopnji sej uporabnikov. Vendar moramo, še preden lahko prepoznamo seje, prepoznati same uporabnike. Tu se pojavita predvsem dva problema. Prvič, kako prepoznamo zahteve posameznega uporabnika po določeni strani (angl. page request) v enem samem obisku, z namenom, da potem ustvarimo seje. In drugič, kako lahko prepoznamo uporabnika skozi večkratne obiske določene strani, da potem lahko analiziramo njegovo obnašanje v daljšem časovnem obdobju (dnevih, mesecih, letih).

Najboljša rešitev bi bila, da bi se uporabniki z uporabo uporabniškega imena in gesla, identificirali že ob samem vstopu na spletno stran. Vendar pa je danes večina obiskov spletnih strani opravljena anonimno, kar nas sili v uporabno raznih »zvijač«, da bi lahko odkrili, katere zahteve po spletni strani so bile zahtevane od istega uporabnika. Za vsako zahtevo strani, datoteka prijav (angl. log file) zabeleži IP naslov (angl. IP adress) tistega, ki je stran zahteval in kateri brskalnik je bil pri tem uporabljen. Žal to običajno ni dovolj za identifikacijo uporabnikov, ker je lahko več uporabnikov nekega podjetja logiranih z istim IP naslovom proksi (pooblaščenega) strežnika. V takem primeru obstaja več načinov, kako prepoznati različne uporabnike. Zahteve strani z istim IP naslovom, vendar pa z različnim brskalnim programom, verjetno pripadajo različnim uporabnikom. Zahteve strani,

ki se prikažejo kot strani, ki med seboj niso povezane z nikakršnimi hiperpovezavami, verjetno prav tako pripadajo različnim uporabnikom.

Kot vidimo, je že težko prepoznati istega uporabnika med večimi zahtevami strani v okviru ene same seje. Potem je še toliko težje preučevati uporabnikovo vedenje v daljšem časovnem obdobju, ker moramo pri tem zopet prepoznati istega uporabnika, ko se vrne nazaj po obdobju tedna ali meseca. Edini zanesljiv sistem bi bila uporaba registracije uporabnikov, vendar je to močno odvisno od pripravljenosti uporabnikov na kaj takega.

Danes je eden najbolj uveljavljenih sistemov za prepoznavanje vračajočih se uporabnikov, uporaba tako imenovanih piškotkov (angl. cookies). Piškotki so digitalni paketi, prilepljeni na elektronske dokumente, za pošiljanje po internetu. Izdani so s strani spletnega strežnika in shranjeni na spletnem brskalniku. Samo spletnemu strežniku, ki je izdal piškotek, je dovoljeno, da ga lahko prebere. Piškotki lahko vsebujejo karkoli si oblikovalec spletnih strani zamisli. Ko obiskovalec pride na neko stran, bo strežnik zahteval piškotek, ki je bil predhodno »podtaknjen« v prisotnosti brskalnika. Če je obiskovalec že bil na tej strani in če ima omogočene piškotke ter uporablja isti brskalnik na istem računalniku kot prejšnjič, potem ga bomo dejansko prepoznali.

Kot vidimo, je za dosego tega cilja potrebno izpolniti precej pogojev. Uporabniki mnogokrat menjajo brskalnike ali dostopajo do istih strani z različnih računalnikov. To ne pomeni, da so piškotki v celoti neuporabni, a so dosti bolj, če imamo opravka z znanimi registriranimi uporabniki. Ta sistem lahko naši strani omogoča, da prepoznamo vračajoče se uporabnike in jim na ta način lahko postrežemo s prilagojeno obliko spletne strani ali jim lahko celo predlagamo kakšen nasvet.

OBLIKOVANJE SEJ

Oblikovanje sej je proces ugotavljanja, katere serije spletnih pogledov so bile zahtevane s strani iste osebe med enim samim obiskom. Seje so zelo pomembne, kajti ustrezajo pojmu opazovanja uporabnikov določene spletne strani. Lahko rečemo, da pojem seje enačimo s pojmom obiska.

ZAKLJUČEVANJE POTI

Eden izmed problemov, ki zadeva oblikovanje sej ali kakršnekoli analize poti (angl. path analysis), ki temelji na spletnih prijavah, je v tem, da mnogo zahtev strani sploh nikoli ni zabeleženih v zbirki dostopov (angl. server log). Glavni razlog za to izgubo je tako imenovano predpomnjenje (angl. caching), ki se pojavlja na več nivojih. Sam brskalnik shranjuje predpomnjenje zadnje obiskanih strani. To predpomnjenje nato obstaja za neko uporabniku ustrezno dolžino časa, običajno merjeno v dnevih. Če uporabnik zahteva stran, ki je v predpomnilniku, potem se

brskalnik ne ubada z zahtevo do strežnika in tako ne dobimo nobene zaznave na samem strežniku. V precej primerih imamo še dodatno stopnjo predpomnjenja, ki je zagotovljena s proksi strežnikom. Zaključevanje poti je proces zapolnitve manjkajočih korakov, ki nastanejo zaradi predpomnjenja.

UPORABNIŠKE PRIJAVE

Uporaba metod, kot so filtriranje, odstranjevanje pajkovih sledi, oblikovanje sej, prepoznava uporabnika in zaključevanje poti, nam pripravijo seje prijav (angl. session logs), ki nam opisujejo obnašanje obiskovalcev v smislu zahtev URL. Ne povedo pa nam dosti o tem, kaj dejansko je uporabnik počel ali katere izdelke je kupil oziroma si jih je ogledoval, a je nakup nato zavrnil. Da bi lahko popolnoma razumeli, kaj dejansko uporabniki počnejo, je potrebno, da podatke o poteku povezav dopolnimo s podatki iz uporabniških (aplikacijskih) strežnikov.

V moderni e-tržni arhitekturi vedenje o tem, kateri URL-ji so bili zahtevani, dejansko ne pove kaj dosti, kajti sama vsebina strani je sestavljena na uporabniškem strežniku. Isti URL lahko tako predstavlja popolnoma različno vsebino ob različnem času. S pomočjo uporabniškega strežnika pa lahko izvemo, katera vsebina je bila kdaj zahtevana in kaj nam predstavlja. Torej so uporabniške prijave tiste, ki natančno vedo, kateri izdelki se nahajajo v prodajni ponudbi, oziroma kdaj so bili odstranjeni. To vedo na podlagi tega, kdaj se je uporabnik prijavil in kdaj se je nato odjavil. Obnašanje strank lahko na ta način merimo na nivoju, ki je zanimiv za same trgovce in prodajalce, ne pa samo za oblikovalce spletnih strani. Podatki s spletnih kanalov se dejansko dopolnjujejo s podatki drugih kanalov, kot so prodajalne, klicni centri in katalogi.

RUDARJENJE VZORCEV UPORABE ZA IZBOLJŠANJE UPORABNOSTI SPLETNIH STRANI

Eden najpomembnejših razlogov za uporabo izkopavanja spletne uporabe, je zagotovo izboljšanje uporabnosti neke strani. Prvi korak pri takšni analizi je zbiranje opazovanih poti, ki jih je uporabnik napravil. Vsaka seja je posledica zahteve po strani. Tako kot so izdelki, ki jih je kupil določen kupec ali pa so bili kupljeni ob enkratnem obisku strani, združeni v nakupovalni košarici, lahko na enak način sestavimo tudi zbirke teh strani. Te zbirke lahko analiziramo s pomočjo asociacijskih pravil, da odkrijemo, katere strani gredo skupaj. Te asociacije nam lahko predlagajo še dodatne povezave. Če npr. najdemo dve strani večkrat skupaj v različnih sejah, vendar pa med njima do sedaj ni obstajala nobena povezava, potem se bo verjetno zdelo uporabniku bolj funkcionalno, če to povezavo tudi naredimo. V analizi nakupovalne košarice ponavadi nismo pozorni na to, v kakšnem vrstnem redu so bili izdelki dani v voziček. Vrstni red ogledanih strani pri spletni seji pa je dosti bolj pomemben, saj seje ponavadi analiziramo kot posledice

(ogled ene strani je posledica ogleda predhodne strani) (Linoff, Berry, 2001, str. 42).

Seje so lahko grupirane na različne načine, tako da sorodne seje formirajo gručo (angl. cluster). Te gruče nato predstavljajo različne vrste uporabnikov: izkušeni uporabniki nasproti novim uporabnikom, kupci nasproti tistim, ki so samo obiskovalci naših strani itd. Različne skupine obiskovalcev pridejo na določeno spletno stran z različnimi nameni. Spletno stran neke velike prodajalne lahko obiščemo z namenom (Linoff, Berry, 2001, str. 42, 43):

- kupiti nekaj,
- dobiti informacije o najbližji »fizični« prodajalni,
- pozanimati se o možnostih zaposlitve,
- dobiti informacije o naložbah,
- preveriti število točk lojalnosti,
- pozanimati se, kako vrniti izdelek s katerim nismo bili zadovoljni,
- izvedeti naslov in telefonsko številko vodstva podjetja.

Rudarjenje vzorcev uporabe nam pomaga pri odkrivanju različnih načinov uporabe neke strani in nam predlaga načine, kako izboljšati to stran.

Spletno rudarjenje vzorcev uporabe je vrsta spletnega rudarjenja, ki vsebuje avtomatično odkrivanje uporabnikovih vzorcev dostopa na enem ali več spletnih strežnikov. Glavni namen tega je, da bolje razumemo reakcije strank, ki kupujejo na spletni strani podjetja ali pa zgolj uporabnike, ki »deskajo« po spletni strani podjetja. Nekatere študije pravijo, da lahko rezultate spletnega rudarjenja uporabimo za izboljšanje dizajna spletne strani, sposobnosti sistemske analize in mrežne komunikacije ali pa celo izgradnje prilagodljivih spletnih strani. Na splošno lahko rečemo, da imamo pri uporabi odkritega znanja na področju rudarjenja vzorcev uporabe dva glavna cilja (Fong, 2002, str. 39-62):

- splošno sledenje vzorcev dostopa za njihovo razumevanje in razumevanje trendov ter
- prirejeno sledenje uporabe (angl. customized usage tracking) za prilagajanje in personalizacijo brskalnih izkušenj uporabnikov.

3.3.4 RUDARJENJE PO SPLETNIH VSEBINAH

Rudarjenje po spletnih vsebinah (angl. web content mining) je samodejno odkrivanje vsebinskih vzorcev s spletnih dokumentov (Srivastava, 2000, str. 14).

Brez spletnih iskalnikov si danes težko predstavljamo uporabo svetovnega spleta. Kar se neposredno tiče samih uporabnikov, je zanje zagotovo najbolj pomembna vsebina na svetovnem spletu. Iskanje prave vsebine je s pomočjo spletnih

iskalnikov precej bolj enostavno. Naloge, ki jih opravljajo spletni iskalniki, so zelo uporaben primer rudarjenja po vsebini. Temu, kar počnejo iskalniki, pravimo tudi pridobivanje informacij (angl. information retrieval) (Linoff, Berry, 2001, str. 43-54).

Svetovni splet vsebuje ogromno število informacij, dezinformacij in popolnoma neuporabnih informacij (angl. junk). Za iskanje zahtevane informacije, je to pri večini tem lahko zelo zahtevno opravilo (lahko rečemo, da je svetovni splet dokaj nečista ruda). Če bi bili vsi dokumenti na spletu jasno markirani (označeni) s ključnimi besedami in če bi bili vsi uporabniki vešči iskanja (v slogu knjižničarjev), potem za iskanje informacij ne bi potrebovali zapletenih algoritmov, ampak bi do njih prišli z dosti bolj enostavnimi poizvedbami (angl. queries).

Danes v svetovnem spletu najdemo v glavnem HTML in XML dokumente, vendar pa je pričakovati, da se bo povečal delež strukturiranih meta podatkov. HTML je standard za opis tega, kako naj bi bili dokumenti prikazani; XML pa je raztegljiv standard, ki omogoča skupini uporabnikov, da se dogovorijo za določeno uporabniško zbirko oznak, ki nosijo informacijo o tem, kaj določen dokument pomeni. Trenutno še ne obstaja prav veliko takšnih meta podatkov za dokumente; potrebno jih je nekako sklepati iz same vsebine. Izziv podatkovnega rudarjenja za pridobitev informacij je prav gotovo v kreiranju meta podatkov, ki nam bodo omogočili preproste zahteve, kot so: »najdi mi informacije o alternativnem zdravljenju visokega krvnega pritiska« ali pa »najdi mi še več strani na to temo«. Lahko rečemo, da je danes (zaenkrat) še večina rudarjenja po spletnih vsebinah pravzaprav tekstovno rudarjenje.

Glavni cilj pridobitve informacij je v tem, da najdemo le tiste informacije, ki jih iščemo. Pri tem raziskovalci uporabljajo dva kriterija, ki merita učinkovitost iskanja: odzivnost (angl. recall) in natančnost (angl. precision). Natančnost odgovori na vprašanje: »Od strani, ki smo jih dobili, kolikšen je delež takšnih z iskano vsebino?«, medtem ko odzivnost odgovori na vprašanje: »Od vseh strani z iskano vsebino, kakšen je delež tistih, ki smo jih našli?«. Ta dva cilja se na nek način izključujeta. Iskalnik, ki nam bo izpisal prav vsako stran, povezano z določeno temo, bo gotovo imel popolno odzivnost, vendar pa zelo slabo natančnost. Če pa bi nam poiskal samo eno stran na pravo temo, bi imeli zelo dobro natančnost, a tudi zelo slabo odzivnost.

Kaj je torej bolj pomembno: dobra odzivnost ali dobra natančnost? Odgovor se skriva v sami naravi poizvedbe. Na nekatera vprašanja je zelo lahko odgovoriti samo z ogledom ene strani, medtem ko je nekje potrebno več ogledov. Spletni iskalniki stremijo k temu, da skušajo zagotavljati oboje, tako dobro odzivnost kot natančnost: Tako eno kot drugo pa je odvisno od same klasifikacije strani po temah, kar predstavlja velik izziv za podatkovno rudarjenje.

3.4 UPORABA SPLETNEGA RUDARJENJA V POSLOVNEM SVETU

Zabeleženi podatki (to kar neko podjetje ve o svojih strankah, o svojih trgih, zaposlenih in do česar nimajo dostopa njihovi konkurenti) so lahko v poslovnem svetu pomembna prednost. Danes vidimo predvsem tri področja, kjer ima svetovni splet pomemben vpliv na poslovanje. Ta področja definiramo glede na način, kako se s poslovanjem služi denar. Konec koncev je najpomembnejši faktor uspeha prejemek, ki ga dobimo kot plačilo naših strank za naše storitve. Ta tri področja so (Linoff, Berry, 2001, str. 3):

- e-trgovina,
- e-medij in
- e-trg.

Na koncu tega poglavja bom nekaj povedal tudi o vse pomembnejši vlogi spletnega rudarjenja v trženju.

3.4.1 E–TRGOVINA

Pojem e-trgovine razumemo kot prodajo izdelkov preko svetovnega spleta, kjer kupci plačajo direktno podjetjem, ki se ukvarjajo z e-trgovino. Kaj je prednost takega poslovanja v primerjavi z navadnim »staromodnim« poslovanjem? Prednost je predvsem v tem, da podjetja e–trgovine lahko veliko bolj izkoristijo uporabo spletnega rudarjenja in s tem pridobijo koristne informacije glede na podatke, ki so se zbrali med poslovanjem. Z uporabo teh podatkov, lahko npr. (Linoff, Berry, 2001, str. 3, 4):

- prenovijo spletne strani, glede na zaznavo uporabnikovih preferenc in njihovega obnašanja,
- zaznajo uporabnikove nakupe in jim predlagajo nove še v istem nakupnem procesu,
- si zapomnijo uporabnikove preference iz preteklih obiskov in to uporabijo pri njegovem naslednjem obisku ali
- se osredotočijo tako na uporabnikove preference, kot na ažurno kontroliranje zalog.

Predstavljajte si, da bi imeli strokovnjaka za nakupovanje, ki bi se posvetil vsaki stranki, ji predlagal nove proizvode, medtem ko stranka nakupuje ali samo že kaže interes, da si nekaj želi. Ali pa, da se stranki, vsakič ko stopi v trgovino, le-ta samodejno prilagodi in se ji prilagaja tudi takrat, ko se sprehaja po trgovini. Recimo, da se na polici z izdelki, za vsakim izdelkom, ki ga bo kupec kupil, nahaja tudi nov izdelek, ki naj bi ga kupec kupil z malce prepričevanja. Ta revolucionarna ideja prinaša koristi tako kupcu (z zbiranjem na kup dobrin za katere naj bi se

zanimal), kot prodajalcu (z večanjem obsega prodaje). Tako nam svetovni splet kot trgovski kanal, omogoča nove možnosti poslovanja.

3.4.2 E–MEDIJ

Za razliko od običajnih medijev, e-mediji podjetja dejansko vedo, katera vsebina je najbolj zanimiva za bralce njihovih strani. Prav tako tudi vedo, kateri bralci se zanimajo za katere vsebine. Opazna je tudi razlika na strani oglaševanja, saj lahko zasledujemo bralce, ki berejo določeno vsebino na več različnih straneh in na podlagi tega lahko ugotovimo, katere oglase so lahko opazili. Tako lahko precej bolj natančno ugotavljamo odzivnost na določene oglase. To nam omogoča ogromne možnosti analiziranja učinkovitosti oglaševanja (Linoff, Berry, 2001, str. 5).

3.4.3 E–TRG

E-trg razumemo kot vmesnega posrednika, ki poveže kupce in prodajalce ter nato zahteva plačilo za opravljeno transakcijo. Z uporabo svetovnega spleta, lahko podjetja spletno rudarjenje in analizo uporabijo za ugotavljanje, kateri kupci se ujemajo s katerimi prodajalci (Linoff, Berry, 2001, str. 6).

E–trgovina in e–medij sta zanimiva predvsem zato, ker vplivata na življenje samega potrošnika.

Spletno rudarjenje ne more spremeniti slabo poslovanje v dobro poslovanje. Lahko nam samo omogoči, da dobro poslovanje spremenimo v še boljše poslovanje in da se to osredotoči na bolj pomembne zadeve, na njegove stranke.

3.4.4 UPORABA SPLETNEGA RUDARJENJA V TRŽENJU

Ne gre pa tudi zanemariti vse pomembnejše vloge spletnega rudarjenja v trženju. Spletno rudarjenje je postalo precej popularno orodje, še posebej po uvedbi interneta kot novega distribucijskega kanala dobrin, promocije, oglaševanja, opravljanja transakcij in koordiniranja poslovnih procesov. Splet je tako postal pomemben in primeren vir podatkov o kupcih. Vendar so množični formati podatkov in lastnosti znanja na spletu dosegli, da je postalo zanimivo in koristno zbiranje, urejanje in organiziranje znanja na spletu predvsem za namene trženja in kot podpora pri odločanju. Trženje postaja, zaradi podatkov o kupcih, vse bolj odvisno od spleta in tako postaja tudi spletno rudarjenje vse bolj trženjska funkcija (Hrastar, Krnc, Škoberne, 2003, str. 13-15).

Odločitve v trženjskih akcijah pri promocijah, distribucijskih kanalih, oglaševalskih medijih in podobno, ki so izpeljane na »tradicionalen« način, prinašajo kot rezultat slab odziv in dokaj visoke stroške. Danes imajo kupci zelo različne okuse, zahteve in potrebe, tako da ni mogoče vseh umestiti v eno homogeno skupino oziroma ciljno populacijo za razvijanje trženjskih strategij. V bistvu bi lahko rekli, da je danes vse bolj pomembna individualizacija trženjskih storitev, saj hoče vsak kupec »biti postrežen« glede na svoje lastne potrebe. Trženje s pomočjo podatkov (angl. database marketing) nam nudi možnost uporabljanja podatkov iz baze transakcij in baze kupcev. Tako so si mnoga podjetja ustvarila velike baze podatkov o svojih kupcih in njihovih nakupnih transakcijah.

Z ustvarjanjem profila kupcev je tako mogoče ugotoviti, kakšne so nakupne navade potrošnikov. Ustvarjanje profila potrošnikov je dokaj pogosto uporabljena metoda za različno pomembne trženjske odločitve. Profil kupca je model potrošnika, na osnovi katerega lahko tržnik v podjetju izpelje pravilne odločitve glede strategije in taktike zato, da bi čimbolj spoznal kupčeve želje in potrebe ter jih tako tudi čimbolj zadovoljil. Na osnovi profila kupca lahko izvemo:

- pogostost (frekvenco) nakupa,
- obseg nakupa,
- koliko časa je preteklo od zadnjega nakupa,
- iskanje tipičnega potrošnika v skupini,
- kupčeve vrednote.

V praksi se spletno rudarjenje in njegova uporaba kažeta v množični uporabi spletnih brskalnikov in deskanja po spletu. Tako se lahko preko datotek dostopov (angl. log file) ugotovi, s katere strani je prišel obiskovalec na našo spletno stran. To sledenje poteku povezav je uporabno predvsem pri trženjskih akcijah, saj lahko tako naročniki hitro izvedo, s katere predhodne strani prihajajo najboljši kupci in s katere najslabši. Na primer podjetje Q, ki ima e-trgovino, ima povezave na svojo stran iz strani X in Y. Dokaj hitro lahko izvedo, koliko odstotkov obiskovalcev, ki prihajajo s strani X, bo kupilo v določenem znesku v njihovi spletni trgovini. Tako se potem lahko odločijo za usmerjeno in bolj učinkovito trženje.

Velika prednost spletnega trženja je v tem, da lahko merimo obisk bolj kot pri katerikoli drugi tehniki direktnega trženja. Rudarjenje po podatkih deluje najbolje, če imamo jasne, merljive cilje, npr. (Greening, 2000):

- povečati povprečni ogled strani,
- povečati dobiček na ogled,
- zmanjšati število zavrnjenih proizvodov,
- povečati število strank preko referenc,
- povečati zavedanja znamk,

- povečati stopnjo uporabe izdelka,
- zmanjšati zapiranje programa (angl. clicks-to-close).

3.5 VPETOST UPORABNIKA PRI PODATKOVNEM RUDARJENJU

Odkrivanje znanja je mogoče z današnjimi orodji za podatkovno rudarjenje izvesti sorazmerno samodejno. Predpogoj je, da imamo na voljo podatke v ustrezni obliki: prečiščene, dosledne, z razrešenimi manjkajočimi vrednostmi. Uporabnik tako lahko uporabi ustrezno orodje v skladu s ciljem podatkovnega rudarjenja (npr. prihodnje gibanje delnic na borznem trgu...), ki ga zanima in metodo, s katero želi priti do tega cilja (časovno rudarjenje podatkov, klasifikacija, odkrivanje izjem ipd.). Ponavadi je potrebno vsaj minimalno sodelovanje uporabnika (Horvat, 2003, str. 75-79).

Po drugi strani pa je zaželeno čim večje sodelovanje uporabnika pri podatkovnem rudarjenju. Pri tem je izrednega pomena vizualni komunikacijski vmesnik. Uporabnik je lahko vključen v (skoraj) vsak proces pri rudarjenju podatkov, pri čemer imamo v mislih zdravo mero uporabnikove interaktivnosti v primerjavi s samodejno izvedbo algoritma. Uporabnik in računalnik naj bi delovala v navezi. Konec koncev mora strokovnjak (angl. domain-knowledge expert) dobljene rezultate pri rudarjenju podatkov preveriti in ovrednotiti.

Ponudnik spletne strani si želi, da bi jo uporabljalo čim več uporabnikov in da se le-ti nanjo navadijo ter jo čim pogosteje uporabljajo. V uvajalni fazi je zato pomembno, da ima tak ponudnik dovolj zanimivih storitev, ki pritegnejo vedno nove in nove uporabnike in ki glavnino le-teh obdržijo kot redne obiskovalce ter uporabnike. Začetna ponudba so tako lahko: športne novice, spremljanje dogajanja na borzi, vozni redi ipd. Ko ponudnik doseže želeno število uporabnikov, ščasoma širi svoje storitvene ponudbe in prične nekatere od njih tudi tržiti. Pri tem ni važno ali gre za lastne storitve ali pa jih samo posreduje. Da bo spletna družba preživela in da bo pri tem uspešna, mora vsakemu posameznemu uporabniku ponujati prave storitve zanj.

Osnovne podatke o uporabniku pridobimo tako, da ga nekako motiviramo, da se prijavi, tj. da se vključi med posebej priviligirane uporabnike. Pri tem je pomembno, da nam uspe pridobiti čim več za nas pomembnih podatkov o uporabniku. Ti podatki so za nas trajne narave. Predvsem bi nas lahko zanimal spol, izobrazba, starost, količina denarja, s katerim uporabnik razpolaga, katere jezike razume, kaj ga je pritegnilo, da je obiskal spletno stran in kako je izvedel zanj. Ob prijavi je dobro, če uporabnika izprašamo še o: interesih, načinu obveščanje o novostih itd. Pri tem pa seveda upamo, da so njegovi podatki verodostojni. Podatke lahko dopolnimo tudi iz drugih virov, če so nam na voljo.

Poleg statičnih podatkov nas pri uporabniku zanima predvsem transakcijski del oziroma njegovo obnašanje na spletni strani. Zanima nas: katere storitve uporablja, kakšna navigacija po strani mu ustreza ipd. Na tak način lahko zgradimo profil za vsakega vpisanega uporabnika spletne strani. Pri anonimnih obiskovalcih pa je to oteženo. V najboljšem primeru se lahko naslonimo na sprotno spremljanje njihovih dejavnosti (transakcijski del). Če se uporabnikovemu računalniku IP prireja dinamično, pa je tudi to oteženo.

Na podlagi zgrajenih profilov lahko uporabnikom zagotovimo personaliziran dostop. Poleg tega, da jim omogočimo svojevrstne nastavitve vmesnika, je bistveno to, da jih zalagamo s ponudbo, za katero menimo, da jih verjetno zanima. To je neke vrste neposredno trženje (angl. direct marketing)

Sedaj je vprašanje, kako naj se lotimo izdelave profilov. Lahko izdelamo profil iz statičnega in transakcijskega dela. Statični del profila dobimo tako, da upravljalec spletne strani pregleda vnesene podatke, jih potrdi in tako le-ti že lahko postanejo statični del profila. Pri določitvi transakcijskega dela profila pa uporabimo metode rudarjenja podatkov.

Ko imamo določene profile uporabnikov, lahko ob vstopu na stran, oziroma ob obisku, uporabnika obvestimo o (novih) ponudbah, ki bi ga utegnile zanimati. Ko se torej uporabnik odloči za uporabo neke storitve ali da v e-košarico nek artikel, je smotrno, da mu ponudimo storitve ali izdelke, za katere so se odločili tudi ostali (sorodni) uporabniki, čeprav mu po njegovem profilu sodeč, tega niti ne bi pripisali. To dosežemo npr. tako, da poiščemo asociacijska pravila med vsemi (sorodnimi) uporabniki. Seveda je tu spet potrebno pravilo prečistiti s pomočjo upravljalca. Če odkrijemo, da ta artikel/storitev implicira nakup nekih drugih artiklov (uporabo še kakšne storitve), potem uporabniku le-te na ustrezen način ponudimo. Morda pa je to mogoče odkriti že pri pregledu profilov ostalih uporabnikov.

Sedaj se pojavlja vprašanje: ali ne gre pri vsem tem za korak nazaj pri rudarjenju podatkov. V podatkovni analizi ločujemo rudarjenje podatkov od analize OLAP ravno po tem, da gre pri prvem za samodejno pridobivanje znanja, pri drugem pa uporabnik prek grafičnih predstavitev z lastno interpretacijo pride do novega znanja (tj. ugotovi nepravilnosti ali zanimivosti, ki prej niso bile znane, ali pa so določena predvidevanja potrjena).

Ne samo, da lahko strokovnjak z izgradnjo ustreznega modela znanja veliko prispeva k orodju za rudarjenje podatkov, temveč s tem, ko se poglobi v sam ustroj, tudi sam veliko pridobi od sistema. Vpletenost uporabnika v proces odkrivanja znanja torej veliko pripomore k razumevanju (angl. understandability).

Seveda pa uporabnika ne kaže obremenjevati s preveč podrobnostmi. V tem primeru se možnost napake zelo poveča, kar prednost takega pristopa izniči. Zato bi bilo takrat najbrž bolje vse delo prepustiti računalniškemu algoritmu in njegovi hevristiki.

4 PRIMER UPORABE

Za lažjo predstavitev postopka spletnega rudarjenja in možnosti, ki nam jih ponuja, bom postopek rudarjenja prikazal na praktičnem primeru. Podatke sem dobil na Ekonomski fakulteti v Ljubljani, in sicer v obliki prijav na računalnikih Centralne ekonomske knjižnice (CEK). Študentje uporabljajo te računalnike za več aktivnosti: lahko iščejo literaturo po knjižnici(ah), izdelujejo seminarske naloge, brskajo po spletu, ipd. Za začetek izvajanja aktivnosti iskanja, se morajo na računalnik prijaviti s svojo vpisno številko. To pa nam omogoča natančno spremljanje prijav in posledično pridobivanje informacij o tem, kateri študentje so ob določenem času uporabljali računalnike v CEK. Poleg teh podatkov, sem o študentih dobil tudi nekatere demografske in ostale podatke, s pomočjo katerih sem nato ugotavljal, od česa je najbolj odvisno število oziroma pogostost prijav na računalnikih CEK.

Teorijo rudarjenja podatkov, ki sem jo opisal v prejšnjih poglavjih, bom skušal povezati s praktičnim primerom. Razložil bom rezultate in ugotovitve, opisal težave, ki sem jih pri tem imel, ter razložil, kaj ni bilo v redu oziroma kako bi lahko celoten postopek izboljšal ali popravil.

Prva težava je nastala že pri »umeščanju« teorije v moj primer. Spletno rudarjenje je na križišču več znanstvenih področij. To je tudi razlog za mnoge nejasnosti, saj gre za novo vejo, kjer se prepletajo že obstoječe discipline. Tako smo priče pomanjkanju definicij in standardov ter celovite strukturiranosti področja. Sam sem imel težave že s tem, kako razdeliti postopek spletnega rudarjenja. Predvsem zaradi relativne mladosti tematike, so se razdelitve posameznih avtorjev, ki sem jih zasledil v literaturi, med seboj dokaj razlikovale. Postopek spletnega rudarjenja sem zato razdelil, glede na naš primer, na:

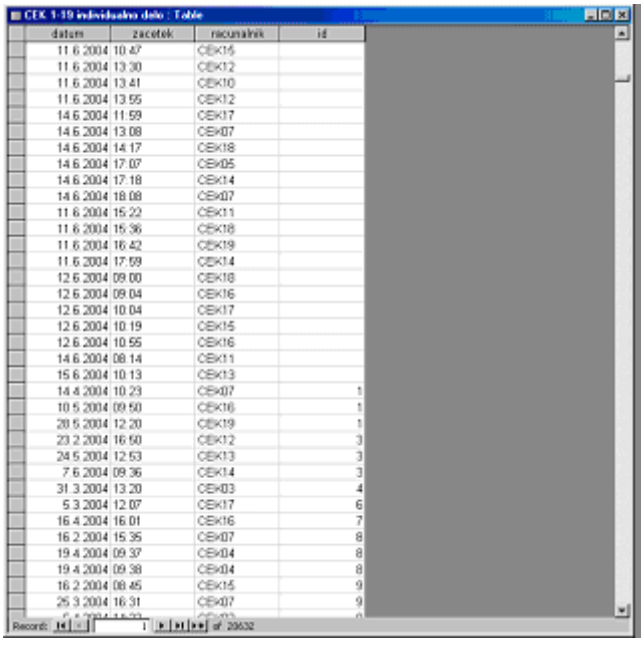
- zbiranje podatkov,
- opredelitev problema,
- pripravo podatkov,
- analizo podatkov in
- razlago ugotovitev.

4.1 ZBIRANJE PODATKOV

Kot sem že omenil, sem podatke za svoj primer dobil v knjižnici Ekonomske fakultete v Ljubljani. Na voljo so mi bili podatki o prijavah študentov na računalnikih v Centralni ekonomski knjižnici. Podatke sem dobil v obliki tabel v Access-ovi datoteki. Bazo podatkov sestavljajo štiri tabele: tabela *CEK 1-19 individualno delo*, tabela *CEK 20-46 dve ali več oseb + miza*, tabela *Dodatne osebe na računalnikih* ter tabela *Studenti*.

Prve tri tabele nam povedo, kateri uporabnik (v našem primeru je to študent) se je ob določenem dnevu in ob določeni uri prijavil na določen računalnik v knjižnici. Računalniki se med seboj razlikujejo po številkah (CEK številka), študentje pa se med seboj razlikujejo glede na id (identifikacijsko številko študenta). Vsak posamezen id predstavlja nekega študenta (glej Sliko 1).

Slika 1: Prijave študentov na računalnikih CEK: tabela CEK 1-19 individualno delo

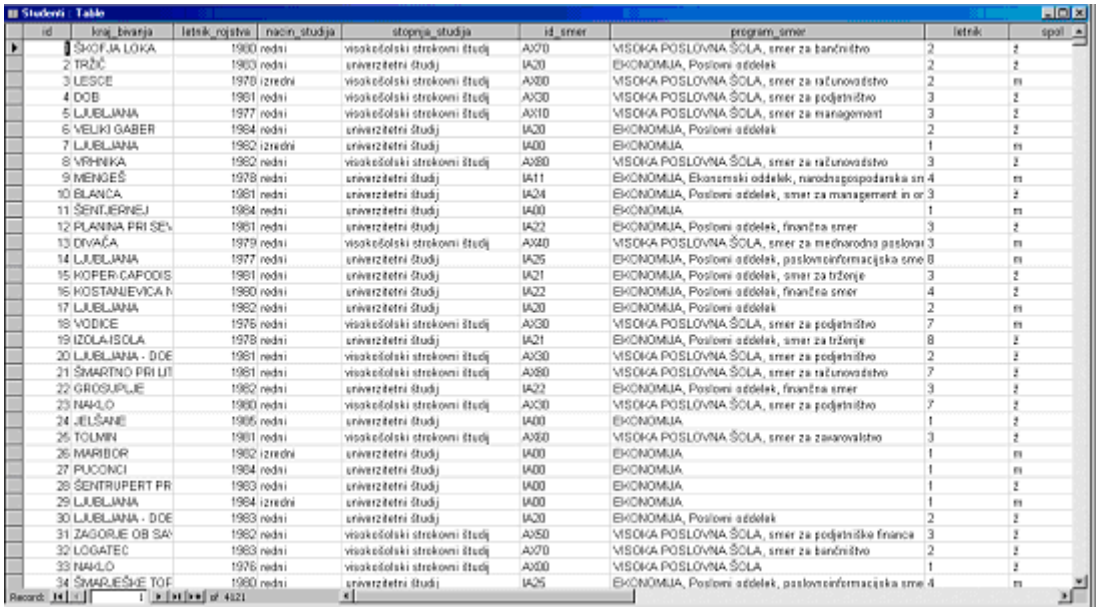


Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

Tabela *CEK 20-46 dve ali več oseb + miza* in pa tabela *Dodatne osebe na računalnikih* sta praktično enaki, kot je tabela *CEK 1-19 individualno delo*.

Tabela *Studenti* nam prikazuje vse študente, ki so se v opazovanem obdobju prijavili na računalnike v knjižnici. Za vsakega posameznega študenta je na voljo več podatkov: id (identifikacijska številka študenta), kraj bivanja, letnik rojstva, način študija (redni ali izredni), stopnja študija (univerzitetni ali visokošolski strokovni študij), id smer (identifikacijska številka smeri), program_smer, letnik in spol študenta (glej Sliko 2, na str. 26).

Slika 2: Tabela Studenti



id	kraj_bivanja	letnik_rojstva	nacin_studija	stopnja_studija	id_smer	program_smer	letnik	spol
1	ŠKOFJA LOKA	1980	redni	visokošolski strokovni študij	IA70	VISOKA POSLOVNA ŠOLA, smer za bančništvo	2	ž
2	TRŽIČ	1983	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni oddelek	2	ž
3	LESCE	1978	izredni	visokošolski strokovni študij	IA60	VISOKA POSLOVNA ŠOLA, smer za računovodstvo	2	m
4	DOB	1981	redni	visokošolski strokovni študij	IA30	VISOKA POSLOVNA ŠOLA, smer za podjetništvo	3	ž
5	LJUBLJANA	1977	redni	visokošolski strokovni študij	IA10	VISOKA POSLOVNA ŠOLA, smer za management	3	ž
6	VELIKI GABER	1984	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni oddelek	2	ž
7	LJUBLJANA	1982	izredni	univerzitetni študij	IA00	EKONOMIJA	1	m
8	VRHNIKA	1982	redni	visokošolski strokovni študij	IA60	VISOKA POSLOVNA ŠOLA, smer za računovodstvo	3	ž
9	MENGEŠ	1978	redni	univerzitetni študij	IA11	EKONOMIJA, Ekonomski oddelek, narodnogospodarska sm	4	m
10	BLANCA	1981	redni	univerzitetni študij	IA24	EKONOMIJA, Poslovni oddelek, smer za management in or	3	ž
11	ŠENTJERNEJ	1984	redni	univerzitetni študij	IA00	EKONOMIJA	1	m
12	PLANINA PRI SEV	1981	redni	univerzitetni študij	IA22	EKONOMIJA, Poslovni oddelek, finančna smer	3	ž
13	DIVAČA	1979	redni	visokošolski strokovni študij	IA40	VISOKA POSLOVNA ŠOLA, smer za mednarodno poslovan	3	m
14	LJUBLJANA	1977	redni	univerzitetni študij	IA25	EKONOMIJA, Poslovni oddelek, poslovnoinformacijska sme	8	m
15	KOPER-CAPODIS	1981	redni	univerzitetni študij	IA21	EKONOMIJA, Poslovni oddelek, smer za trženje	3	ž
16	KOSTANJEVICA N	1980	redni	univerzitetni študij	IA22	EKONOMIJA, Poslovni oddelek, finančna smer	4	ž
17	LJUBLJANA	1982	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni oddelek	2	m
18	VODICE	1976	redni	visokošolski strokovni študij	IA30	VISOKA POSLOVNA ŠOLA, smer za podjetništvo	7	m
19	IZOLA-ISOLA	1978	redni	univerzitetni študij	IA21	EKONOMIJA, Poslovni oddelek, smer za trženje	8	ž
20	LJUBLJANA - DOE	1981	redni	visokošolski strokovni študij	IA30	VISOKA POSLOVNA ŠOLA, smer za podjetništvo	2	ž
21	ŠMARTNO PRI LIT	1981	redni	visokošolski strokovni študij	IA60	VISOKA POSLOVNA ŠOLA, smer za računovodstvo	7	ž
22	GROSUPLJE	1982	redni	univerzitetni študij	IA22	EKONOMIJA, Poslovni oddelek, finančna smer	3	ž
23	NAKLO	1980	redni	visokošolski strokovni študij	IA30	VISOKA POSLOVNA ŠOLA, smer za podjetništvo	7	ž
24	JELŠANE	1985	redni	univerzitetni študij	IA00	EKONOMIJA	1	ž
25	TOLMIN	1981	redni	visokošolski strokovni študij	IA60	VISOKA POSLOVNA ŠOLA, smer za zavarovalstvo	3	ž
26	MARIBOR	1982	izredni	univerzitetni študij	IA00	EKONOMIJA	1	m
27	PUCONCI	1984	redni	univerzitetni študij	IA00	EKONOMIJA	1	m
28	ŠENTRUPERT PR	1983	redni	univerzitetni študij	IA00	EKONOMIJA	1	ž
29	LJUBLJANA	1984	izredni	univerzitetni študij	IA00	EKONOMIJA	1	m
30	LJUBLJANA - DOE	1983	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni oddelek	2	ž
31	ZAGORJE OB SA	1982	redni	visokošolski strokovni študij	IA50	VISOKA POSLOVNA ŠOLA, smer za podjetniške finance	3	ž
32	LOGATEC	1983	redni	visokošolski strokovni študij	IA70	VISOKA POSLOVNA ŠOLA, smer za bančništvo	2	ž
33	NAKLO	1976	redni	visokošolski strokovni študij	IA00	VISOKA POSLOVNA ŠOLA	1	ž
34	ŠMARJEŠKE TOP	1980	redni	univerzitetni študij	IA25	EKONOMIJA, Poslovni oddelek, poslovnoinformacijska sme	4	m

Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

4.2 OPREDELITEV PROBLEMA

Ko so podatki zbrani, se moramo odločiti, kaj želimo iz njih izvleči, oziroma do katerih informacij želimo priti. Proces bi bil lahko tudi obraten in sicer bi najprej definirali problem, nato pa bi začeli zbirati potrebne podatke. To je odvisno predvsem od tega, kateri podatki in v kakšni obliki so nam na voljo. Glede na to, da se podatki v današnjem času večinoma sproti shranjujejo v raznih bazah, sem postopek zbiranja podatkov umestil pred opredelitvijo problema.

V mojem primeru so mi bili najprej na voljo zbrani podatki, šele nato sem lahko začel ugotavljati, do kakšnih informacij bi lahko na podlagi le-teh prišel. Ugotavljali bi lahko predvsem dvoje. Kot prvo, kakšna je odvisnost pogostosti obiskov od posameznih značilnosti (atributov) študentov, oziroma kateri atribut sploh vpliva na pogostost obiskov. Ko govorim o obiskih študentov, mislim na prijave študentov na računalnikih CEK (predpostavimo, da vsako prijavo na računalnikih CEK štejemo kot obisk). Lahko bi tudi ugotavljali, kateri izmed vseh teh atributov ima največji vpliv na pogostost obiskov študentov. Kot drugo pa lahko ugotavljamo, kakšna je odvisnost števila obiskov glede na časovne komponente, tj. mesec obiska (v katerem mesecu je dan) in/ali dan v tednu.

4.3 PRIPRAVA PODATKOV

Ko sem zbral podatke in definiral, kaj želim iz njih izvleči, sem jih moral urediti tako, da bi čim bolj enostavno prišel do želenih ugotovitev. Dobljene podatke je bilo zato potrebno urediti, oziroma prilagoditi za enostavnejšo analizo. Že takoj na

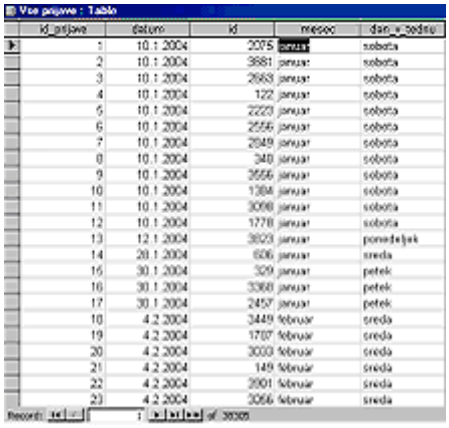
začetku je bilo jasno, da imamo po nepotrebnem tri različne tabele, ki nam opisujejo isto zadevo; to so tabele *CEK 1-19 individualno delo*, tabela *CEK 20-46 dve ali več oseb + miza* in tabela *Dodatne osebe na računalnikih*. Razlikujejo se po tem, da je vsaka tabela omejena le na določene računalnike, s katerimi se je dostopalo do podatkov. Zato sem te tri tabele združil v novo tabelo pod skupnim imenom *Vse prijave*. V prenovljeni bazi podatkov imamo tako tri tabele: tabelo *Studenti*, tabelo *Vse prijave* ter tabelo *Datum*, ki sem jo dodal kasneje.

4.3.1 Tabela Vse prijave

Po izdelavi nove tabele o prijavah, sem se lotil priprave podatkov v tabeli. Najprej sem se znebil podatkov, ki jih nisem potreboval (niso pomembni za analizo). Tako sem zbrisal podatke o tem, ob kateri uri se je študent prijavil na računalnik CEK. Zabeleženih je bilo tudi nekaj obiskov, kjer ni bilo zapisane identifikacijske številke študenta. Tudi te podatke sem zbrisal, čeprav to za analizo časovne odvisnosti niti ne bi bilo potrebno.

Po »čiščenju« podatkov, sem se lotil njihovega urejanja. Napravil sem nekaj novih tabel, v našem primeru stolpcev (od sedaj naprej bom uporabljal ta izraz), za katere sem menil, da mi bodo omogočili lažjo analizo. Za potrebe časovne analize sem dodal dva stolpca in sicer stolpec *mesec* (pove nam v katerem mesecu v letu je bil dan v tabeli datum) in stolpec *dan_v_tednu* (kateri dan v tednu je bil dan iz tabele datum). Kot ključ sem dodal stolpec *id_prijave* (identifikacijsko številko prijave) (glej Sliko 3).

Slika 3: Tabela Vse prijave



Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

4.3.2 Tabela Studenti

Tabelo *Studenti* sem uredil tako, da sem tabeli dodal nekaj novih stolpcev. Stolpec *stevilo_prijav* pove za vsakega študenta posebej, kolikokrat v našem časovnem

obdobju se je prijavil v knjižnici preko računalnikov CEK. Vrednosti sem dobil v tabeli *Vse prijave* in s pomočjo uporabe programa Excel (z uporabo funkcije »COUNTIF«).

Tabeli sem dodal tudi stolpec *okolica*. V tabeli sicer že imamo stolpec *kraj_bivanja*, vendar pa je ta delitev, za potrebe naše analize, mnogo preveč podrobna. Zato sem vse kraje razdelil na dve kategoriji: kraje iz Ljubljane (vrednost LJUBLJANA) in kraje izven Ljubljane (IZVEN LJUBLJANE) v stolpcu *okolica*. Tudi tu sem si pomagal z uporabo programa Excel in funkcijo »IF«.

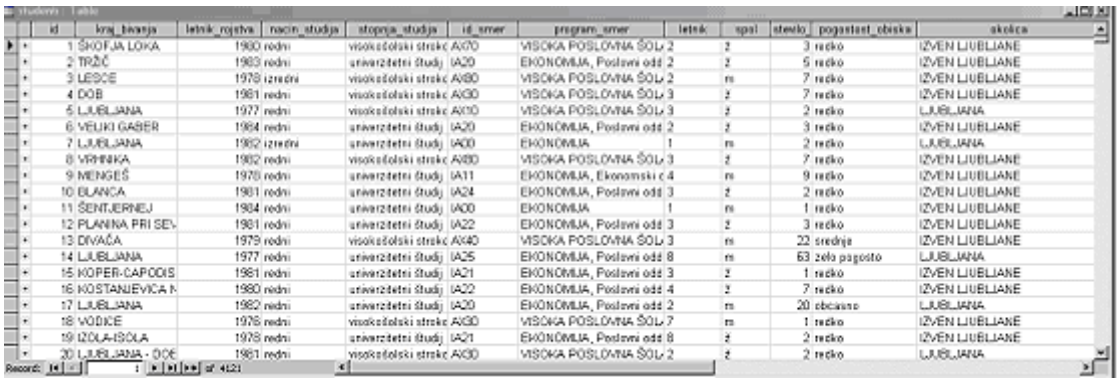
Tretji dodatek v tabeli *Student* je stolpec *pogostost_obiska*. To je v bistvu prirejen stolpec *stevilo_prijav.* Namen te tabele je združevanje študentov v skupine, glede na število prijav. Študente sem razdelili v 5 skupin, kriterij za to delitev pa sem za potrebe diplomske naloge določil sam (glej Tabelo 1).

Tabela 1: Kriterij za združevanje v skupine

POGOSTOST OBISKA	ŠTEVILO OBISKOV
Redko	0 do 10
Občasno	nad 10 do 20
Srednje	nad 20 do 30
Pogosto	nad 30 do 40
Zelo pogosto	nad 40

Vir: Lastna zasnova

Slika 4: Prirejena tabela Student



id	kraj_bivanja	letnik_rojstva	nacin_studija	stopnja_studija	id_smer	program_smer	letnik	spol	stevilo	pogostost_obiska	okolica
1	ŠKOFJA LOKA	1980	redni	visokošolski strokc	AV70	VISOKA POSLOVNA ŠOL	2	ž	3	redko	IZVEN LJUBLJANE
2	TRŽČ	1983	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni odd	2	ž	5	redko	IZVEN LJUBLJANE
3	LESCE	1979	izredni	visokošolski strokc	AV80	VISOKA POSLOVNA ŠOL	2	m	7	redko	IZVEN LJUBLJANE
4	DOB	1981	redni	visokošolski strokc	AV30	VISOKA POSLOVNA ŠOL	3	ž	7	redko	IZVEN LJUBLJANE
5	LJUBLJANA	1977	redni	visokošolski strokc	AV10	VISOKA POSLOVNA ŠOL	3	ž	2	redko	LJUBLJANA
6	VELIKI GABER	1984	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni odd	2	ž	3	redko	IZVEN LJUBLJANE
7	LJUBLJANA	1982	izredni	univerzitetni študij	IA00	EKONOMIJA	1	m	2	redko	LJUBLJANA
8	VRHNIKA	1982	redni	visokošolski strokc	AV80	VISOKA POSLOVNA ŠOL	3	ž	7	redko	IZVEN LJUBLJANE
9	MENGEŠ	1978	redni	univerzitetni študij	IA11	EKONOMIJA, Ekonomski c	4	m	9	redko	IZVEN LJUBLJANE
10	BLANCA	1981	redni	univerzitetni študij	IA24	EKONOMIJA, Poslovni odd	3	ž	2	redko	IZVEN LJUBLJANE
11	ŠENTJERNEJ	1984	redni	univerzitetni študij	IA00	EKONOMIJA	1	m	1	redko	IZVEN LJUBLJANE
12	PLANINA PRI SEV	1981	redni	univerzitetni študij	IA22	EKONOMIJA, Poslovni odd	3	ž	3	redko	IZVEN LJUBLJANE
13	DIVAČA	1979	redni	visokošolski strokc	AV40	VISOKA POSLOVNA ŠOL	3	m	22	srednje	IZVEN LJUBLJANE
14	LJUBLJANA	1977	redni	univerzitetni študij	IA25	EKONOMIJA, Poslovni odd	8	m	63	zelo pogosto	LJUBLJANA
15	KOPER-CAPODIS	1981	redni	univerzitetni študij	IA21	EKONOMIJA, Poslovni odd	3	ž	1	redko	IZVEN LJUBLJANE
16	KOSTANJEVICA N	1980	redni	univerzitetni študij	IA22	EKONOMIJA, Poslovni odd	4	ž	7	redko	IZVEN LJUBLJANE
17	LJUBLJANA	1982	redni	univerzitetni študij	IA20	EKONOMIJA, Poslovni odd	2	m	20	obcasno	LJUBLJANA
18	VODICE	1978	redni	visokošolski strokc	AV30	VISOKA POSLOVNA ŠOL	7	m	1	redko	IZVEN LJUBLJANE
19	IZOLA-ISOLA	1978	redni	univerzitetni študij	IA21	EKONOMIJA, Poslovni odd	8	ž	2	redko	IZVEN LJUBLJANE
20	LJUBLJANA - ODE	1981	redni	visokošolski strokc	AV30	VISOKA POSLOVNA ŠOL	2	ž	2	redko	LJUBLJANA

Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

4.3.3 Tabela Datum

Tabelo *Datum* sem oblikoval naknadno, saj sem jo potreboval pri časovni analizi števila prijav. Sestavljajo jo stolpci: *datum, id_datum, stevilo_prijav_datum, mesec_d* in *dan_v_tednu_d*.

4.4 ANALIZA PODATKOV

Po pripravi in ureditvi podatkov v relacijski bazi podatkov, sem se lotil analize oziroma postopka rudarjenja podatkov. Pri tem sem uporabil program SQL Server. Najprej je bilo potrebno izdelati podatkovno kocko, ki sem jo napravil v obliki zvezdnate sheme (angl. star schema). Zvezdnata shema vsebuje eno tabelo dejstev (angl. fact table) in več dimenzijskih tabel (angl. dimension tables), ki so z njo povezane.

Glede na razpoložljive podatke, sem sestavil večdimenzijsko podatkovno strukturo (kocko) ter opravil prvo analizo samih podatkov o prijavah študentov v knjižnici.

4.4.1 SESTAVLJANJE KOCKE

Na začetku je bilo potrebno določiti mero (angl. measure), torej tisto količinsko vrednost, ki jo želimo v bazi podatkov analizirati. V našem primeru je to število prijav (*stevilo_prijav*), ki jih najdemo v tabeli *Studenti*.

Naslednji korak je definiranje dimenzij kocke. Dimenzije (angl. dimensions), ki sem jih določil za potrebe analize so:

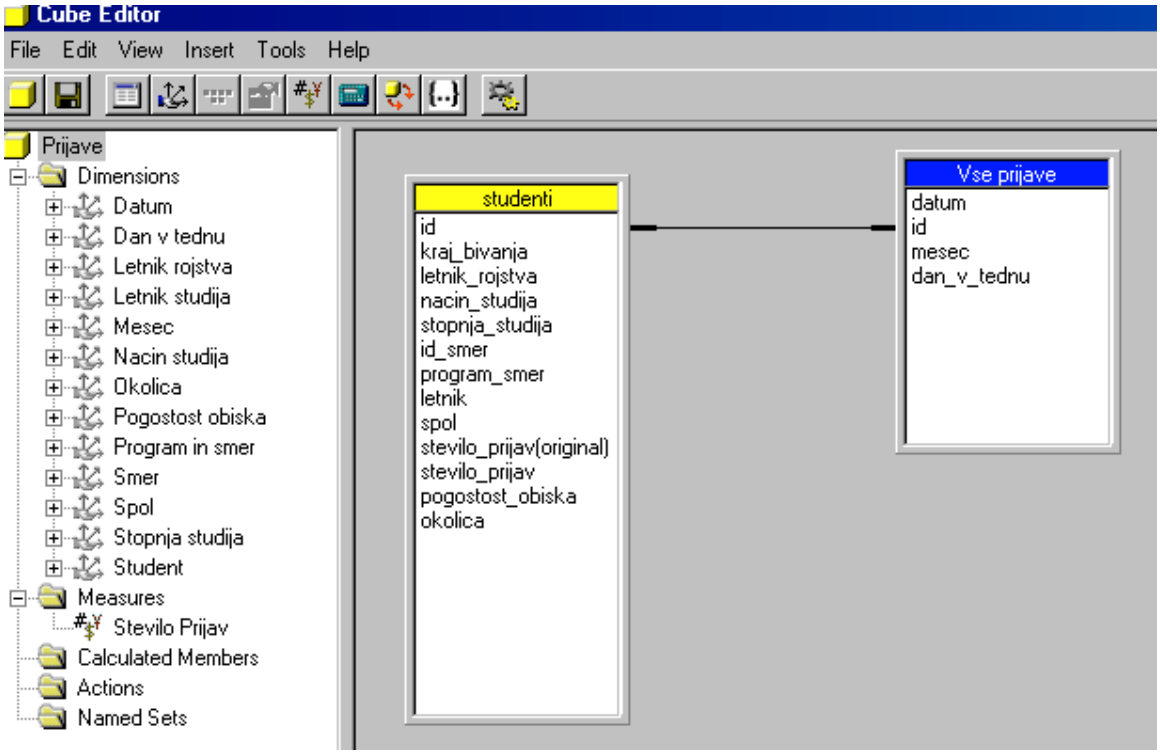
- Datum (leto, mesec, dan v mesecu), ki je dejansko tudi edina časovna dimenzija v naši analizi
- Mesec
- Dan v tednu
- Letnik rojstva
- Nacin studija
- Okolica
- Pogostost obiska
- Program in smer
- Smer (tu je mišljena identifikacijska številka smeri)
- Spol
- Stopnja studija.

Izgled strukture naše kocke *Prijave* lahko vidimo na Sliki 5, na str. 30. (Upoštevati moramo dejstvo, da še nisem uporabil tabele *Datum*). Tabela *Studenti* je naša tabela dejstev, ki je povezana z dimenzijsko tabelo *Vse prijave*. Kot vidimo sta tabeli povezani preko atributa *id* (id študenta). Na levi strani slike pa so naštete vse definirane dimenzije (na sliki Dimensions), kot tudi mera (Measures; število prijav).

Za potrebe analize je bilo potrebno podatke agregirati (združiti) in kocko obdelati. Način shranjevanja (angl. storage mode), ki sem ga uporabil, je bil MOLAP

(večdimenzijska sprotna analitična obdelava podatkov). S tem, ko sem kocko obdelal, sem jo »napolnil« s podatki iz ODBC vira (angl. ODBC source) in tako omogočil pregled predvidenih vrednosti.

Slika 5: Struktura podatkovne kocke



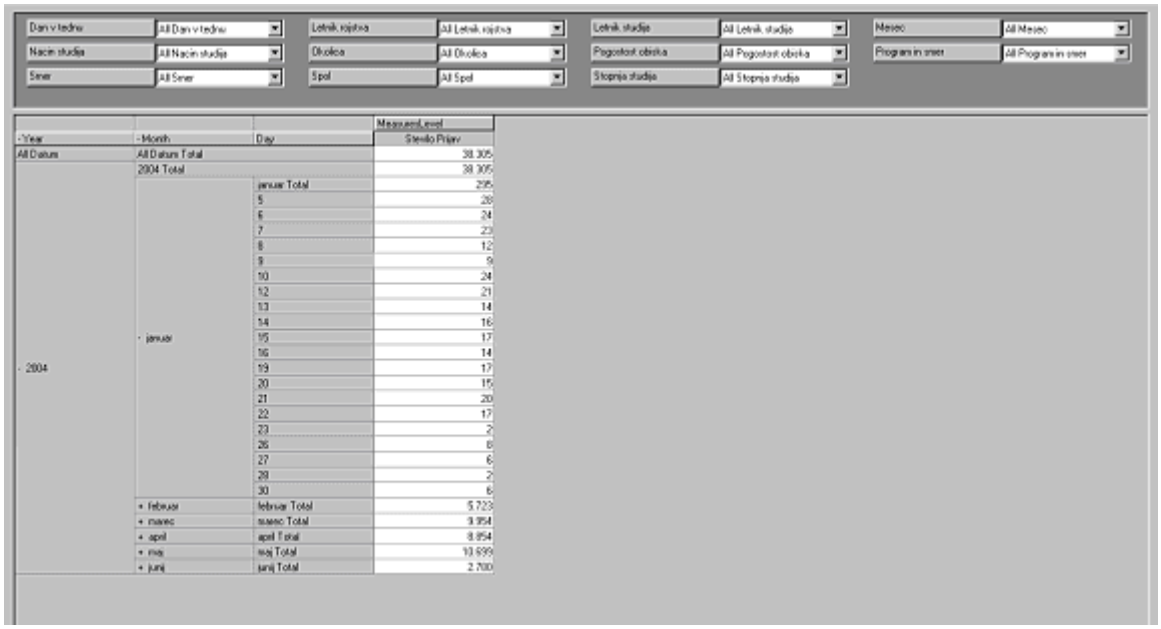
Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

4.4.2 AGREGIRANJE PODATKOV

Podatke lahko analiziramo po vseh tistih dimenzijah, ki sem jih določil v strukturi kocke. Na Sliki 6, na str. 31 vidimo razdelitev števila prijav študentov glede na datum prijave. Tako lahko najprej razberemo število prijav v določenem letu (v našem primeru leto 2004). Nato pa, če je to mogoče, po dimenziji še nadaljnje »vrtamo« (angl. drill down). Lahko vidimo, da imamo v našem primeru za leto 2004 podatke o številu prijav za obdobje od januarja do junija. Podobno lahko opravimo postopek vrtanja še za vsak mesec posebej.

Že tu je razvidno, da obseg podatkov v naši bazi ni dovolj reprezentativen. Za podrobnejšo analizo bi vsekakor potreboval podatke za vsaj nekaj let. V mojem primeru imam na voljo le podatke za 6 mesecev leta 2004, s čimer sem izgubil možnost analize sezonskih nihanj, ki bi bila vsekakor smiselna. A ne glede na to menim, da je primer uporaben za predstavitev postopka rudarjenja podatkov.

Slika 6: Agregirani podatki



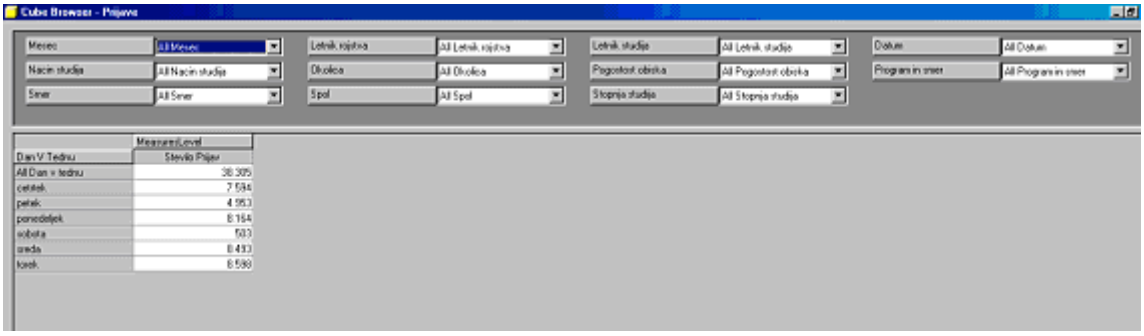
Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

Pri analizah, kadar imamo agregiranih veliko število podatkov, je težko priti do konkretnejših ugotovitev. Večja kot je količina podatkov in večje kot je število opazovanih dimenzij, bolj je otežen pregled nad agregiranimi podatki. Prav gotovo pa ima pomemben vpliv na preglednost analize tudi raznolikost podatkov znotraj dimenzije. Če imamo za nek atribut veliko različnih podatkov, potem je zelo težko »s prostim očesom« iz danih podatkov ugotoviti kaj uporabnega. Še posebej bi bilo to zahtevno takrat, kadar bi iskali neke nove značilnosti, torej če bi iskali nekaj, česar vnaprej nismo predvideli. V našem primeru, razen ko imamo opravka s časovnimi dimenzijami, ni mogoče natančno ugotoviti odvisnost števila prijav študentov od izbranih dimenzij. Npr. če bi delali analizo glede na kraj bivanja (kraji v Sloveniji, kjer živijo uporabniki računalnikov CEK), kjer imamo veliko raznolikost podatkov, bi z analizo OLAP težko prišli do (uporabnih) ugotovitev. Tabela, ki bi jo dobili kot rezultat, bi bila velika in preglednost nad rezultati bi bila premajhna. Tu sem si na nek način pomagal s tem, da sem dimenzijo *kraj bivanja* zamenjal z dimenzijo *okolica*, pri kateri so rezultati bolj pregledni, saj se nam v tem primeru izpišeta samo dve vrednosti: Ljubljana in izven Ljubljane. V primerih, ko je preglednost nad rezultati analize premajhna, moramo uporabiti rudarjenje podatkov. Predvsem, da odkrijemo tiste značilnosti, ki jih z analizo OLAP, zaradi prevelike raznolikosti podatkov, težko opazimo.

Z analizo OLAP pa lahko dokaj enostavno ugotovimo, kakšna je odvisnost števila prijav glede na dan v tednu. Kot vidimo na Sliki 7, na str. 32, je obisk (oziroma prijave na računalnikih CEK) od ponedeljka do četrtka konstanten, v petek upade za približno polovico, v soboto pa je že zelo majhen. Te ugotovite so razumljive. Dosti študentov, ki ne živijo v Ljubljani, odide domov že v četrtek, kar je zagotovo

razlog za močan upad obiskov v petek. V soboto v glavnem ni predavanj in posledično tudi ni študentov v prostorih fakultete, v nedeljo pa je knjižnica na Ekonomski fakulteti zaprta. Tako lahko predvidimo, da je število prijav relativno odvisno tudi od tega, ali študentje prihajajo (živijo) iz Ljubljane ali izven Ljubljane. Vendar tega ne moremo zagotovo trditi, vse dokler ne opravimo rudarjenja podatkov.

Slika 7: Število obiskov glede na dan v tednu



Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

4.4.3 RUDARJENJE PODATKOV

Za podrobnejšo in natančnejšo analizo agregiranih podatkov, se moramo poslužiti postopka rudarjenja podatkov. Lotil sem se ga na dva načina. Najprej sem ugotavljal, kakšna je odvisnost pogostosti obiska glede na atribute (lastnosti) študenta, nato pa sem se lotil še odvisnosti števila prijav glede na časovni dimenziji, ki sta nam na voljo (torej meseca in dneva v tednu).

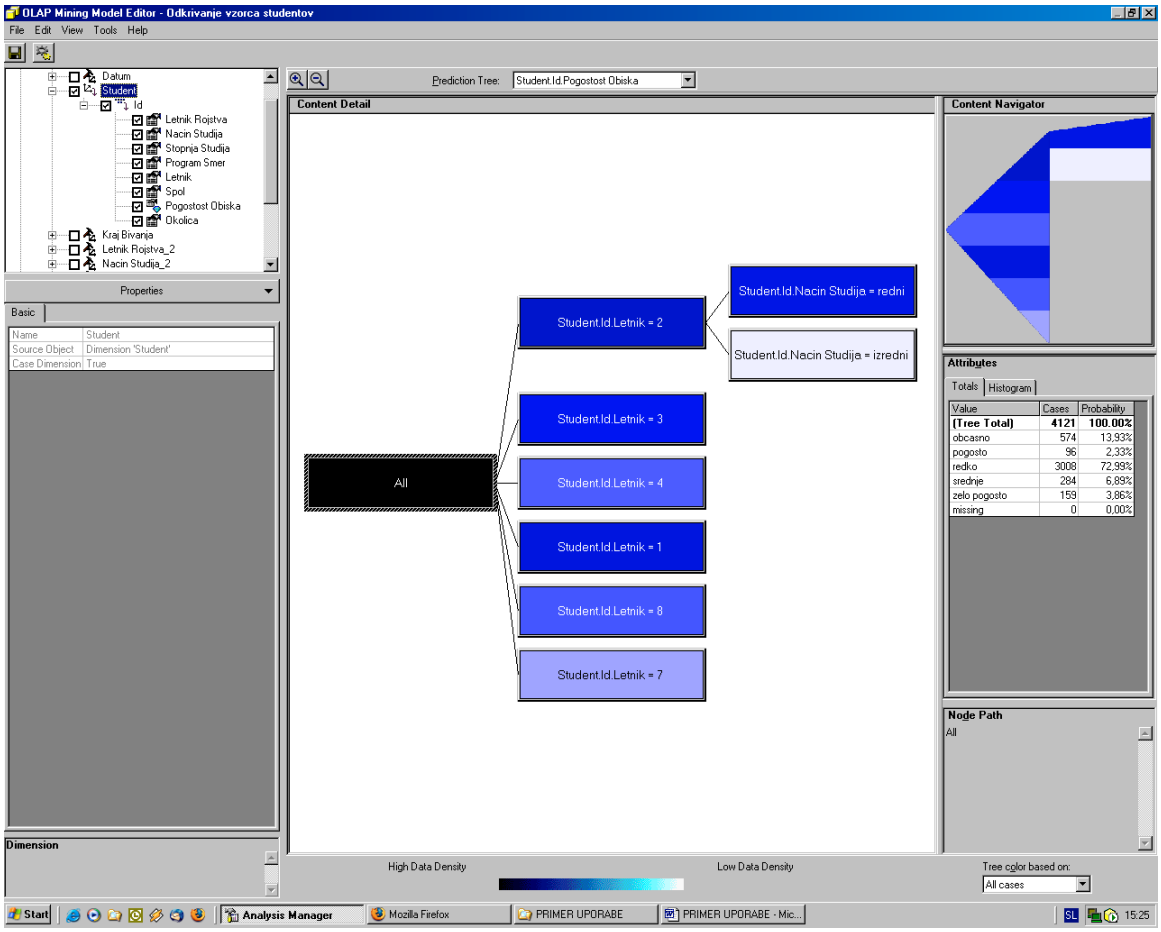
Rudarjenje podatkov je uporabno pri odkrivanju in opisovanju skritih vzorcev v določeni podatkovni kocki. Ker je količina podatkov dostikrat zelo obsežna, je »ročno« zelo težko najti kakšne koristne informacije. Rudarjenje podatkov nam za to ponuja algoritme, ki nam omogočajo avtomatično odkrivanje vzorcev in interaktivno analizo.

Tukaj se pojavi vprašanje smiselnosti rudarjenja podatkov v primeru ugotavljanja odvisnosti števila prijav glede na časovni dimenziji. Analiza OLAP kocke bi bila lahko dovolj, saj so rezultati pregledni in ne preveč obsežni. Vseeno pa sem, tudi v tem primeru, izvedel postopek rudarjenja in to predvsem iz enega razloga. Z analizo OLAP sicer lahko ugotovimo, v katerih mesecih in dnevih v tednu je večje število prijav na računalnikih CEK, vendar pa lahko le s pomočjo rudarjenja podatkov ugotovimo, katera izmed teh dveh dimenzij ima večji vpliv, tj. ali na število prijav bolj vpliva mesec ali dan v tednu.

ODVISNOST POGOSTOSTI OBISKA

Kot sem že omenil, sem ugotavljal tudi odvisnost pogostosti obiska oziroma prijave študentov na računalnikih CEK, glede na njihove demografske in študijske značilnosti. Zanimalo me je, kateri od teh dejavnikov ima največji vpliv na pogostost obiskov, oziroma kateri sploh vpliva nanje. Napravil sem model rudarjenja podatkov z uporabo algoritma drevesa odločanja (Slika 8).

Slika 8: Drevo odločanja: odvisnost pogostosti obiska



Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

V levem zgornjem kotu so predstavljene dimenzije (vhodi), na podlagi katerih sem zgradil model podatkovnega rudarjenja. Potrebno je bilo definirati, katere dimenzije kocke sem želel vključiti v model. V našem primeru je to dimenzija Student, ki je za nas zanimiva predvsem na drugem nivoju, kjer imamo atribute, po katerih želimo opazovati odvisnost pogostosti obiska: letnik rojstva, način študija, stopnja študija, program-smer, letnik, spol in okolica. Potrebno je bilo tudi definirati opazovano lastnost (member property). V našem primeru je bila to pogostost obiska.

Kako lahko razložimo dobljene rezultate? Različna obarvanost posameznih polj pri odločitvenem drevesu pomeni, da bolj kot je barva intenzivna (temna), več primerov polje vsebuje. Zato je torej polje *All (Vsi)* obarvano s črno barvo.

Poglejmo si še strukturo našega drevesa. Prvi nivo drevesa je »razvejan« glede na letnik. Struktura drevesa je določena z algoritmom, ki temelji na pomembnosti dejavnika na vrednost, ki sem jo določil. Trdim lahko, da je letnik študija dejavnik, ki ima največji vpliv na pogostost prijav študentov na računalnike CEK. Lahko tudi opazimo, da ima naše odločitveno drevo samo še en nadaljnji nivo. Tako lahko ugotovimo, da pri študentih drugega letnika, na pogostost obiska vpliva še način študija, tj. redni ali izredni.

REZULTATI ANALIZE
Na pogostost obiska vplivata letnik študija in način študija, pri čemer ima prvi močnejši vpliv. Ostali dejavniki imajo na pogostost obiska zanemarljiv vpliv, oziroma ga sploh nimajo. S tem se podre naše predvidevanje, da na pogostost obiska vpliva tudi okolje bivanja (to ali živi študent v ali izven Ljubljane). Z nadaljnjo analizo lahko dobimo še druge informacije, ki pa jih lahko ugotovimo tudi z analizo OLAP (npr. ugotovimo lahko, da izmed vseh študentov, ki so se prijavljali na računalnike CEK, se je prijavilo največ študentov drugih letnikov in da se je večina (skoraj 73%) študentov vseh letnikov, ki so se prijavljali, le redko prijavila na računalnike).

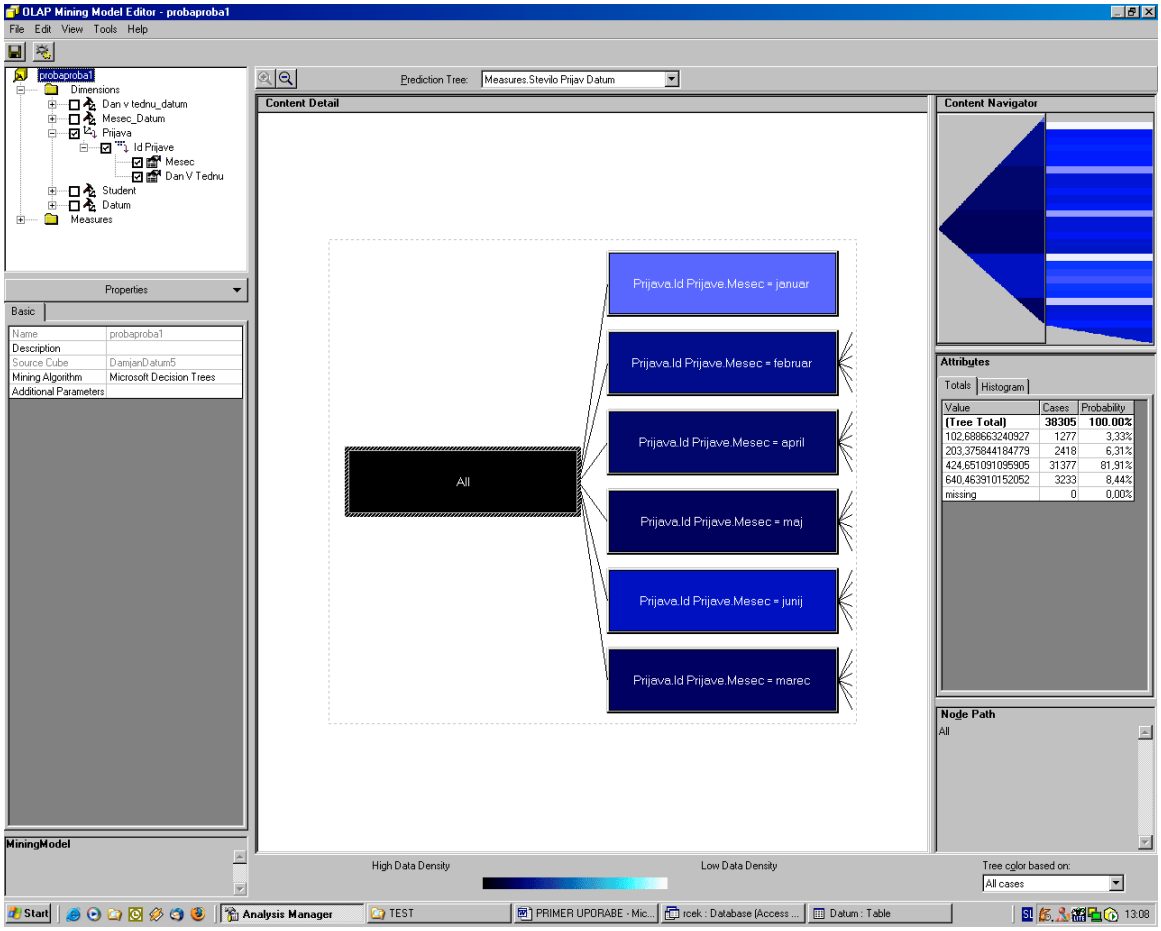
ČASOVNA ODVISNOST

V tem primeru sem ugotavljal, kakšna je odvisnost števila obiskov glede na mesec in glede na dan v tednu. Kot sem že omenil, ta analiza v našem primeru ni najbolj zanesljiva, saj imam količinsko premalo podatkov, vendar jo bom, predvsem zaradi prikaza možnosti tovrstne analize, vseeno opravil. Za razliko od analize OLAP, bomo tu dobili odgovor tudi na vprašanje, kateri izmed teh dejavnikov je pomembnejši.

Za izvedbo postopka rudarjenja, sem moral pri strukturi podatkovne kocke dodati še tabelo *Datum.* S tem pa sem se moral ponovno vrniti nazaj na postopek priprave podatkov.

V model podatkovnega rudarjenja sem vključil dimenzijo *Prijava*, opazoval pa sem odvisnost števila prijav glede na mesec in dan v tednu.

Slika 9: Drevo odločanja: časovna odvisnost



Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

REZULTATI ANALIZE

Dobljeno drevo odločanja si lahko razlagamo enako kot v prejšnjem primeru. Število prijav na računalnikih CEK je bolj odvisno od tega, katerega meseca v letu smo, kot pa od tega, kateri dan v tednu je. Naše drevo je namreč na prvi stopnji razvejano glede na mesec, na drugi pa glede na dan v tednu. Trdimo lahko, da je sezonska komponenta bolj pomembna z vidika števila obiskov kot pa tedenska komponenta. Kar je po neki strani tudi razumljivo, saj je zelo verjetno, da se obisk v knjižnici v obdobju počitnic zmanjša (predvsem poleti) in poveča v obdobju pred izpitnimi roki. Še enkrat pa moram poudariti, da bi za bolj zanesljivo analizo potreboval podatke za nekoliko daljše obdobje, menim, da vsaj za obdobje nekaj let. Ne gre zanemariti tudi nihanja števila obiskov znotraj obdobja enega tedna, saj je opazen upad števila prijav na računalnikih CEK proti koncu vsakega tedna (Slika 9).

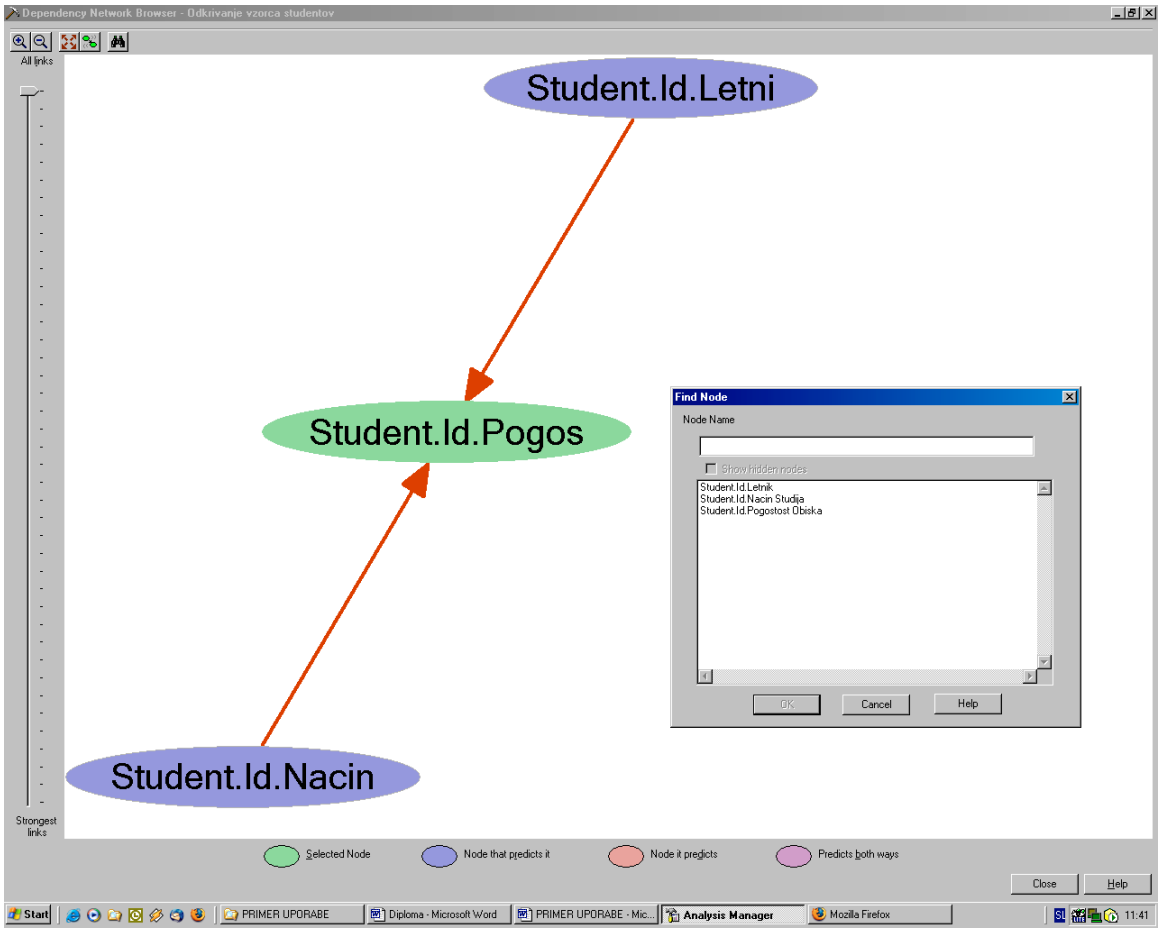
Rezultate bi lahko ponovno še podrobneje analizirali. Kot vidimo, je v našem zelo omejenem časovnem obdobju, največji obisk mogoče zaznati v mesecu maju. Verjetno se je takrat dosti študentov že pripravljalo na junijske izpitne roke, oziroma se je ukvarjalo z izdelavo seminarskih nalog. Morda nas zato toliko bolj

čudi najnižji obisk v mesecu januarju, vendar pa moramo tu upoštevati čas praznikov in počitnic ter posledično manjše število aktivnosti glede predavanj in izdelav seminarskih nalog. Podobno lahko vidimo upad obiskov tudi v mesecu juniju. Tako lahko trdimo, da se v času izpitnih rokov zmanjša obisk računalnikov CEK. Tudi do teh informacij bi lahko prišli že samo na podlagi analize OLAP, vendar pa se mi zdijo v tej obliki vseeno bolj pregledne.

MREŽA ODVISNOSTI

Z mrežo odvisnosti si pomagamo takrat, ko želimo (na drugačen način) prikazati, kakšna je povezava med atributi, oziroma od česa (katerega atributa) je odvisna opazovana vrednost. Pomembno je predvsem to, da lahko prikažemo relativno moč posameznega atributa na neko vrednost.

Slika 10: Mreža odvisnosti in pogostost obiska



Vir: Podatki o prijavah na računalnikih CEK, 28.6. 2004.

V našem primeru lahko to prikažemo tako za odvisnost pogostosti obiska, kot za časovno odvisnost. Kot vidimo na Sliki 10, je pogostost obiska (Student.Id.Pogos) odvisna od letnika študija (Student.Id.Letni) in načina študija (Student.Id.Nacin). Ker sta ti dve polji obarvani z vijolično barvo, to pomeni, da napovedujeta (angl.

predict) vrednost v zelenem polju; torej letnik in način študija vplivata na pogostost obiska. Če pogledamo na levo stran slike, vidimo drsnik, katerega funkcija je, da nam pokaže moč odvisnosti. Bolj ko drsimo z drsnikom navzdol, bolj izginjajo »šibkejše« povezave in na koncu ostane samo najmočnejša. V našem primeru izgine povezava med pogostostjo obiska in načinom študija. Tako dokažemo, da je pogostost obiska (naj)bolj odvisna od letnika študija.

Podobno sem storil tudi za časovno odvisnost in ugotovil, da je število prijav odvisno prav tako od meseca kot od dneva v tednu, s tem, da ima prvi dejavnik večji vpliv.

4.5 RAZLAGA UGOTOVITEV

Po opravljeni analizi lahko ugotavljam ali imajo pridobljene informacije za nas kakšen smisel? Sedaj sicer vemo, kdo so naši bolj zvesti obiskovalci, vendar pa v samem neposrednem odnosu do študentov, ne moremo kaj dosti storiti. Rezultate analize lahko interpretiramo na mnogo načinov, odvisno predvsem od naših potreb in sposobnosti razumevanja rezultatov. Vsekakor bi lahko napravili mnogo boljšo analizo, če bi imeli na voljo več podatkov. Predvsem bi bilo zanimivo, če bi imeli podatke o literaturi, ki so jo pregledovali študentje. S tem bi na primer lahko ugotavljali, ali je izbor literature odvisen od letnika študija ter se tako že približali ideji personalizacije; npr. študentu tretjega letnika bi ponudil povezavo do literature, ki so jo ostali študenti tretjih letnikov največkrat uporabili pri izdelavi seminarskih nalog.

Sicer je tema moje diplomske naloge spletno rudarjenje, vendar pa bi glede na primer težko določil, kaj je tisto, kar mojo analizo loči od navadnega podatkovnega rudarjenja. Vseeno menim, da obstaja nekaj značilnosti, zaradi katerih bi to analizo lahko opredelili kot spletno rudarjenje. Preučeval sem zbirko dostopov in prijav uporabnikov, ki so se morali identificirati, če so želeli uporabljati računalnik CEK. Tako so sicer odpadle težave v zvezi z razpoznavo uporabnikov, prav tako zaradi enostavnosti problema tudi nisem imel dela z razpoznavanjem uporabniških sej. Čeprav svojih podatkov nisem dobil na svetovnem spletu, lahko pri naši podatkovni bazi najdemo dovolj podobnosti z običajno »online« podatkovno bazo. Tako sem lahko, na podlagi vzorca dostopov uporabnikov, izvedel proces izkopavanja spletne uporabe. Na nek način je logiranje na računalnikih CEK podobno logiranju na spletni strani.

5 SKLEP

Rudarjenje in spletno rudarjenje podatkov postajata nedvomno vse bolj zanimiva, predvsem v poslovnem svetu. Uporabnost rudarjenja se pokaže najbolj takrat, ko imamo opravka z velikimi količinami podatkov. Informacije, ki jih na ta način dobimo, morajo biti razumljive, da so lahko koristne in uporabne predvsem za management v podjetju. Prav zato je pomembno, da imamo primerno izobražen kader, z znanjem tako s področja analize podatkov, kot tudi s področja obvladovanja poslovnih znanj. Ravno tako ne smemo spregledati, da postajajo orodja za rudarjenje podatkov uporabniku vse prijaznejša.

Uporabniki imajo na svetovnem spletu različne cilje, zahteve in želje, zato moramo do njih pristopati individualno. To lahko dosežemo s pomočjo uporabe spletnega rudarjenja, ki v širšem smislu pomeni odkrivanje in razlago uporabnih informacij ter znanj s svetovnega spleta. Spletno rudarjenje je torej rudarjenje podatkov na svetovnem spletu. Bistvena razlika med rudarjenjem podatkov in spletnim rudarjenjem, je časovni okvir analize. Hitrost je pri spletnem rudarjenju dosti bolj pomembna. Eden glavnih ciljev spletnega rudarjenja je tudi izboljšava oblike spletne strani, in sicer tako z vidika privlačnosti strani za samega uporabnika, kot tudi boljšega omogočanja nadaljnjega procesa rudarjenja.

Spletno rudarjenje lahko razdelimo na več aktivnosti, vendar je bilo, z vidika našega primera, najbolj zanimivo spletno rudarjenje vzorcev uporabe, pri katerem pa se v mnogih primerih srečujemo s problemom razpoložljivosti informacij. Podatki so navadno zabeleženi na številnih strežnikih, ki so v lasti različnih podjetij, in mnoga od njih nimajo ne možnosti in ne interesa deliti teh informacij. Težave se pojavijo tudi pri prepoznavanju uporabnikov. Tu si lahko pomagamo z uporabo gesla oziroma z registracijo uporabnikov, prepoznavanjem IP naslova in uporabo piškotkov. Rudarjenje vzorcev uporabe nam pomaga pri odkrivanju različnih načinov uporabe neke strani in nam predlaga načine, kako to stran izboljšati. Z vidika razširjenosti uporabe, je zelo pomembna tudi uporaba rudarjenja po spletnih vsebinah, saj si brez spletnih iskalnikov težko predstavljamo uporabo svetovnega spleta.

Pri procesu podatkovnega rudarjenja je zelo zaželeno sodelovanje uporabnika. Tu imamo v mislih zdravo mero uporabnikove interaktivnosti, saj ga nočemo preveč obremenjevati in posledično izgubiti njegovo zanimanje. Uporabnik in računalnik naj bi delovala v navezi. Za sam proces spletnega rudarjenja je pomembno, da uspemo o uporabniku pridobiti čim več uporabnih informacij. Poleg statičnih podatkov, nas pri uporabniku zanima predvsem transakcijski del oziroma njegovo obnašanje na spletišču. Zanima nas katere storitve uporablja, kakšna navigacija po spletni stani mu ustreza ipd. Na tak način lahko zgradimo profil za vsakega

vpisanega uporabnika in na podlagi tega zagotovimo uporabnikom personaliziran dostop na spletno stran.

Z izdelavo primera uporabe sem ugotovil, na kaj moramo biti pozorni pri postopku spletnega rudarjenja. Pomembno je, da so podatki, ki so nam na voljo, dovolj obsežni in vsebinsko primerni. Sam sem imel pri tem težave, saj sem imel količinsko premalo podatkov za boljšo izdelavo časovne analize. Prav tako je pomembno, da na podlagi zbranih podatkov, smiselno definiramo problem, saj je od tega odvisno, kako uporabne bodo naše ugotovitve. Resno se moramo lotiti tudi priprave samih podatkov, saj si s tem omogočimo enostavnejši postopek analize. V predstavljenem primeru porabe sem imel največ težav z razlago ugotovitev, saj bi na podlagi rezultatov analize težko predlagal kakšno uporabno rešitev v praksi.

Podatkovno rudarjenje je smiselno torej takrat, ko imamo opravka s kompleksnimi podatki. Za enostavnejše raziskave pa nam zadostuje že uporaba analize OLAP, ki je enostavnejša ter povzroča manj dela in stroškov. V primerih, ko imamo agregiranih količinsko veliko število podatkov, pa z analizo OLAP težko pridemo do konkretnejših ugotovitev. Več kot je podatkov in čim večje je število dimenzij, bolj je otežen pregled nad agregiranimi podatki. V takšnih primerih je bolj primerno uporabiti postopek podatkovnega rudarjenja.

LITERATURA

1. Brandel Mary: Spinning data into gold. Computerworld, Framingham, 35(2001), 13, str. 67.

2. Fong Joseph, Wonk H. K.: Online analytical mining of path traversal patterns for Web measurement. Journal of Database Management, Hershey, 13(2002), 13, str. 39-62.

3. Greening Dan R.: Data Mining on the Web: There`s Gold in that Mountain of Data [URL: http://www.webtechniques.com/archives/2000/01/greening/], 26.5.2003.

4. Horvat Branko: Spletno rudarjenje in personalizacija. Magistrsko delo. Maribor : Fakulteta za elektrotehniko, računalništvo in informatiko, 2003. 118 str.

5. Hrastar Urška, Krnc Matej, Škoberne Iva: Spletno rudarjenje. Seminarska naloga. Ljubljana : Ekonomska fakulteta, 2003. 17 str.

6. Kimball Ralph, Merz Richard: The Data Webhouse toolkit: building the web-enabled data warehouse. New York: J. Wiley cop, 2001. 401 str.

7. Linoff Gordon, Berry Michael J. A.: Mining the Web: transforming customer data into customer value. New York: J. Wiley cop, 2001. 348 str.

8. Mattison Rob: Web Warehousing and Knowledge Management. New York: McGraw-Hill, 1999. 336 str.

9. Mobasher Bamshad, Cooley Robert, Srivastava Jaideep: Automatic personalization based on Web usage mining. Association for Computing Machinery. Communications of the ACM, New York, 43(2000), str. 142-152.

10. Robinson Brian: Buidling better Web sites. Federal Computer Week**,** Falls Church, 14(2000), 3, str. 34-35.

11. Srivastava J. et al.: Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, 1(2000), str. 12-23.

VIRI

1. Islovar. [URL: http://www.islovar.org], 1.3. 2005.

2. Podatki o prijavah na računalnikih CEK, 28.6. 2004.

ANGLEŠKO SLOVENSKI SLOVARČEK IZRAZOV

ANGLEŠKI IZRAZ	SLOVENSKI PREVOD
ad server	oglasniški strežnik
application logs	uporabniške prijave
application server	uporabniški strežnik
application server logs	uporabniške spletne prijave
caching	predpomnjenje
chi-square	hi kvadrat
clickstream	potek povezav
cluster	gruča
commerce server	trgovinski strežnik
cookies	piškotki
CRM (customer relationship management)	ravnanje odnosov s strankami
Cross tabulation	tehnika kontingenčnih tabel
customized usage tracking	prirejeno sledenje uporabe
data mining	rudarjenje podatkov, podatkovno rudarjenje
direct marketing	neposredno trženje
dynamic web page technology	dinamična tehnologija spletnih strani
e-commerce	e-trgovina
e-market	e-trg
e-media	e-medij
file	datoteka
fuzzy logic	mehka logika
heuristic reasoning	hevristična metoda reševanja problemov
hyperlink	hiperpovezava
information retrieval	pridobivanje informacij
inteligent agents	inteligentni agentje
IP (internet protocol)	internetni protokol
IP adress	IP naslov
log	zbirka prijav, prijaviti
log file	datoteka dostopov, zbirka prijav
market basket	tržna košarica
market basket analysis	analiza tržne košarice
mining content	rudarjenje po vsebini
mining structure	rudarjenje po strukturi
mining usage	rudarjenje uporabniških vzorcev
MOLAP (multidimensional online analytical processing)	večdimenzijska sprotna analitična obdelava podatkov

ANGLEŠKI IZRAZ	SLOVENSKI PREVOD
neural networks	nevronske mreže
observations sets	zbirke podatkov
OLAP (online analytical processing)	sprotna analitična obdelava podatkov
page request	zahteva po spletni strani
page view	pregled strani
path completion	zaključevanje poti
recommendation engines	priporočilniki
search engine	iskalnik
server	strežnik
server log	zbirka dostopov
session	seja
sessionalization	oblikovanje sej
storage mode	način shranjevanja
warehousing	skladiščenje podatkov
web content mining	rudarjenje po spletnih vsebinah
web designer	oblikovalec spletnih strani
web log	spletna prijava
web mining	spletno rudarjenje
web usage mining	spletno rudarjenje vzorcev uporabe
webhouse	spletno skladišče podatkov