Univerza v Ljubljani

Fakulteta za računalništvo in informatiko

Andrej Remškar

ANALIZA OBISKA V SPLETNI TRGOVINI S POMOČJO PODATKOVNEGA SKLADIŠČA

Diplomsko delo

Mentor:
prof. dr. Viljan Mahnič

Ljubljana, avgust 2002

Povzetek

Glavni cilj diplomske naloge je bil zgraditi podatkovno skladišče, s katerim lahko analiziramo različne parametre obiska v spletni trgovini.

Preučili smo dogajanje v spletni trgovini, še posebej sledi, ki jih uporabniki puščajo na spletnem mestu v dnevniku spletnega strežnika. Ugotovili smo, da je za zasnovo podatkovnega skladišča najprimernejša tehnika dimenzijskega modeliranja. Določili smo tri najbolj tipične procese, povezane z obiskom spletne trgovine, in zanje izdelali dimenzijske podatkovne modele: za analizo uporabniških sej, za analizo obiskov strani in za analizo nakupov. Navedli smo vprašanja, na katera nam ti modeli lahko pomagajo odgovoriti. Definirali smo tabele dejstev in dimenzijske tabele posameznih modelov ter določili vire podatkov zanje.

V praksi smo uporabili podatke iz spletne trgovine podjetja Merkur *(http://www.merkur.si)*. Natančno smo preučili razpoložljive vire, podatke smo uredili in jih prenesli v dimenzijski model za analizo uporabniških sej. S pomočjo orodja Microsoft SQL Server 2000 Analysis Services smo izvedli analizo in prišli do zanimivih ugotovitev o vplivu različnih faktorjev (čas, viri obiskovalcev, vstopne strani, nagradne igre itd.) na obiskanost spletnega mesta, število pogledanih strani, trajanje obiskov, verjetnost nakupa in povprečno vrednost nakupa.

Rezultati opravljene analize lahko urednikom spletne trgovine pomagajo pri dopolnjevanju prodajnega programa, kakovostnejšem razvoju vsebin in funkcionalnosti spletne trgovine, pri odločanju za različne oblike promocije spletnega mesta in podobnih vprašanjih.

Ključne besede

dimenzijsko modeliranje, podatkovno skladišče, spletna trgovina, analiza obiska, analiza dnevnika spletnega strežnika, sledi obiskovalcev, podpora odločanju

Summary

The main goal of this thesis was to build a data warehouse in which we could capture, analyze, and understand the behavior of users visiting an online store.

We studied different aspects of online store, especially the traces that visitors leave in the web server log (clickstream). We found dimensional modeling to be the most appropriate logical design technique for the clickstream data warehouse. We selected three most typical processes from online store (user sessions, page views, and orders) and built respective dimensional models. We stated questions which those models should provide answers to; we defined fact tables, dimensional tables and data sources for each table.

In the practical part of this thesis we used data from the Merkur's online store (*http://www.merkur.si)*. We thoroughly examined all available data – logs and other data sources. We extracted and transformed the data and then loaded it into the dimensional model for analyzing user sessions. With the Microsoft SQL Server 2000 Analysis Services we performed the analysis to found many interesting conclusions about the different parameters' impact (time, referrers, entry pages, online games, etc.) on number of visits, average visit length, conversion rate (visitors to buyers), and average order value.

The results of the analysis are very helpful to online store managers when they make changes in product range, develop new content, choose between different advertising options, etc.

Keywords

dimensional modeling, data warehouse, online store, web log analysis, clickstream analysis

Zahvala

Mentorju prof. dr. Viljanu Mahniču se zahvaljujem za podporo, pomoč in usmerjanje pri izdelavi diplomske naloge. Merkurju, še posebej Juretu Cviklu, hvala za podatke o obisku spletne trgovine in dodatno študijsko gradivo.

Hvala Evi, Ajdi in Maši za prizanesljivost in razumevanje.

Nenazadnje hvala vsem, ki ste verjeli vame in vztrajno spraševali, kdaj bo. Brez vas te naloge morda sploh ne bi bilo.

Kazalo

1. Uvod

Internet je od povezave prvih računalnikov leta 1969, predvsem pa v zadnjih desetih letih, doživel nesluten razmah. Postal je globalno komunikacijsko orodje, na njem poleg elektronske pošte najpogosteje pregledujemo spletne strani. Idejo o dokumentih, med seboj povezanih s hipertekstnimi povezavami, je leta 1989 začel v praksi uresničevati Tim Berners-Lee, nastajajočo mrežo je poimenoval *World Wide Web (WWW)* [4]. Svetovni splet danes sestavlja nepregledna množica med seboj bolj ali manj povezanih strani (samo v iskalniku Google je vključenih čez tri milijarde dokumentov [7]), število v internet vključenih strežnikov pa se je že povzpelo čez 162 milijonov [10].

Po podatkih raziskave RIS [5] je v Sloveniji junija 2002 internet že uporabljalo okoli 850.000 oseb, slaba četrtina prebivalstva (470.000) pa ga uporablja tedensko. Delež uporabnikov interneta med aktivnim prebivalstvom (kot tudi ostali pokazatelji uporabe interneta) z določenim zamikom sledijo trendom v zahodni Evropi.

Podjetja so spoznala priložnost, ki jo je ponudil internet: z relativno majhnim vložkom lahko svojo ponudbo predstavijo globalni javnosti, predstavitev lahko nadgradijo tudi z možnostjo opravljanja transakcij (naročanje) in dvosmerno komunikacijo s strankami. Izkazalo se je, da je meč dvorezen – globalno se lahko za podobno nizko ceno s podobnimi orodji predstavijo vsa podjetja, na lokalno tržišče, ki smo ga obvladovali, pa je prišla globalna konkurenca. Podjetja brez prave strategije in primerjalnih prednosti so praviloma končala v povprečju – v izgubah in stečajih. Mit o internetu kot idealnem prodajnem kanalu se je razblinil spomladi leta 2000, ko so vlagatelji, katerih denar so internetna podjetja pridno trošila, začeli razmišljati o donosnosti svojih vložkov. [6]

Kljub streznitvi ostaja dejstvo, da je internet komunikacijsko povezal ves svet, da z njim res lahko dosežemo globalno občinstvo in ga zato kaže izkoristiti. Značilnost sodobnih načinov uporabe je, da podjetja prek interneta komunicirajo z različnimi deležniki (poleg kupcev in dobaviteljev so to tudi mediji, delničarji, lokalna, splošna in strokovna javnost, zaposleni itd.). Včasih najbolj poudarjeni cilj – prodaja – se je umaknil bolj realnim ciljem: izboljšanju in olajšanju komunikacije z deležniki, izgradnji blagovne znamke, informiranju potencialnih kupcev, ohranjanju komunikacije z obstoječimi kupci, v ozadju (manj vidno splošni javnosti, čeprav za podjetja

zelo pomembno) je elektronsko poslovanje s poslovnimi partnerji *(B2B – Business to Business)*.

Nakupovanje ni aktivnost, zaradi katere bi se ljudje primarno odločali za uporabo interneta. (V ospredju med razlogi za vstop na internet sta komuniciranje po elektronski pošti in iskanje informacij.) Nakupovanje zaenkrat ostaja v domeni bolj izkušenih spletnih uporabnikov, ki mediju tudi bolj zaupajo, čeprav se je v internetno najbolj razvitem delu sveta, v ZDA, demografski profil uporabnikov interneta že zelo približal profilu povprečnega obiskovalca. Tako se je denimo delež žensk med spletnimi nakupovalci že povzpel nad polovico. [8]

Spletno nakupovanje nikakor ni nepomembno – samo v predbožičnem času leta 2001 so ameriški kupci prek interneta opravili za okoli 13,8 milijard USD nakupov [10]; po nekaterih ocenah bo (ravno tako v ZDA) spletno nakupovanje v letu 2005 doseglo 199 milijard USD, zbiranje informacij na osebnih računalnikih in drugih napravah, povezanih v internet, pa bo vplivalo še na dodatne nakupe po drugih kanalih v vrednosti 632 milijard USD [11]. V Sloveniji se s tako impresivnimi podatki sicer še ne moremo pohvaliti, na spletu nakupuje okoli 15 % aktivnih uporabnikov interneta. Večina spletnih trgovcev je hitro spoznala, da bo spletna trgovina lahko preživela le, če ji priznamo tudi vlogo informatorja in generatorja nakupov v fizičnih trgovinah. Po raziskavah namreč tudi v Sloveniji skoraj 70 % uporabnikov interneta na spletnih straneh zbira informacije, ki kasneje vplivajo na nakup. [5]

Uspešna spletna mesta poskušajo njihovi uredniki čim bolje prilagoditi zahtevam in pričakovanjem obiskovalcev. V tej nalogi bomo spoznali, da lahko veliko o navadah in zahtevah obiskovalcev izvemo tudi iz podatkov, zbranih s pomočjo spletnih mest samih. Malce ironično je, da podjetja vlagajo velika sredstva v CRM projekte[1], s katerimi želijo spoznati navade kupcev (samo kupcev, ker v informacijskih sistemih običajno nimajo drugih podatkov) in jim prilagoditi ponudbo, obenem pa ne izkoristijo relativno enostavno dostopnih podatkov, ki so na voljo v dnevnikih spletnih strežnikov. V slednjih lahko najdemo zelo koristne informacije o navadah in zahtevah ne samo kupcev, ampak vseh obiskovalcev, ki nas usmerijo k še boljši ponudbi. Za kakovostno ukrepanje glede na obnašanje uporabnikov pa potrebujemo ustrezno orodje za analizo dogajanja na spletnem mestu.

[1] CRM - Customer Relationships Management (upravljanje odnosov s strankami) je disciplina, ki teži k enotnemu in koordiniranemu nastopu podjetij proti strankam ne glede na komunikacijski kanal; osnova zanj je enotno obravnavanje strank v podjetju (zbiranje in obdelava vseh informacij na enem mestu) in natančno poznavanje strank, njihovih zahtev in pričakovanj. [13]

Cilj te naloge je zasnova podatkovnega skladišča sledi spletnih obiskovalcev[2] *(clickstream data warehouse)*. Uporabili bomo metodo dimenzijskega modeliranja, ki je v praksi bolj primerna za izvedbo analitičnih podatkovnih baz kot entitetni relacijski modeli. Za procese, ki nas zanimajo, bomo zgradili podatkovne modele: določili osnovne tabele dejstev in dimenzije, ki posamezna dejstva natančneje opisujejo. Modele bomo napolnili s podatki iz Merkurjeve spletne trgovine in jih analizirali z orodjem Analysis Services, ki je del podatkovne zbirke Microsoft SQL Server 2000. V nalogi bo manjši poudarek na samem prenosu podatkov iz virov v podatkovno skladišče. Na koncu bomo ugotovili, kako lahko zbrani podatki pripomorejo k izboljšanju spletnega mesta in odločanju pri snovanju dopolnitev in izboljšav.

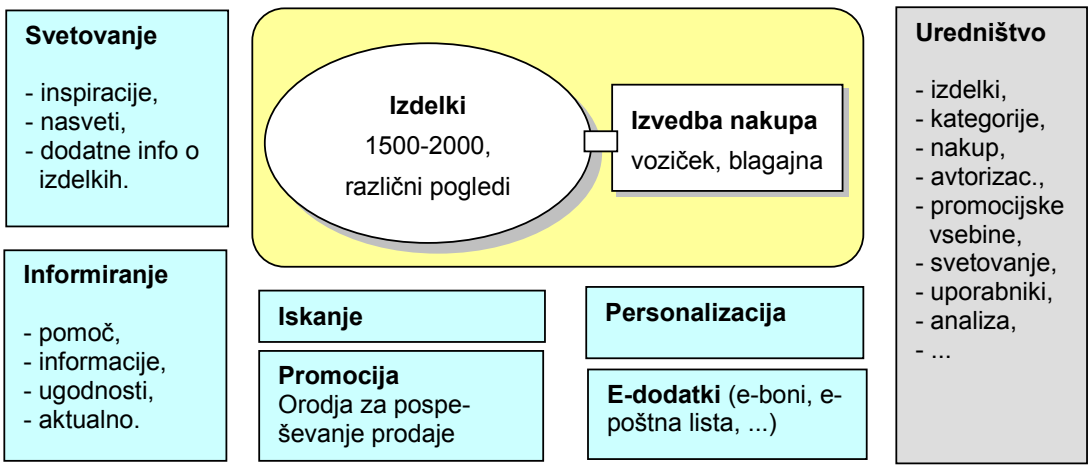
Diplomska naloga obsega osem poglavij. Po uvodnem poglavju bomo spoznali tipično spletno trgovino in predvsem parametre v njej, ki vplivajo na obnašanje obiskovalcev. V tretjem poglavju bomo natančno pregledali sledi, ki jih obiskovalci puščajo na spletnem mestu – te bodo namreč naš glavni vir podatkov. Četrto poglavje natančneje predstavlja dimenzijsko modeliranje, metodo, ki je zelo primerna za gradnjo analitičnih podatkovnih baz. V petem poglavju bomo pogledali, kateri podatki so dejansko na voljo v praktičnem primeru, Merkurjevi spletni trgovini. Nato bomo v šestem poglavju zasnovali podatkovno skladišče za te podatke, ga v sedmem poglavju napolnili s podatki in analizirali. Osmo poglavje povzame bistvene rezultate in pokaže možnosti praktične uporabe in nadaljnjega razvoja.

[2] Izraz *sledi (spletnih) obiskovalcev* bomo v tej nalogi uporabljali kot prevod besede *clickstream,* ki pomeni zapise dejanj spletnih obiskovalcev v dnevniku spletnega strežnika.

2. Spletna trgovina

V tej nalogi želimo analizirati uporabo spletne trgovine, izsledke pa uporabiti pri snovanju izboljšav in dopolnitev tako pri vsebini kot pri samem delovanju sistema. Za uspešno analizo moramo zato spoznati parametre, ki vplivajo na obnašanje obiskovalcev spletnega mesta. Informacije, podane v tem poglavju, so večinoma rezultat praktičnih izkušenj, pridobljenih ob izgradnji Merkurjeve spletne trgovine in drugih spletnih mest.

2.1. Vsebina tipične spletne trgovine



Slika 2.1. Shematski prikaz spletnega mesta nakup.merkur.si [15]

2.1.1. Nabor izdelkov

Jedro spletne trgovine predstavlja nabor izdelkov. Obiskovalci so zelo zahtevni: pričakujejo podoben obseg ponudbe kot v klasični trgovini, ob tem pa še cenovne ali druge ugodnosti, ki jih bodo prepričale v nakup po internetu.

Ob predstavitvi izdelkov v spletni trgovini moramo upoštevati, da so naši obiskovalci 'oropani' večine čutil: izdelek vidijo samo na zaslonu, ne v naravni velikosti, ne

morejo otipati materialov, vonjati, natančneje pogledati detajlov, običajno je težka tudi pot do dodatnih informacij (ki jih v klasični trgovini poiščemo pri najbližjem prodajalcu). Zato je priprava kakovostnih informacij o izdelkih nujna: preko tekstovnih, slikovnih in multimedijskih materialov spoznamo posamezne izdelke, dodatni atributi (npr. velikost zaslona, moč motorja, hitrost zapisovanja, ...) omogočajo primerjavo med več izdelki. Zelo jasno moramo označiti tudi osnovne, prodajalcu trivialne informacije: za kakšno količino velja navedena cena, v kakšnih količinah je mogoče izdelek sploh kupiti in podobno.

S pomočjo spletnih obiskovalcev lahko opise izdelkov dopolnimo z ocenami in mnenji.

Nabor izdelkov na internetu mora biti tak, kot ga obiskovalci od določene blagovne znamke pričakujejo – če nas v fizičnem svetu poznajo kot prodajalca čevljev, bo tako tudi pričakovanje ob obisku naše spletne trgovine. Dejstvu, da so določeni izdelki primernejši za prodajo po internetu kot drugi, sicer ne moremo uiti, a v vsakem primeru kaže spletno mesto izkoristiti vsaj kot orodje za informiranje potencialnih kupcev in ohranjanje lojalnosti obstoječih kupcev.

2.1.2. Urejenost izdelkov v skupine

Izdelki morajo biti primerno urejeni v skupine, če želimo, da bodo obiskovalci v spletni trgovini našli tisto, kar iščejo. Prednost interneta je, da je vsaka urejenost navidezna in enostavno spremenljiva – izdelkov ni potrebno fizično prenašati med policami. Iz tega izhaja tudi dodatna prednost: izdelke lahko poljubno združujemo v skupine (t.i. virtualne police), posamezen izdelek se lahko pojavlja na poljubnem številu le-teh. Kreiramo jih glede na trenutne zahteve in pričakovanja uporabnikov.

Vse spletne trgovine so za obiskovalce ob vstopu enako velike: zavzemajo eno okno v spletnem brskalniku. Zato moramo s pregledno razvrstitvijo polic poskrbeti, da bodo hitro dobili občutek, kaj jim ponujamo (in tudi, česa tu ne bodo dobili, da ne bodo po neuspešnem iskanju razočarani).

2.1.3. Iskalnik

Iskalnik je eno najpomembnejših orodij v spletni trgovini. Z njim predvsem pri velikem naboru izdelkov skrajšamo pot tistim obiskovalcem, ki že vsaj približno vedo, kaj jih zanima. Ti obiskovalci so tudi najbolj zanimivi za prodajalca, saj je pri njih verjetnost nakupa veliko večja kot pri drugih, ki se samo sprehajajo med policami.

Dober iskalnik poleg ključnih besed, ki se nahajajo v opisih izdelkov in kategorij, pozna sinonime in najpogostejše napačno črkovane besede. Rezultate iskanja zna urediti po več kriterijih, ravno tako zna izvesti iskanje samo znotraj dobljene množice rezultatov.

Zelo pomembne informacije upravitelju spletne trgovine razkrije statistika uporabe iskalnika: iz najpogosteje najdenih pojmov lahko izlušči najbolj iskane izdelke v določenem obdobju (in jih po potrebi še bolj izpostavi ali dopolni ponudbo), neuspešna iskanja pa pomagajo pri gradnji baze sinonimov in pri dopolnjevanju prodajnega programa.

2.1.4. Postopek nakupa, prijava

Večina truda, vloženega v pripravo spletne trgovine, je zaman, če kupec zaradi kakršnegakoli razloga izbranih izdelkov ne prinese do virtualne blagajne in ne zaključi nakupa. Razlogov za nedokončanje nakupov je precej, nekaterim najbolj kritičnim pa se lahko izognemo. Postopki ob zaključku nakupa morajo biti enostavni, pregledni in jasno opisani. Vidno morajo biti izpostavljeni najpomembnejši dejavniki nakupa, kot so cene, davki, pogoji in stroški dostave, možnost preklica in vračila. Od kupca zahtevamo le tiste informacije, ki so nujne za izvedbo in dostavo naročenih izdelkov. Sporočila o napakah pri vnosu morajo biti vidna in nedvoumna. Z enostavno prijavo omogočimo kupcem, da ob naslednjih nakupih samo vpišejo uporabniško ime in geslo, ostalih podatkov pa jim ni potrebno vnašati.

2.1.5. Svetovalno-izobraževalne vsebine

Velikokrat moramo kupca za nakup določenega izdelka šele navdušiti, pri čemer nam pomagajo svetovalni in izobraževalni članki. Pripravimo vsebine, kjer določene izdelke dodatno opišemo, primerjamo posamezne značilnosti, najpogosteje pa izhajamo iz življenjskih situacij uporabnika in ponudimo konkreten izdelek kot 'rešitev' za nek problem. Tovrstne vsebine imajo lahko odličen stranski učinek: naše spletno mesto lahko za zadovoljnega obiskovalca postane prva postaja, na katero se bo vrnil ob naslednjem problemu – in bo tako zopet prišel v stik z našim prodajnim programom.

2.1.6. Notranji promocijski elementi

Zaradi velike količine informacij mora dobra spletna trgovina vsebovati orodja, s katerimi lahko določene vsebine dodatno izpostavljamo. Ne moremo namreč pričakovati, da bodo obiskovalci po vrsti raziskovali police, izdelke in informativne vsebine. Zato na samih straneh oblikovalci predvidijo določen prostor, sistem v ozadju pa samodejno ali s pomočjo urednikov izbira aktualne vsebine.

Ob predstavitvi izdelka se tako pojavijo sorodni izdelki (dopolnilni in alternativni, običajno boljši – t.i. *cross-sell* in *up-sell)*, povezani informativni članki, tudi mnenja in ocene uporabnikov. Na prodajnih policah izpostavimo prodajne akcije, najpopularnejše izdelke, informativne vsebine. Ob svetovalno-izobraževalnih vsebinah priporočimo povezane izdelke ali kategorije. V celotni spletni trgovini pa izpostavi-

mo prodajne akcije, trenutno aktualne kategorije, aktualne svetovalne vsebine, zaradi narave spletnih kupcev (predvsem njihovega nezaupanja v medij) je pomembna tudi promocija prednosti nakupa po internetu in skrbi za varnost podatkov.

2.1.7. Ostali dejavniki

Na celovito uporabniško izkušnjo (t.i. *user experience)* ne vpliva samo spletna aplikacija in vsebina spletnega mesta. Za uporabnika so zelo pomembni tudi faktorji, kot so podpora (npr. po telefonu ali elektronski pošti), kakovostna dostava (območje dostave, hitrost, prilagodljiv čas), uspešno reševanje reklamacij in poprodajne aktivnosti.

Uspešna blagovna znamka v fizičnem svetu pomeni, da nas obiskovalci tudi na internetu že bolj poznajo in nam zaupajo. Manj znana podjetja (ali popolnoma nova spletna podjetja) morajo veliko več truda vložiti v prepričevanje obiskovalcev, da so vredna zaupanja.

Za spletno trgovino je pomembno pridobivanje novih obiskovalcev in ohranjanje stika z obstoječimi. Za nove obiskovalce poleg zelo učinkovite dobre besede zadovoljnih kupcev poskrbi oglaševanje (na spletu, v drugih medijih, v lastnih sredstvih, kot so katalogi, letaki, oglasi v fizičnih trgovinah itd.), z obstoječimi obiskovalci in kupci pa ohranimo stik predvsem prek elektronske pošte.

2.2. Upravljanje spletne trgovine

Spletna trgovina je dinamičen sistem. Stalno dopolnjujemo prodajni program in svetovalno-izobraževalne vsebine, občasno dodajamo tudi funkcionalne izboljšave. Za razvoj običajno skrbi uredniški odbor, ki se odloča na osnovi poročil o prodaji, obisku, učinkovitosti promocijskih akcij, pa tudi splošnih trendov v panogi.

Pri določanju smernic (predvsem na področju dopolnjevanja ali prilagajanja vsebin, navedenih v prejšnjih odstavkih) so nam lahko zelo v pomoč sledi, ki jih obiskovalci puščajo na našem spletnem mestu. Ob spremembah in dopolnitvah prodajnega programa uporabimo podatke o najbolj prodajanih in o največkrat iskanih izdelkih, razširimo ponudbo v najbolj gledanih kategorijah, upoštevamo vpliv prodajnih in promocijskih akcij. Za splošno izboljšanje delovanja poiščemo strani ali področja, s katerih obiskovalci največkrat zapustijo spletno mesto, pregledamo najbolj iskane vsebine in preverimo, ali so tudi brez iskalnika dovolj izpostavljene itd.

Praktičen problem, ki nastopi ob obdelavi uporabniških sledi, je pretvorba ogromnih količin razpoložljivih podatkov v obliko, ki bo primerna za analizo, in dejanska izvedba analize. V nadaljevanju naloge bomo pogledali podatke, ki so na voljo v spletni trgovini, in s pomočjo dimenzijskega modeliranja zasnovali praktičen model za njihovo ureditev in analizo.

3. Sledi obiskovalcev

Podatki, ki so na voljo v transakcijskih informacijskih sistemih podjetij, večinoma opisujejo le zadnji korak pri nekem poslu; v primeru trgovine je to izdan račun. Nikakršnih informacij nimamo o tem, zakaj je kupec izbral določene izdelke, ali je dobil tisto, kar je iskal, zakaj je nakup izvedel v danem trenutku, kolikokrat se je že prej oglasil pri nas in poizvedoval o izdelkih, zakaj se je odločil za našo trgovino itd.

V tem poglavju si bomo pogledali vir podatkov, ki je na voljo na spletu in razkriva odgovore na večino zgornjih vprašanj: sledi obiskovalcev *(clickstream)*. V dnevniku spletnega strežnika (in v nekaterih spremljajočih virih) so z mikroskopsko natančnostjo opisane vse aktivnosti uporabnikov. Aktivnosti posameznih uporabnikov združimo v seje, ki opisujejo en obisk spletnega mesta. Z analizo sej lahko ugotovimo, kakšne poti vodijo uporabnike do dejanj, ki nas posebej zanimajo, od nakupa do predčasnega odhoda s spletnega mesta.

Informacije v tem poglavju so večinoma povzete po viru [2].

3.1. Nameni in vedenjski vzorci

Samo zapis obiskovalčevih dejanj na spletnem mestu ni dovolj za analizo obnašanja, saj zaradi gostih dreves običajno ne moremo videti celega gozda. Zelo koristen podatek, ki dopolnjuje seznam posameznih dejanj, je obiskovalčev namen. Običajno lahko že na osnovi ogleda neke strani prepoznamo namene, kot so:

- zbiranje splošnih informacij,
- zbiranje informacij o izdelku,
- pregled pogostih vprašanj in odgovorov,
- reševanje specifičnih težav,
- naročilo izdelka ali storitve,
- pregled statusa naročila,
- iskanje,
- branje novic, nasvetov, člankov,
- zabava,
- prenos datotek,
- kontakt.

Če namene, ki smo jih razbrali iz obiskov posameznih strani, pogledamo celovito, lahko ugotovimo tudi bolj splošne vzorce obnašanja obiskovalcev, npr.:

- uspešen nakup,
- prekinjen ali neuspešen nakup,
- našel iskane informacije,
- ni našel iskanih informacij,
- *session killer* – dogodek, po katerem obiskovalec zapusti spletno mesto,
- prehod med stranmi po napačni poti,
- jezen ali zadovoljen obiskovalec.

Čeprav je že samo ugotavljanje vedenjskih vzorcev zanimivo, se prava vrednost njihove analize skriva v izboljšanju celotne interakcije (ne samo spletnega mesta) med obiskovalcem in našo organizacijo. Ta posledično vpliva na povečano učinkovitost spletnega mesta, večjo lojalnost kupcev, povečan prihodek in dobiček.

3.2. Kako nam sledi obiskovalcev pomagajo pri odločitvah?

Zaradi ogromnih količin podatkov in neskončnih možnosti analiziranja je pomembno, da se že dovolj zgodaj vprašamo, kaj je cilj našega raziskovanja sledi obiskovalcev. Samo tako se lahko usmerimo in ostanemo na pravi poti ves čas analize.

Končni cilj našega podatkovnega skladišča je podpora odločitvam, povezanimi tako z vsakdanjim delovanjem kot z večjimi strateškimi spremembami spletnega mesta. V nadaljevanju so naštete nekatere od takih odločitev, ki jih lahko podpremo z analizo sledi obiskovalcev (včasih tudi v povezavi z drugimi viri podatkov).

Prepoznavanje obiskovalcev

Obiskovalcem lahko glede na njihovo 'zgodovino' ponudimo prilagojeno vsebino, od izdelkov ali člankov (glede na področja zanimanja, ki smo jih razbrali iz dosedanjih obiskov) do prilagojenih cen; ravno tako lahko bolj učinkovito usmerimo marketinške aktivnosti. Nove obiskovalce lahko obravnavamo drugače, jim ponudimo pomoč ali enostavnejšo različico spletnega mesta. Podatki o obiskovalcih nam pomagajo tudi pri ocenjevanju zunanjih povezav – ugotovimo namreč lahko, da nam določena spletna mesta prinašajo bolj 'kakovostne' obiske kot druga. Nenazadnje lahko obiskovalce, ki so nas zapustili (nekdaj pogosti obiskovalci, ki nas v zadnjem času ne obiskujejo), ponovno povabimo k obisku ter vprašamo za razlog odhoda.

Komuniciranje

Spletno mesto je predvsem komunikacijsko orodje. Komunikacija je obojestranska: na pripravljeno vsebino se obiskovalci odzivajo s klikanjem, lahko tudi izpolnjujejo spletne obrazce. Zelo hitro lahko iz sledi obiskovalcev v podatkovnem skladišču ocenimo učinkovitost spletnih oglasov, analiziramo lahko tudi vpliv promocijskih akcij. Vsebino lahko prilagodimo posebnim priložnostim (sezonske spremembe, spremembe v življenju obiskovalcev). Ugotovimo lahko tudi precej dejstev, povezanih s splošno učinkovitostjo spletnega mesta: kakšne strani (tako oblikovno kot vsebinsko) obiskovalci bolje sprejemajo, ali so najpogostejša opravila res hitro pri roki, na katere strani se obiskovalci pogosto vračajo, katere najhitreje zapustijo.

Poslovanje

Podatkovno skladišče nam hitro da odgovor na vprašanje, katere izdelke ali storitve dobro prodajamo preko interneta in katerih ne. Ugotovimo lahko, katere opise obiskovalci berejo, ali je kakšna povezava med kakovostjo opisa in nakupno odločitvijo, koliko klikov potrebujejo do nakupa ponovni in koliko novi obiskovalci, koliko začetih nakupov obiskovalci ne dokončajo itd. Pod črto lahko v povezavi s podatki o stroških ugotovimo dobičkonosnost posameznih skupin uporabnikov, skupin izdelkov in storitev, promocijskih akcij in na koncu dobičkonosnost celotnega spletnega mesta.

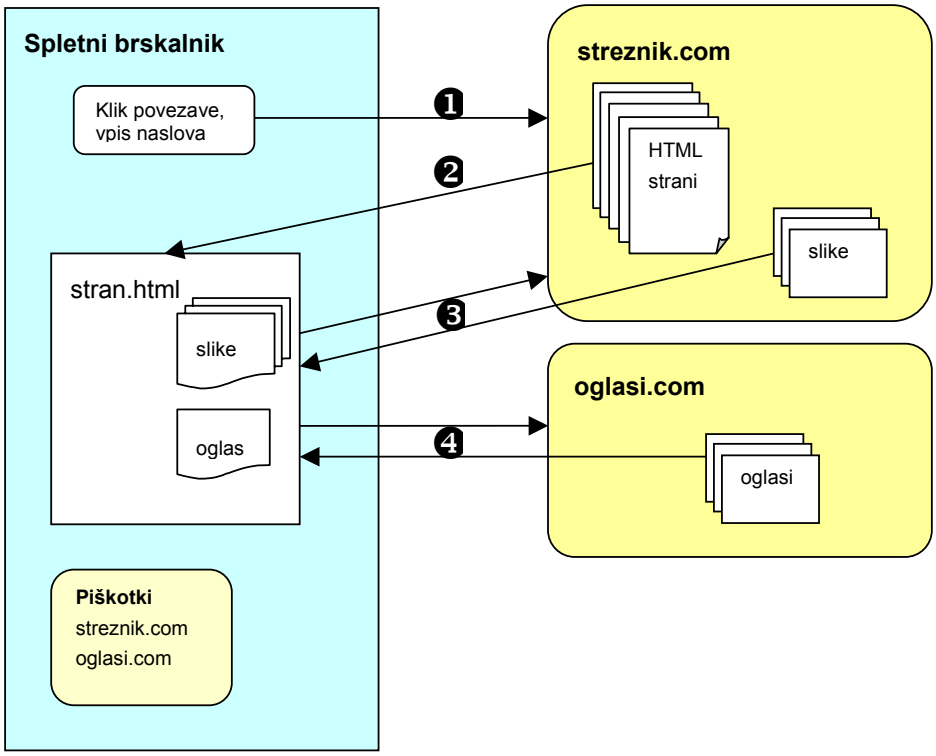
3.3. Kako komunicirata spletni strežnik in odjemalec – brskalnik?

Glavni vir podatkov, ki jih bomo v nalogi uporabili, je dnevnik spletnega strežnika. Za boljše razumevanje podatkov v njem bomo v tem razdelku na kratko opisali komunikacijo med strežnikom in odjemalcem – spletnim brskalnikom. Postopek je natančneje prikazan na sliki 3.1.

Komunikacija poteka po protokolu HTTP (HyperText Transfer Protocol). Ob vpisu novega naslova ali kliku na povezavo brskalnik posreduje zahtevo za novo stran strežniku (1). Ta vrne zahtevani dokument (2). Ko ga odjemalec v celoti sprejme (običajno je to HTML koda, ki opisuje stran), v njem poišče kazalce na dodatne elemente (slike, stilske predloge, skripte) in po vrsti (lahko zaradi večje hitrosti tudi v več vzporednih procesih) pošlje zahteve strežniku (3); ta mu vrne prave elemente. Osnovna stran lahko vsebuje tudi elemente z drugih strežnikov; med najpogostejšimi tovrstnimi vsebinami so oglasne pasice. Tudi tem strežnikom brskalnik pošlje zahtevo in nato sprejme zahtevani element (4).

Ob sprejemu zahteve za posredovanje strani strežnik lahko tudi prebere piškotke (t.i. *cookije,* natančneje so opisani v nadaljevanju), shranjene v brskalniku, in/ali vanje

zapiše določene informacije, ki so običajno namenjene kasnejšemu prepoznavanju uporabnika.



Slika 3.1. Komunikacija med spletnim brskalnikom (odjemalcem) in spletnim strežnikom

3.4. Sledi v dnevniku spletnega mesta

Sledi obiskovalcev na spletnem mestu so zapisane v dnevniku spletnega strežnika. Ta si podrobno zabeleži vse HTTP zahteve – tiste za prikaz strani in tudi ostale, s katerimi uporabnikov brskalnik zahteva dodatne elemente strani (slike, multimedijske vsebine itd.). Vsaka zahteva predstavlja eno vrstico dnevnika. Strežnik zapise v dnevnik dodaja glede na časovno zaporedje, zato je potrebno kar nekaj (programske) spretnosti, da združimo zapise, ki tvorijo eno uporabnikovo sejo oziroma en obisk.

Na sliki 3.2 je izsek tipičnega dnevnika spletnega strežnika. Zaradi preglednosti smo podatke delno uredili – posamezni zapisi so razdeljeni na več vrstic.

```
- 2002-05-01 09:23:39 217.72.71.186 - NAKUPWEB GET
  /pic/promocija/ozadje-pikce.gif - 200 448 390 968 HTTP/1.1
  Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98;+Win+9x+4.90)
  VisitorId=VRURAXJG6MMYHCMNQY4G;+ASPSESSIONIDQQGQQHAY=GAMEJFJCINLIPMLBGOOCHPLG
  http://nakup.merkur.si/home.asp?MySes=39SJVXYT1VYKQ1Y3U0FT
- 2002-05-01 09:23:40 217.72.71.186 - NAKUPWEB GET
  /katalog-kop.asp MySes=39SJVXYT1VYKQ1Y3U0FT 200 20743 561 422 HTTP/1.1
  Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98;+Win+9x+4.90)
  VisitorId=VRURAXJG6MMYHCMNQY4G;+ASPSESSIONIDQQGQQHAY=GAMEJFJCINLIPMLBGOOCHPLG
  http://nakup.merkur.si/home.asp?MySes=39SJVXYT1VYKQ1Y3U0FT
- 2002-05-01 09:23:40 193.90.144.165 - NAKUPWEB GET
  /oddelek.asp id=0101&amp;MySes=8T9ORHMN2ACGUFP7KT63 200 25897 330 860 HTTP/1.0
  Mozilla/4.0+(HIH+-+BabelServer+1.01) - -
- 2002-05-01 09:23:43 193.90.144.165 - NAKUPWEB GET
  /oddelek.asp id=0103&amp;MySes=8T9ORHMN2ACGUFP7KT63 200 27207 330 547 HTTP/1.0
  Mozilla/4.0+(HIH+-+BabelServer+1.01) - -
- 2002-05-01 09:23:43 194.249.41.111 - NAKUPWEB GET
  /oddelek.asp id=0204&MySes=90NGKMGBD594GMX8DQH6 200 22439 598 469 HTTP/1.1
  Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98;+Win+9x+4.90;+Hotbar+3.0)
  ASPSESSIONIDQQGQQHAY=OOKEJFJCJEANDFGFNEBFHBHO;+VisitorId=1JQNYN1ZIWYI8WLS9N6T
  http://nakup.merkur.si/search.asp?isciKaj=hladilniki&isciKje=2&MySes=90NGKMGBD594GMX8DQ
- 2002-05-01 09:23:44 194.249.41.111 - NAKUPWEB GET
  /nakup2-mojster.css - 200 1201 401 156 HTTP/1.1
  Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98;+Win+9x+4.90;+Hotbar+3.0)
  ASPSESSIONIDQQGQQHAY=OOKEJFJCJEANDFGFNEBFHBHO;+VisitorId=1JQNYN1ZIWYI8WLS9N6T
  http://nakup.merkur.si/oddelek.asp?id=0204&MySes=90NGKMGBD594GMX8DQH6
- 2002-05-01 09:23:44 217.72.71.186 - NAKUPWEB GET
  /pic/home-detajl-dz3.gif - 200 386 390 328 HTTP/1.1
  Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98;+Win+9x+4.90)
  VisitorId=VRURAXJG6MMYHCMNQY4G;+ASPSESSIONIDQQGQQHAY=GAMEJFJCINLIPMLBGOOCHPLG
  http://nakup.merkur.si/katalog-kop.asp?MySes=39SJVXYT1VYKQ1Y3U0FT
```

Slika 3.2. Urejen izsek iz dnevnika spletnega strežnika

Vsebina dnevnika je delno standardizirana (formata *CLF – Common Log Format* in *ECLF – Extended Common Log Format* določata osnovna polja zapisa), delno pa je odvisna od nastavitev in dodatnih zmožnosti spletnega strežnika. V naslednjih odstavkih so opisani najpomembnejši podatki, shranjeni v dnevniku (povzeto po [2]).

Strežnik (Host)

Spletni naslov (IP naslov, redkeje ime domene) brskalnika ali drugega agenta, ki je sprožil HTTP zahtevo. S pomočjo IP naslova lahko običajno ugotovimo obiskovalčevo domeno, ki včasih razkrije tudi lokacijo (državo) ali organizacijo uporabnika. Zaradi velikega števila klicnih dostopov do interneta z dinamično dodeljenimi IP naslovi (ob klicu dobi uporabnik enega od trenutno prostih naslovov, isti naslov se po sprostitvi klica lahko dodeli nadaljnjim uporabnikom) in dostopa izza požarnih zidov podjetij (v tem primeru so pogosto vsi računalniki navzven predstavljeni z istim IP naslovom) popolna identifikacija obiskovalcev samo s pomočjo IP naslova ni mogoča. Če nimamo boljšega orodja, pa lahko predpostavimo, da so zahteve, ki v določenem časovnem obdobju prihajajo z istega IP naslova, del ene uporabniške seje.

Čas (Time)

Čas, ko je strežnik prejel zahtevo. Nekateri strežniki namesto tega zapisujejo čas, ko so zahtevo izpolnili. Običajno v formatu [dd/Mmm/yyyy:hh:mm:ss zone].

Zahteva (Request)

Najpomembnejši del strežnikovega zapisa: vsebuje ime elementa (URI – Uniform Resource Identifier), ki ga uporabnik zahteva, ob tem še oznako metode (najpogosteje GET in POST) in protokola (npr. HTTP/1.0), ki sta pomembna za sporazumevanje strežnika in uporabnikovega agenta.

Status (Status)

Trištevilčna koda, s katero strežnik brskalniku sporoči status zahteve. Najpogostejši vrednosti sta 200 (OK) in 404 (stran ne obstaja).

Velikost (Bytes)

Velikost strežnikovega odgovora v bajtih.

Vir (Referrer)

Brskalnik v tem polju sporoči strežniku naslov strani, s katere je prišla zahteva. Zelo koristen podatek pri analizi povezav med stranmi in spletnimi mesti.

Odjemalec (User-Agent)

Oznaka spletnega brskalnika ali drugega odjemalca, ki je strežniku posredoval zahtevo. Vsebuje ime programa, verzijo in operacijski sistem.

Piškotek (Cookie)

Vsebina piškotka, veljavnega v tej seji. Več o piškotkih je napisano v nadaljevanju.

Obstaja kar nekaj programskih rešitev za analizo dnevnikov spletnih strežnikov (med bolj poznanimi sta izdelka *WebTrends* in *FastStats)*. Tovrstni programi omogočajo koristen vpogled v gibanje števila obiskovalcev, obiskanost posameznih strani, iskanje najpogostejših vstopnih in izstopnih točk in druge zbirne statistične pokazatelje. Zaradi njihove splošne narave, omejenosti zgolj na podatke iz dnevnika spletnega strežnika in nepoznavanja zakonitosti v ozadju spletnega mesta pa nam običajno ne morejo pomagati niti pri natančnejši analizi obiskanosti vsebin niti pri analizi obnašanja obiskovalcev.

3.5. Kako prepoznati obiskovalca?

Med zahtevnejšimi cilji, ki jih imajo pred očmi snovalci sodobnih spletnih mest, je prilagoditev vsebine prikazanih strani posameznemu obiskovalcu. Temeljni pogoj za

to je, da obiskovalca prepoznamo – samo tako mu bomo lahko glede na njegovo dosedanjo aktivnost ponudili vsebino, za katero menimo, da ga najbolj zanima, ob tem pa podatke o vseh obiskih združevali v celoto, ki nam bo dala še boljšo sliko o njegovih navadah in potrebah.

Uporabniki interneta zelo neradi razkrijejo svojo identiteto. Glavni vzroki so nezaupanje (do interneta, do upraviteljev spletnih mest), strah pred zlorabami (elektronskega naslova, osebnih podatkov, številke plačilne kartice) in nelagodje pred dejstvom, da bi lahko njihovo početje zasledovali (in kasneje povezali s konkretnim imenom). V primeru, ko od obiskovalcev brez konkretnega razloga zahtevamo vpis podatkov ali elektronskega naslova, jih bo vsaj polovica podala neresnične podatke. [2]

Več kot očitno torej potrebujemo mehanizem, ki bo omogočal bolj ali manj prikrito označevanje spletnih obiskovalcev. HTTP protokol, ki je osnova za dogodke v dnevniku spletnega strežnika, ima s tega stališča veliko pomanjkljivost: nikjer ni razvidno, katere zahteve sodijo skupaj v eno uporabnikovo sejo, ravno tako je nemogoče prepoznati različne seje istega uporabnika. Programsko sicer lahko serije zahtev, ki prihajajo v določenem časovnem zaporedju z istega naslova, združimo v posamezne seje, a s tem ne presežemo anonimnosti uporabnikov.

3.5.1. Piškotki

Kot rešitev za to pomanjkljivost je bila vpeljana izmenjava piškotkov (t.i. *cookies)*. Mehanizem omogoča spletnemu strežniku, da v obiskovalčevem brskalniku shrani kratek tekstovni zapis in ga kadarkoli kasneje tudi prebere. Piškotki so lahko stalni *(persistent)*, shranjeni na disku, ali začasni *(session level)*, ki jih brskalnik hrani v spominu, dokler programa ne zapremo. Zaradi varovanja zasebnosti lahko posamezne piškotke pišejo, berejo in spreminjajo le strežniki iz domene, ki je vpisana v samem piškotku. Mehanizem piškotkov se najpogosteje uporablja ravno za enolično označevanje uporabnikov.

Piškotek vsebuje naslednja polja: ime, vrednost (dejansko vsebino), domeno in pot (do vsebine piškotka lahko prek spleta dostopajo samo aplikacije iz navedene domene in poti), rok veljavnosti ter varnostno oznako, ki pove, ali naj brskalnik in strežnik vsebino piškotka izmenjata po varni (SSL) povezavi.

V odvisnosti od zmožnosti spletnega strežnika, nastavitev brskalnika in nenazadnje pripravljenosti obiskovalca, da se nam predstavi, lahko obiskovalce prepoznamo bolj ali manj natančno. Najmanj nam pomaga podatek o eni seji, ki prihaja z določenega določenega računalnika in spletnega brskalnika na njem – ne moremo namreč prepoznati ponovnih obiskov. Nekaj boljši možnosti sta prepoznavanje ponovnega obiska z istega računalnika in prepoznavanje ponovnega obiska neke

osebe; naš končni cilj pa je natančno prepoznati, kdaj se je na spletno mesto vrnil poznan obiskovalec.

3.5.2. Drugi načini prepoznavanja

Predvsem zaradi zahtev po varstvu zasebnosti nikoli niso zaživele ideje o uporabi enoličnih oznak računalnikov, čeprav te v praksi že obstajajo: vsaka mrežna kartica ima že več kot 20 let svoj strojni naslov *(hardware Ethernet address)*, tudi novejši procesorji imajo enolično določene serijske številke.

Spletna mesta, pri katerih je enolična identifikacija uporabnika tudi v interesu slednjega (elektronsko bančništvo, elektronsko poslovanje med podjetji in podobne aplikacije), v zadnjem času uporabljajo za prepoznavanje in kodiranje podatkov digitalne certifikate. Tudi uporaba teh sredstev v našem primeru, ko uporabniki praviloma ne želijo razkriti svoje identitete, ni mogoča.

3.6. Prilagoditev spletnega mesta za enostavnejšo analizo

Na tej točki bomo ob poznavanju vsebine dnevnika spletnega strežnika opisali, kako lahko tudi z ustrezno zasnovo spletnega mesta pomagamo, da bo kasnejša analiza obiskov enostavnejša.

3.6.1. Označevanje strani

Klasifikacija strani je redko prioriteta ob nastajanju spletnega mesta – takrat so vse moči usmerjene v pripravljanje in polnjenje vsebin ter odkrivanje slepih in napačnih povezav. Kljub temu moramo strani spletnega mesta označiti in razporediti v kategorije, ki nam bodo pomagale pri analizi vedenja obiskovalcev, saj samo iz zahtev za prikaz strani, zapisanih v dnevniku strežnika, le redko lahko ugotovimo uporabnikov namen.

Iz spodnjih dveh praktičnih primerov vidimo, da sam zahtevek običajno ne pove kaj dosti o vsebini strani. Prvi zahtevek sicer razkriva nekaj podrobnosti (gledamo oddelek s kodo 020603), drugi pa čisto nič:

```
http://nakup.merkur.si/oddelek.asp?ID=020603
http://www.nlb.si/cgi-bin/nlbweb.exe?doc=502
```

Naivno bi bilo torej pričakovati, da bomo ob prenosu podatkov iz dnevnika v podatkovno skladišče na osnovi URL zahtevkov znali določiti tudi vrsto strani. Podatke moramo pripraviti vnaprej in jih ob razvoju spletnega mesta tudi vzdrževati.

Osnovna lastnost, ki nas o določenem zahtevku zanima, je tip strani. Čeprav so možni tipi odvisni od posameznega spletnega mesta, lahko običajno izhajamo iz naslednjih skupin:

- vstopna stran,
- osnovna stran,
- pregled prodajnega programa,
- seznam izdelkov,
- detajli o izdelkih,
- informativne vsebine,
- novice,
- predstavitev podjetja.

Za analizo je nujno potreben tudi dopolnilni podatek o konkretnem elementu (izdelku, kategoriji, članku, novici itd.), ki je bil prikazan.

3.6.2. Zapisovanje dodatnih informacij

Večkrat že ob snovanju spletnega mesta vemo, da nekaterih informacij ne bomo mogli pridobiti iz dnevnika ali iz drugih virov. Tako denimo v dnevniku spletnega strežnika ni natančnih podatkov o izvedenem nakupu (vrednost nakupa, številka računa), s pomočjo katerih bi lahko spletne podatke bolje ovrednotili ali povezali s transakcijskimi bazami. Takrat si lahko s pomočjo spletne aplikacije ustvarimo dodaten vir informacij tako, da ob določenih dogodkih sami zapisujemo ustrezne podatke v lasten dnevnik. Zapis nakupa v dnevniku lahko izgleda takole:

```
10.08.2002:15:20:25; 4536271890; 5210; 83590.20; fdagcdebebadeb
```

V poljih so navedeni datum in ura nakupa, oznaka uporabnika (iz piškotka), številka naročila, znesek naročila in oznaka seje.

Alternativna možnost, za katero ne potrebujemo niti dodatnega dnevnika, je proženje zahtev do spletnega strežnika po 'praznih' vsebinah, kjer v samem zahtevku skrijemo dodatne informacije. Primer shranjevanja istih informacij kot v zgornjem primeru:

```
<img src="null.gif?type=nakup&t=10082002-152025&u=4536271890&
     n=5210&v=8359020&s=fdagcdebebadeb" width=0 height=0 alt="">
```

Od strežnika v bistvu zahtevamo prazno sliko, ki jo vrne v minimalnem času neodvisno od ostalih parametrov, v dnevnik pa si vpiše našo celotno zahtevo. Ta je kasneje na voljo za analizo.

V obeh primerih lahko ob prenosu podatkov v podatkovno skladišče na osnovi skupne oznake (oznake seje) podatke priključimo k ostalim podatkom, povezanim z dano uporabniško sejo.

3.6.3. Označevanje uporabnikov

Osnovni pogoj za prepoznavanje obiskovalcev, ki se na naše spletno mesto vrnejo, je njihovo označevanje. Zaželeno je, da obiskovalca označimo samo enkrat in nato iz te oznake izpeljemo vse akcije; celotno spletno mesto (ali širše vsa spletna mesta korporacije) naj imajo enotno politiko in osrednji mehanizem dodeljevanja oznak obiskovalcem.

Za označevanje v praksi uporabimo trajne piškotke (persistent cookies). V izogib morebitnim kasnejšim problemom naj bo vsebina kodirana in zaščitena s kontrolno vsoto, da enostavno lahko izločimo podatke, ki so bili ročno (ali programsko) spremenjeni na strani uporabnika.

Med dodatne zahteve glede prilagoditve spletnega mesta sodi tudi sinhronizacija sistemskega časa strežnikov, če spletno mesto teče na več strežnikih in bomo morali podatke z njih združevati v enotne seje.

3.7. Ostali viri podatkov o obiskovalcih

Analiza podatkov o obisku spletne trgovine in navadah obiskovalcev bo bolj temeljita, če v podatkovnem skladišču zapisom iz spletnega dnevnika dodamo informacije iz drugih virov. Ti so lahko zelo raznoliki:

- podatki o nakupih v fizičnih trgovinah,
- podatki o vračilu blaga in reklamacijah,
- podatki o drugih kontaktih (povpraševanje, klicni center za podporo),
- demografski podatki.

Šele združeni podatki so res dobra osnova za segmentacijo kupcev, prepoznavanje dobičkonosnih obiskovalcev in v končni fazi individualno prilagojeno ponudbo.

4. Dimenzijsko podatkovno skladišče

V tem poglavju bomo spoznali dimenzijsko modeliranje – tehniko logičnega urejanja podatkov, ki nas pripelje do enotne logične sheme in logični model za organizacijo velike količine podatkov, denimo v podatkovnem skladišču. Vsebina tega poglavja se navezuje na eno temeljnih del s tega področja [1].

4.1. Podatkovno skladišče

Ralph Kimball takole definira podatkovno skladišče:

> **»Podatkovno skladišče je kopija transakcijskih podatkov, posebej strukturirana za izvajanje poizvedb in analiz.«**

V praksi so podatkovna skladišča največkrat zbirke podatkov, namenjene podpori odločanju, tako strateškemu kot operativnemu. V velikih količinah podatkov, uvoženih in predelanih iz transakcijskih sistemov *(OLTP – on line transaction processing)* in drugih virov (v našem primeru je to dnevnik spletnega strežnika), se običajno izvajajo poizvedbe, ki obdelajo veliko število zapisov, kot rezultat pa praviloma vrnejo manjše število zbirnih vrstic.

Uporabniki lahko podatkovno skladišče učinkovito uporabljajo, če natančno razumejo logično strukturo in relacije med podatki. Ob tem je mogoče poizvedbe, ki temeljijo na podatkovnih shemah z enotno strukturo, bolje optimizirati in zato hitreje izvesti kot tiste, ki jih izvajamo nad poljubnimi entitetami s poljubnimi odnosi med njimi. Predvsem prva potreba – narediti tako shemo podatkovne baze, da jo bodo uporabniki razumeli in bodo mogli v njej izvajati poljubne poizvedbe – je pripeljala do enostavnih podatkovnih modelov, ki jih oblikujemo s tehniko dimenzijskega modeliranja.

4.2. Zakaj dimenzijsko modeliranje?

Z razcvetom relacijskih podatkovnih zbirk so se v začetku osemdesetih let pojavili entitetni podatkovni modeli. Temeljijo na odpravljanju redundanc; glavni cilj postopka normalizacije je, da se vsak podatek v bazi pojavi samo enkrat. Tovrstni modeli so zelo prikladni za obdelavo velikega števila transakcij, saj v normaliziranem modelu vsaka transakcija praviloma vpliva le na minimalno število tabel in zapisov v bazi.

Ima pa tovrstna zasnova podatkovne baze tudi veliko slabost. V iskanju čim večje učinkovitosti procesiranja transakcij smo se precej oddaljili od enega najpomembnejših ciljev – nastale so podatkovne baze, v katerih je praktično nemogoče delati poizvedbe. Entitetni modeli so postali zelo kompleksni; vrhunski poslovni informacijski sistemi, kot je SAP, vsebujejo več tisoč logičnih entitet, ki ob implementaciji postanejo samostojne tabele. Zaradi take kompleksnosti končni uporabniki entitetnih modelov ne morejo razumeti ali si jih zapomniti, niti ne obstajajo splošni grafični vmesniki, ki bi modele približali uporabnikom in jim omogočili enostavne poizvedbe. Na drugi strani programi poizvedb ne znajo dobro optimizirati, saj iz modela niso vidne dejanske relacije med podatki. Uporaba entitetnih modelov pravzaprav onemogoči bistvo podatkovnih skladišč – intuitiven in hiter dostop do informacij.

Ob spoznanju, da končnim uporabnikom s tako kompleksnimi shemami ne morejo dosti pomagati, so se snovalci podatkovnih modelov začeli zatekati k enostavnejšim rešitvam. Novi, poenostavljeni modeli so si bili večinoma presenetljivo podobni in izkazalo se je, da jih lahko skoraj vse pojmujemo kot dimenzijske modele. [3]

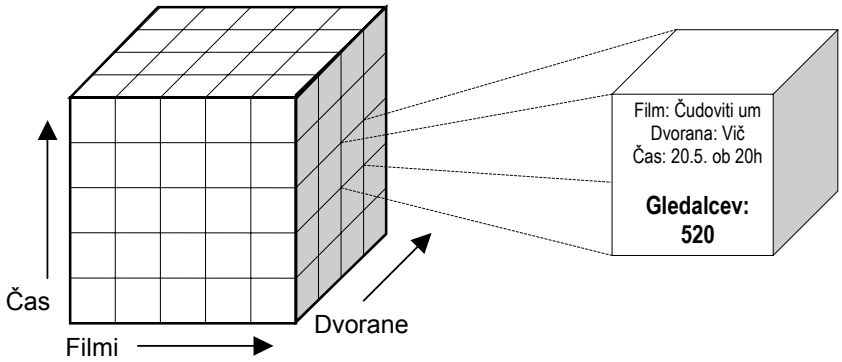
Ker je cilj take učinkovite podatkovne baze čim večja usmerjenost k uporabniku, v središču modela niso entitete ampak posamezni poslovni procesi. Dimenzijsko modeliranje zaradi take usmerjenosti ni nič manj uporabno od kompleksnih entitetnih modelov – razumeti moramo le, da entitetno shemo podatkovne baze podjetja lahko pretvorimo v množico enostavnih dimenzijskih modelov, ki opisujejo posamezne procese.

4.3. Osnove dimenzijskega modeliranja

Dimenzijsko modeliranje je tehnika, pri kateri podatke oblikujemo v standardnem, intuitivnem ogrodju, ki omogoča hiter dostop in poizvedbe. Vsak dimenzijski model opisuje svoj proces. Sestavljen je iz osrednje tabele večdelnim ključem (tabela dejstev – *fact table)* in iz množice manjših tabel, od katerih vsaka vsebuje po en ključ osnovne tabele (dimenzijske tabele – *dimension tables)*. Različni dimenzijski modeli so lahko povezani med seboj v celoto preko skupnih dimenzij *(conforming dimensions)*, ravno tako imajo lahko dejstva v različnih modelih enak pomen.

Za ilustracijo dimenzijskega podatkovnega modela bomo uporabili enostaven model, ki vsebuje podatke o gledanosti filmov v kinematografih. Predpostavljamo, da na različnih lokacijah predvajamo različne filme in dogajanje spremljamo skozi čas.

Podatke v dimenzijskem modelu si najlažje predstavljamo kot večdimemzionalno kocko. Stranice te podatkovne kocke so posamezne dimenzije, na presečiščih dimenzij pa so posamezna dejstva.



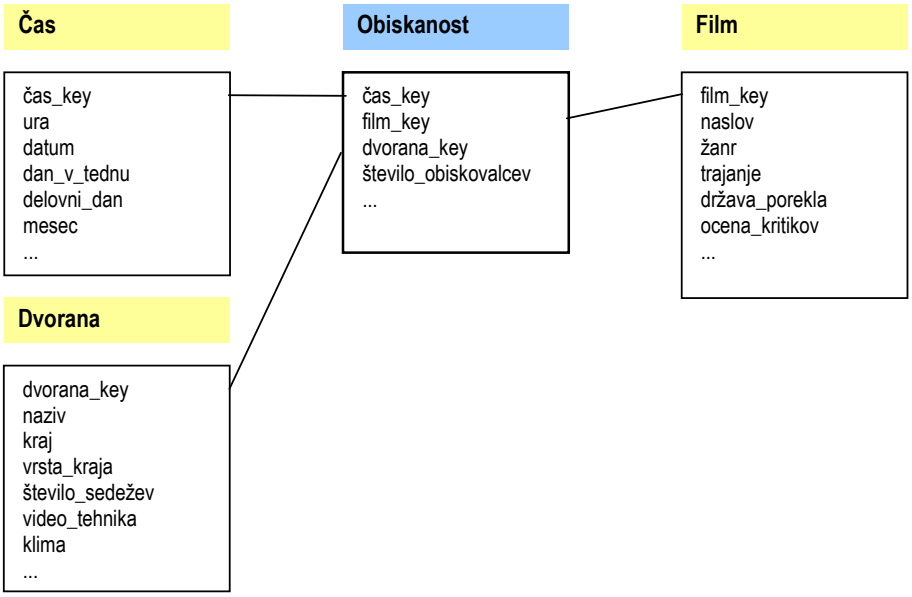
Slika 4.1. Podatkovna kocka: vsaka točka kocke vsebuje dejstva za dano kombinacijo dimenzij

Proces, ki ga modeliramo, je obisk predstav. Osrednje dejstvo, ki nas zanima, je število obiskovalcev na posamezni prestavi. Predstavo definirajo čas, lokacija (dvorana) in film. Za potrebe analize potrebujemo kakovostne informacije o vseh dimenzijah. Tako lahko filme opremimo z atributi, kot so žanr, dolžina, država porekla, ocena kritikov in podobno. Pri dvoranah na obisk gotovo vpliva tehnična opremljenost, število sedežev, lokacija (večji kraj, manjši kraj, nakupovalno središče, ...). Zelo zanimiva je tudi časovna dimenzija – poleg običajnih podatkov (datum, ura) jo opremimo z značilnostmi, kot so dan v tednu, praznik, vikend, celo vreme.

Podatkovni model za opisani proces je prikazan na sliki 4.2, v nadaljevanju pa so predstavljeni njegovi posamezni elementi.

4.3.1. Tabela dejstev

V tabeli dejstev so shranjene vrednosti, povezane z osnovnimi dejstvi *(grain)* procesa, ki ga modeliramo. Posamezni zapisi vsebujejo vrednosti na presečiščih vseh dimenzij. (V primeru obiskanosti predstav nas tako predvsem zanima število obiskovalcev ob nekem času v neki dvorani za nek film.)



Slika 4.2. Dimenzijski podatkovni model za opis obiskanosti kinopredstav

Aditivna, semiaditivna in neaditivna dejstva

Zapisi v tabeli dejstev poleg ključev za vse dimenzije vsebujejo posamezne vrednosti – dejstva *(facts)*. Najuporabnejša so aditivna numerična dejstva, katerih vrednost se v odvisnosti od dimenzij spreminja *(continuously valued)*. Le ob takih dejstvih so mogoče poizvedbe, ki iz številnih zapisov pripravijo seštevke (tudi povprečja in druge izračune) ter omogočajo primerjave. Take poizvedbe predstavljajo glavni del aktivnosti v podatkovnih skladiščih.

Semiaditivna dejstva so tista, pri katerih seštevki dajo smiselne rezultate samo v določenih dimenzijah, v drugih pa ne. Primer tovrstnih dejstev so posnetki stanj (stanje na bančnem računu, zaloga določenega izdelka). Seštevek stanj na računih strank v določenem trenutku je pomemben podatek, večina drugih seštevkov (npr. seštevek stanj določene stranke skozi čas) pa nima praktičnega pomena.

Še najmanj koristna so neaditivna dejstva, saj nad njimi ne moremo izvajati nobene računske operacije; edino, kar lahko z njimi storimo, je izpis posameznih vrednosti in ročen pregled ali analiza, kar pa ob velikih tabelah dejstev ni preveč koristno početje.

V praksi tabela dejstev običajno vsebuje le delček vseh možnih kombinacij (presečišč) dimenzij. Zapisujemo namreč samo tista dejstva, ki so se dejansko zgodila. (V našem primeru torej v tabeli dejstev niso vpisane ničle za obisk filma, ki ga ob neki uri v neki dvorani ni bilo.)

Tabela dejstev brez dejstev

Ob modeliranju procesov lahko včasih ugotovimo, da v tabeli dejstev ob ključih, ki določajo dimenzije, drugi podatki pravzaprav niso potrebni. Tipičen primer tovrstnih procesov je evidentiranje dogodkov (npr. prisotnosti študentov na posameznih predavanjih). Dimenzijski modeli s tovrstnimi tabelami (t.i. *factless fact tables)* ohranjajo vse splošne lastnosti in so ravno tako primerni za opravljanje poizvedb. Zaradi enostavnosti včasih v tabelah dejstev uvedemo numerično polje, ki ima konstantno vrednost 1, saj tako lahko namesto štetja v SQL stavkih uporabljamo enostavnejše seštevanje.

Zbirna dejstva

Ob poznavanju zahtev končnih uporabnikov lahko tipične poizvedbe znatno pohitrimo z uporabo seštevkov *(aggregates)*. Vnaprej namreč lahko pripravimo zapise s seštevki aditivnih dejstev preko izbranih dimenzij. (V našem primeru obiska kinopredstav bi lahko sešteli mesečni obisk neke dvorane, dnevni in mesečni obisk nekega filma v vseh dvoranah in podobno.) Zbirna dejstva lahko shranimo na dva načina:

- uvedemo dodatne tabele dejstev, ki vsebujejo samo zbirne podatke,
- uvedemo oznake v osnovni tabeli dejstev, ki povedo, ali je zapis osnovno dejstvo ali seštevek na višjem nivoju.

Za pravilno vključitev seštevkov v našo podatkovno zbirko moramo nove zapise dodati tudi v dimenzijske tabele.

4.3.2. Dimenzijske tabele

V dimenzijskih tabelah so opisane dimenzije modeliranega procesa. Zapisi opisujejo vse možne vrednosti, ki jih lahko zavzamejo elementi neke dimenzije. (Tako ima vsak film v našem primeru svoj zapis v dimenziji filmov.)

Če izhajamo s stališča uporabnikov (npr. analitikov, ki izvajajo poizvedbe), potrebujemo take opise dimenzij, ki bodo enostavno razumljivi, primerni za določanje robnih pogojev poizvedb in v končni fazi primerni tudi za označevanje vrstic v poročilih. Torej moramo dimenzije opisati s čim več diskretnimi tekstovnimi podatki, katerih pomen uporabniki dobro poznajo. Večina pomislekov v zvezi s trošenjem prostora v podatkovni bazi, ki ga zavzemajo ponavljajoči se tekstovni podatki, je povsem odveč, saj v tipičnem podatkovnem skladišču v primerjavi s tabelo dejstev tudi večje dimenzijske tabele zavzemajo minimalen prostor, izdvajanje opisov v ločene tabele ali celo kreiranje dodatnih hiearhij (kreiranje strukture, podobne snežinki – *snowflaking*) pa zelo oteži pregledovanje dimenzij in optimizacijo poizvedb.

Dimenzije brez atributov

Ob snovanju dimenzijskih tabel včasih naletimo na dimenzije, ki poleg ključa nimajo dodatnih atributov in dejansko sploh ne potrebujejo svoje tabele *(degenerate dimensions)*. Večkrat se tak primer pojavi, ko imamo opravka s klasičnimi dokumenti, npr. naročili ali računi. Številka originalnega dokumenta je podatek brez dodatnih atributov; koristi nam za grupiranje dejstev (izdelkov na računu) in na določenih izpisih kot povezava s pravimi dokumenti, sicer pa so vsi pomembni podatki že shranjeni v drugih dimenzijah.

Dimenzije, ki se počasi spreminjajo

Pogosto imamo opravka z dimenzijami, ki se počasi spreminjajo *(slowly changing dimensions)*. Takrat se moramo odločiti, kako se bomo spopadli z nastajajočimi spremembami. (Primer take spremembe je npr. sprememba demografske skupine kupca ali sprememba lastnosti izdelka. Pri vključitvi sprememb v bazo se običajno odločamo med naslednjimi tremi možnosti:

- Prepišemo lahko stare vrednosti v dimenzijski tabeli z novimi. S tem poenostavimo vzdrževanje, a obenem izgubimo vse zgodovinske podatke; vedno imamo na voljo le trenutno stanje.
- Dodamo lahko nov zapis v dimenzijsko tabelo in s tem ohranimo vse starejše podatke.
- V dimenzijski tabeli lahko uvedemo polja za trenutno in prejšnje vrednosti atributov, ki pa prinesejo kompleksne probleme pri izvajanju poizvedb.

Na odločitev, za katero od navedenih možnosti se bomo odločili pri obravnavi počasi spreminjajočih se dimenzij, vpliva predvsem poznavanje zahtev in pomena podatkov za končne uporabnike. Če podatka za resne analize ne uporabljajo, ga lahko prepišemo, v primeru potrebe po jasno vidnih spremembah se odločimo za drugo možnost, tretja pa predstavlja pot, za katero se odločimo le v primeru, ko ob spremembah ne želimo imeti dveh zapisov v dimenzijski tabeli, vseeno pa moramo spremembo evidentirati.

Minidimenzije

V primeru res velikih dimenzij lahko uporabimo prijem, ki zmanjša število potrebnih sprememb v dimenzijski tabeli: pogosto uporabljan atribut, ki se v dimenziji občasno spreminja, izločimo dimenzije v njegovo lastno minidimenzijo, namesto opisa v prvotni dimenziji pa ključ do atributa dodamo kar v tabelo dejstev. Tako lahko npr. iz tabele kupcev izločimo minidimenzijo z demografskimi podatki. Namesto da bi bila oznaka demografske skupine navedena pri posameznih kupcih, dodamo v tabelo dejstev ključe nove minidimenzije, v majhni dimenzijski tabeli pa

shranimo opise demografskih skupin. Posledično občasne spremembe demografskih podatkov tako ne vplivajo neposredno na tabelo kupcev (in ob nerodni uporabi spremenijo tudi pomen podatkov za nazaj). Izločene minidimenzije pozitivno vplivajo tudi na hitrost izvajanja poizvedb.

4.3.3. Zasnova dimenzijskega podatkovnega skladišča

Gradnja dimenzijskega podatkovnega skladišča je proces, v katerem potrebe uporabnikov preslikamo na dejansko razpoložljive podatke. Proces lahko razdelimo na devet tipičnih korakov, v katerih določamo:

- procese, ki jih modeliramo,
- najmanjše enote *(grain)*, ki bodo shranjene v posameznih tabelah dejstev,
- dimenzije posameznih tabel dejstev,
- vrednosti, ki bodo shranjene v tabelah dejstev,
- atribute posameznih dimenzij,
- načine spremljanja počasi spreminjajočih se dimenzij,
- seštevke (agregate), minidimenzije in druge posebnosti shranjenih podatkov,
- čas trajanja podatkov v bazi in
- pogostost prenosa podatkov iz transakcijskih baz (in drugih virov) v podatkovno skladišče.

Za naveden pristop je tipično, da začnemo pri vrhu – pri procesih, nato pa se spuščamo v globino in odkrivamo detajle. Večino odločitev sprejmemo šele po temeljitem razgovoru s končnimi uporabniki in s tistimi, ki upravljajo razpoložljive podatke.

Posamezne procese lahko identificiramo na osnovi mest, kjer se v podjetjih zbirajo podatki: blagajna (nakupi), klicni center (reklamacije, povpraševanja), spletni strežnik (sledi klikov). Za izbrani proces določimo najmanjšo enoto, ki bo shranjena v tabeli dejstev. Pogosto je najmanjša enota kar najbolj natančen podatek, ki je na voljo v transakcijskih sistemih, saj le s tako natančnimi zapisi podatkovno skladišče omogoča poljubne prereze in poglobljene poizvedbe. Pri določenih procesih pa lahko kot najmanjšo enoto izberemo tudi zbirne podatke (npr. prodaja izdelka v enem dnevu namesto posameznih vrstic na računih) ali posnetek stanja v določenem trenutku (zaloga izdelka ali stanje na računu v banki). Na osnovi najmanjše enote lahko hitro določimo tudi osnovne dimenzije, ki enoto opisujejo, glede na razpoložljivost podatkov in zahteve uporabnikov pa se lahko odločimo še za dodatne dimenzije.

V tej točki je rezultat navedenega postopka tipična zvezdasta shema, v kateri so definirana vsa polja tabel podatkovnega skladišča. V preostalih korakih natančneje

določimo postopke in parametre, s katerimi bo podatkovno skladišče živelo v praksi.

4.4. Tipične poizvedbe

Prednosti standardizirane strukture dimenzijskih podatkovnih skladišč so najbolje vidne, če si ogledamo končno uporabo – izvajanje uporabniških poizvedb. Že na začetku lahko povemo, da tipične poizvedbe iz velikega števila podatkov pripravijo relativno majhen odgovor, zgrajen na osnovi seštevkov ali drugih izračunov. Videli bomo, da so vse poizvedbe tudi podobno strukturirane.

Končni uporabniki se praviloma ne ukvarjajo s tabelami, SQL stavki in podobnimi tehničnimi podrobnostmi. Na voljo imajo posebna orodja, s katerimi kreirajo svoja poročila.

Osnovni korak pri kreiranju poizvedb je izbira vsebine - stolpcev poročila. V orodju za kreiranje poročil uporabnik iz tabele dejstev izbere polja, ki ga zanimajo (npr. seštevek prodaje, skupno število obiskovalcev), iz dimenzijskih tabel pa atribute, ki bodo osnova za grupiranje rezultatov (npr. žanr filma ali proizvajalec izdelka). Tako dejstev kot prikazanih atributov je lahko v enem poročilu več. Pred izvedbo poročila uporabnik določi še omejitve v dimenzijah, s katerimi omejimo množico dejstev, ki bodo vključena v končno poročilo. Skoraj vedno je ena od omejitev časovno obdobje, lahko se omejimo tudi samo na določeno kategorijo izdelkov ali zvrst filma, lahko gledamo obisk filma samo v določeni dvorani ali samo v dvoranah z manj kot 500 sedeži.

Gledanost filmov v maju 2002

Film	*Sum (število_obisk)*
Pošasti iz omare	25.291
Ocean's Eleven	15.354
Gospodar prstanov	10.598
E.T. vesoljček	9.580
Sestreljeni črni jastreb	5.279
Outsider	3.157

Zasedenost dvoran glede na opremo

Mesec	*Klima*	*Povpr.zasedenost*
April	da	75 %
April	ne	71 %
Maj	da	70 %
Maj	ne	65 %
Junij	da	63 %
Junij	ne	51 %

Vsi podatki so izmišljeni.

Slika 4.3. Tipična rezultata poizvedb v podatkovnem skladišču

Na zgornji sliki sta predstavljena rezultata dveh tipičnih poizvedb. Prvi primer je najenostavnejši – seštevek dejstev glede na eno samo dimenzijo (film) in omejitve v drugi dimenziji (časovno obdobje). Drugi primer je sicer zahtevnejši, a ga analitik z ustreznim orodjem ravno tako izvede samo z določanjem omejitev (časovno

obdobje) in vsebine oz. naslovov vrstic (mesec, klimatizirana dvorana ter določeno računska operacija nad dejstvi in atributi – razmerje med številom obiskovalcev in kapaciteto dvorane).

Ena glavnih prednosti enostavne in standardizirane zasnove dimenzijskih podatkovnih skladišč so enostavne poizvedbe, ki jih je mogoče tudi dobro optimizirati. Poizvedbe so dveh vrst: pregledovanje dimenzij (ob določanju omejitev) in končne poizvedbe, s katerimi kreiramo poročilo.

Pregledovanje dimenzij uporabniki izvajajo ob določanju omejitev. V posameznih dimenzijah določajo vrednosti atributov, ki naj bodo v končnem poročilu upoštevani. Zaradi čim bolj enostavnega pregledovanja dimenzij je pomembno, da se ob snovanju dimenzijskih tabel upremo želji po normalizaciji. Poleg dejstva, da nam normalizacija prinese le zanemarljive prihranke pri velikosti baze, lahko z nenormaliziranimi tabelami vse vrednosti atributov, ki analitika zanimajo in so lahko osnova za omejitve, poiščemo z enostavnim SQL ukazom `SELECT DISTINCT`.

Končna poizvedba je rezultat izbranih polj in določenih omejitev. V prvem od zgoraj navedenih primerov je to SQL stavek na sliki 4.4. Njegova struktura je tipična za poizvedbe v dimenzijskih podatkovnih skladiščih.

```
SELECT f.naslov, sum(o.st_obisk) as obisk      // polja poročila

FROM obiskanost o, filmi f, cas t              // tabele, ki jih potrebujemo

WHERE o.film_key= f.film_key                   // združevanje tabel (join)
  AND o.cas_key = t.cas_key                    // združevanje tabel (join)
  AND t.mesec   = "maj 2002"                   // omejitev v časovni dimenziji

GROUP BY f.naslov                              // določen nivo združevanja rezultatov

ORDER BY obisk DESC                            // določen vrstni red izpisa
```

Slika 4.4. Tipičen SELECT stavek poizvedbe v dimenzijskem podatkovnem skladišču

V seznamu polj `SELECT` stavka so našteti naslovi vrstic in seštevki (ali drugi izračuni), ki nas zanimajo. Del `FROM` je enostaven seznam tabel, iz katerih črpamo podatke in omejitve. `WHERE` pogoj poskrbi za združevanje tabel preko ključev (t.i. *join)* in vključuje omejitve, ki jih je postavil uporabnik glede vrednosti posameznih atributov v dimenzijah. Z določilom `GROUP BY` povemo, na osnovi katerih polj naj bo poročilo grupirano, zadnji stavek `ORDER BY` pa določa vrstni red rezultatov.

Pogosta operacija, ki jo uporabniki izvedejo nad dobljenim poročilom, je poizvedovanje v globino *(drill down)*. V praksi to ne pomeni nič drugega kot dodajanje novih stolpcev v poročilo. Tako bi lahko tabelo s slike 4.3 poglobili z dodatno kolono –

dvorano. Dobili bi poročilo o gledanosti filmov v posameznih dvoranah (znotraj določenega časovnega obdobja). [1]

4.5. Prenos podatkov v podatkovno skladišče

Preden se posvetimo praktičnemu primeru, si bomo na kratko pogledali še problematiko prenosa podatkov iz različnih virov v naše dimenzijsko podatkovno skladišče.

Že definicija govori, da je podatkovno skladišče kopija transakcijskih podatkov (glej poglavje 4.1.). Te moramo pred prenosom urediti tako, da bodo kar najbolj primerni za izvajanje poizvedb. V praksi se pogosto izkaže, da glavne tabele (npr. glavna tabela izdelkov, *product master*) niso primerne za neposreden prenos. Zaradi administrativnih napak ali neupoštevanja pravil se pojavljajo nepravilnosti, kot so ponavljajoče se šifre izdelkov ali spremembe izdelkov brez sprememb šifer. V dimenziji izdelkov (podobno je z dimenzijo kupcev ali obiskovalcev) zato v dimenzijskem podatkovnem skladišču običajno kreiramo svoje, neodvisne šifre. Ravno tako moramo med prenosom dopolniti in urediti tekstualne opise tako, da jih bodo končni uporabniki enostavno razumeli.

V času prenosa podatkovno skladišče običajno uporabnikom ni dostopno, saj se v ozadju izvaja veliko število operacij, podatki pa niso vedno v konsistentnem stanju. Šele po kontroli kakovosti podatkovno skladišče z novimi podatki spet odpremo za uporabo. Začasni nedostopnosti se lahko izognemo z uporabo podvojenih (potrojenih itd.) kopij podatkovnih skladišč, ki pogosto obstajajo že iz varnostnih razlogov, in z delnim indeksiranjem, kjer izkoristimo dejstvo, da so novi podatki običajno tisti s konca časovne dimenzije.

Vsakič, ko izvedemo prenos podatkov v podatkovno skladišče, izvajamo večino od naslednjih korakov:

1. branje podatkov iz produkcijskih virov,
2. identifikacija novih in spremenjenih zapisov,
3. generalizacija ključev v dimenzijah, kjer se podatki lahko spreminjajo,
4. priprava novih zapisov,
5. prenos podatkov v sistem podatkovnega skladišča,
6. urejanje in priprava zbirnih zapisov,
7. določanje ključev zbirnih zapisov,
8. prenos pripravljenih zapisov v tabele,
9. obdelava izjem,
10. kontrola kakovosti,
11. objava.

V primeru neuspeha na katerikoli točki proces prekinemo in podatkovno skladišče povrnemo v prejšnje stanje, saj delni prenosi v praksi običajno niso zaželeni.

5. Razpoložljivi podatki

V praktičnem delu naloge bomo na primeru spletne trgovine podjetja Merkur, d. d., zasnovali dimenzijsko podatkovno skladišče, s pomočjo katerega bomo analizirali obiskanost spletnega mesta in navade spletnih obiskovalcev.

Merkurjeva spletna trgovina

Merkur, d. d., je trgovsko podjetje za prodajo tehničnega blaga na debelo in drobno. Obstaja že več kot 100 let, danes pa se uvršča med najboljše slovenske trgovce. Zaposluje okoli 3.000 ljudi, letno ustvari okoli pol milijarde evrov prometa in 2 milijardi SIT dobička. [16]

Slika 5.1. Merkurjeva spletna trgovina

Na internetu je s spletnimi stranmi Merkur prisoten že od leta 1997 *(http://www.merkur.si)*, elektronskega poslovanja s končnimi kupci pa so se resno lotili v letu 2000. Prvi (poizkusni) korak je predstavljala spletna objava izdelkov iz sezonskih katalogov in izdelkov rednih prodajnih akcij »Vroče cene« z možnostjo naročila. Zaradi pozitivnih odzivov in pridobljenih izkušenj so si hitro postavili nove cilje in že v maju 2001 je bila postavljena prava spletna trgovina z obsežnim prodajnim programom in spremljajočimi vsebinami *(http://nakup.merkur.si)*. Prodajni program danes (avgust 2002) obsega okoli 2.500 izdelkov, spletno mesto pa vsebuje tudi veliko informativno-izobraževalnih vsebin. [15]

Spletna trgovina zaenkrat ni neposredno povezana s produkcijskimi podatkovnimi bazami Merkurja. Povezava v eni smeri poteka preko vmesnih datotek (izdelki, cenik), v drugi smeri pa ročno (vnos naročil).

Ob analizi dnevnika moramo upoštevati, da spletno mesto poleg same trgovine vsebuje tudi nekatera dodatna področja, od katerih so najbolj obiskane strani

spletnih iger (/nogomet/...) in strani za poslovne partnerje (/metalurgija/...), korporativne strani s predstavitvijo podjetja poslovni javnosti pa niso del spletnega mesta, ki ga analiziramo.

S stališča analize so pomembne tudi aktivnosti, ki jih Merkur izvaja za pridobivanje obiskovalcev spletnega mesta. Med najpomembnejšimi so spletno oglaševanje, promocijska elektronska sporočila (»Merkurjeve e-novice« pošiljajo naročnikom, veliko truda je usmerjenega v pridobivanje novih naročnikov), vpisovanje novosti v spletne imenike in iskalnike ter oglaševanje v lastnih promocijskih letakih in katalogih.

5.1. Dnevnik spletnega strežnika

V dnevniku Merkurjevega spletnega strežnika so zapisane vse HTTP zahteve, ki jih je strežnik prejel in izvedel. Vsak zapis je v svoji vrstici, polja v zapisu so med seboj ločena s presledkom. Zapisi vsebujejo naslednja polja:

Oznaka	Pomen	Primer vsebine
Date	datum zahteve	2002-07-20
Time	čas zahteve	07:02:51
C-ip	IP naslov odjemalca	217.72.73.35
Cs-username	uporabnikovo ime	-
S-computername	ime strežnika	NAKUPWEB
Cs-method	HTTP metoda	GET
Cs-uri-stem	zahtevan element	/glava.asp
Cs-uri-query	dodatni parametri zahteve	MySes=04FZKYJVCK0PCBW2C2JU
Sc-status	status zahteve	200
Sc-bytes	prenešenih bajtov (odjemalec-strežnik)	0
Cs-bytes	prenešenih bajtov (strežnik-odjemalec)	520
Time-taken	porabljen čas v milisekundah	31
Cs-version	verzija protokola	HTTP/1.1
Cs(User-Agent)	odjemalec (brskalnik)	Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
Cs(Cookie)	vsebina piškotka	VisitorId=ZHKV30OSRUUNKN4WLCTM; +ASPSESSIONIDQQQGGRAO=IJMIEJMAOFHAM...
Cs(Referrer)	prejšnja stran	http://nakup.merkur.si/default.asp?MySes=04FZKYJV...

Najpomembnejše informacije o zahtevah obiskovalcev se skrivajo v poljih *Cs-uri-stem* in *Cs-uri-query* – iz njiju lahko razberemo splošen tip strani, ki je bila zahtevana (naslovnica, oddelek, izdelek, nasvet itd.), pa tudi konkretno kodo elementa (izdelka, oddelka ali nasveta). Med parametri strani je vedno tudi številka, s katero enolično prepoznamo uporabniško sejo (parameter *MySes*). Sejo lahko pri uporabnikih, ki imajo omogočen prenos piškotkov, identificiramo tudi v polju *Cs(Cookie)*. V njem

sta avtomatska oznaka seje (*ASPSessionID*, vsakemu obiskovalcu jo dodeli strežnik) in oznaka *VisitorID*, s katero spletna aplikacija trajno označi obiskovalca. (Če smo natančni, je s tem piškotkom enolično označen le brskalnik na določenem računalniku, a za naše potrebe bomo privzeli, da je to dovolj za identifikacijo uporabnika.) Na osnovi te vrednosti lahko združimo podatke vseh sej istega obiskovalca.

5.2. Dnevnik prihodov obiskovalcev

Zaradi zahtev po analizi učinkovitosti spletnih oglasov in povezav na drugih spletnih mestih, ki jo je bilo na osnovi dnevnika spletnega strežnika težko izvajati, je bil aprila 2002 uveden tudi t.i. dnevnik prihodov obiskovalcev, v katerem so glavni podatki koda obiskovalca (VisitorID), oznaka referenta ali oglasa in naslov strani, s katere prihaja povezava. Natančneje je vsebina razdelana prikazana v spodnji tabeli:

Oznaka	Pomen	Primer vsebine
Ref	oznaka referenta	0212g157445
VisitorID	interna (trajna) oznaka uporabnika	GM8DR919HZ576P4A3RTU
IP	IP naslov odjemalca	194.249.253.140
LastPage	prejšnja stran	-
UserAgent	odjemalec (brskalnik)	Mozilla/4.0 (compatible; MSIE 5.01; Windows 98)

Hitro lahko opazimo, da manjka zapis časa prihoda (datum je razviden iz imena datoteke, v kateri je shranjen dnevnik) – dobimo ga lahko s križanjem podatkov z dnevnikom spletnega strežnika. Za nameček se je v nekaj mesecih, odkar poteka beleženje prihodov, oblika zapisa že spremenila in so starejši zapisi v drugačnem formatu.

Vsak dan se prihodi beležijo v svojo datoteko. Posamezni zapisi so v eni vrstici, polja so ločena z znakom ^.

5.3. Dnevnik nakupov

Vsak uspešno izveden nakup se zabeleži v dnevnik nakupov. Zapis vsebuje naslednja polja:

Oznaka	Pomen	Primer vsebine
DateStamp	datum, ura	11.8.2002 12:15:27
VisitorID	interna (trajna) oznaka uporabnika	GM8DR919HZ576P4A3RTU
MySess	Interna oznaka trenutne seje	JBDG8SMX7SX9QXM9VC81
InvoiceNum	Številka računa	8185
Amount	Znesek računa	35990
IP	IP naslov odjemalca	194.249.253.140

Tudi dnevnik nakupov je bil uveden aprila 2002 in tudi njegova struktura se je v času od aprila do avgusta 2002 že spremenila, kar bomo upoštevali pri uvozu podatkov.

5.4. Natančnejši podatki o nakupih

Podatki o nakupih so urednikom spletne trgovine kadarkoli na voljo v obliki Excelove preglednice. Ta vsebuje zelo natančne podatke o kupcih (ime, priimek, naslov, pošta, elektronski naslov, številka Merkurjeve kartice lojalnosti), o nakupljenih izdelkih (koda, kategorija, opis, cena, v akciji) in o samih nakupih (datum in ura, način plačila in dostave, skupni znesek nakupa). Na osnovi številke računa lahko te podatke povežemo z drugimi viri.

5.5. Tabela izdelkov in kategorij

V predstavitvi Merkurjeve spletne trgovine je bilo že omenjeno, da ni neposredne povezave med produkcijskimi bazami in spletno aplikacijo. To velja tudi za podatke o izdelkih. Ti so zato pripravljeni v Excelovi preglednici, ki jo sistem zna tako uvoziti (ob sprejemu novih in spremenjenih podatkov) kot izvoziti (ko želimo dobiti zapis trenutnega stanja šifranta).

Obsežna tabela z dogovorjeno strukturo stolpcev in vrstic vsebuje tako podatke o izdelkih kot o njihovi urejenosti (na oddelke, pododdelke in police). Izdelki so predstavljeni s kodami (šestmestna koda Merkurjevega informacijskega sistema, dodatna štirimestna koda za potrebe spletne trgovine, EAN koda), nazivom, blagovno znamko, modelom, podrobnim opisom, redno in akcijsko ceno ter kodo akcije. Kategorije opisujeta koda in naziv. (Nekatere za naše potrebe manj pomembne podatke, kot so šifra dodatnega opisa ali koda slike, namenoma izpuščamo.) Glede na položaj v datoteki lahko ugotovimo, v katero kategorijo izdelek sodi.

5.6. Seznam svetovalnih člankov

Svetovalni članki niso shranjeni v bazi, za potrebe urejanja obstaja le enostaven seznam (v obliki Excelove preglednice) z vpisanimi naslovi, s kodami člankov, iz seznama pa lahko razberemo tudi, v katero kategorijo sodijo posamezni članki.

Kot lahko hitro sklepamo iz opisa razpoložljivih podatkov, nas čaka pri pretvorbi in združevanju v kakršnokoli obliko, ki jo bo moč obdelati v podatkovnem skladišču, precej dela. Vidi se, da so opisani viri podatkov nastajali parcialno, brez jasno določenih skupnih ciljev in kot taki še niso povsem primerni za praktično uporabo.

6. Izgradnja dimenzijskega modela za analizo obnašanja obiskovalcev

V tem poglavju bomo na osnovi poznavanja problematike spletne trgovine in razpoložljivih podatkov zasnovali dimenzijsko podatkovno skladišče, s katerim bomo lahko analizirali obisk in navade spletnih obiskovalcev.

Procesi, ki jih bomo opisali s posameznimi dimenzijskimi modeli, so naslednji:

- uporabniške seje,
- ogledi strani in
- nakupi.

Na začetku bomo definirali nekaj skupnih dimenzijskih tabel, ki jih bomo uporabili v podatkovnem skladišču. Nekatere dimenzije bomo v tej nalogi samo definirali in nakazali smernice za njihovo uporabo, saj jih zaradi nedosegljivosti podatkov ne bomo mogli napolniti in uporabiti pri analizi.

6.1. Skupne dimenzijske tabele

6.1.1. Datum

Podatkovne baze praviloma vsebujejo tipe polj, ki so primerni za shranjevanje ure in datuma, a se v podatkovnem skladišču velikokrat odločimo za dodatno dimenzijo (pravzaprav dve), s katerima lahko veliko enostavneje izvajamo kompleksne poizvedbe, povezane s časovno dimenzijo dejstev.

Vsak datum v obdobju, ki nas zanima, ima v tej dimenzijski tabeli svoj zapis. Čeprav je večino od opisnih atributov z bolj ali manj zahtevnimi računskimi prijemi kadarkoli mogoče izračunati iz osnovnega podatka – datuma, so poizvedbe in analize precej enostavnejše z vnaprej pripravljenimi atributi.

Atribut	Pomen
Datum_key	Neodvisen ključ, 1 .. N
Dan v mesecu	1 .. 31
Mesec v letu	1 .. 12
Leto	Npr. 2002
SQL datum	Celoten zapis datuma
Ime dneva	ponedeljek, torek, ...

Številka dneva v tednu	1 .. 7
Številka dneva v letu	1 .. 366
Številka dneva v obdobju	Od začetka nekega obdobja, lahko negativno.
Številka tedna v letu	1 .. 53
Številka tedna v obdobju	Od začetka obdobja, lahko negativno
Ime meseca	januar, februar, ...
Številka meseca v obdobju	Od začetka obdobja, lahko negativno.
Delavnik	Delavnik / Praznik
Praznik	Opis praznika
Vikend	Vikend / Med tednom
Posebni dogodki	Opis posebnih dogodkov oz. prazno.
Obračunsko obdobje	Oznaka obračunskega obdobja

Tabelo lahko vnaprej napolnimo s podatki za obdobje, ki nas zanima. Kasneje urejamo samo morebitne posebne opise dni. Za potrebe pregledovanja ter naslove vrstic v poročilu je zelo primerno, če v tabelo vnesemo res opisne podatke (npr. »Delavnik« ali »Praznik« namesto »Da« ali »Ne« v polju Delavnik).

6.1.2. Ura

S to dimenzijo bomo opisali podatke o točnem času nekega dejstva. Glede na obseg in naravo podatkov, ki jih bomo analizirali, smo se odločili za precejšnjo natančnost tabele (do sekunde natančno). V tej dimenziji je opisan samo čas dogodka znotraj enega dneva, sam dan je definiran v datumski dimenziji.

Atribut	Pomen
Ura_key	Neodvisen ključ, 1 .. N
Ure	0 .. 23
Minute	0 .. 59
Sekunde	0 .. 59
SQL ura	Skupen zapis ure
Sekunde po polnoči	0 .. 86399
Minute po polnoči	0 .. 1439
Obdobje	Opis obdobja dneva: ponoči, dopoldne, popoldne, zvečer, ...; odvisno od potreb

V primeru, da našo tabelo časa vnaprej napolnimo z vsemi možnimi vrednostmi, bo vsebovala okoli 86.400 zapisov. Zapise lahko dodajamo tudi po potrebi - ob dodajanju novih dejstev. Pogosto je za potrebe analize dovolj natančen tudi zapis ure z minutno natančnostjo.

6.1.3. Strani

Kot smo omenili v poglavju o prilagoditvi spletnega mesta za potrebe analiziranja, strani običajno razdelimo v tipične skupine. S preučevanjem obiskovanja in prehajanja med različnimi skupinami spoznamo pomen tipičnih strani za posamezna

dejanja, ki nas zanimajo in nam lahko razkrijejo navade obiskovalcev. Vsaka stran sodi v eno samo skupino. Te skupine so shranjene v naslednji dimenzijski tabeli:

Atribut	Pomen
Stran_key	Neodvisen ključ, 1 .. N
Opis	Tekstovni opis tipa strani
Tip strani	Neznano, Izdelek, Kategorije, Iskanje, Svetovanje, Splošne informacije, Novice, Voziček, Izvedba nakupa, Konec nakupa, Zabava
Področje	Pri straneh z izdelki in članki vpisano področje
Podpodročje	Pri straneh z izdelki in članki vpisano podpodročje
Detajl	Opcija: natančnejši podatki o vsebini strani – opis izdelka ali članka

Dimenzijo strani lahko naredimo veliko bolj podrobno tako, da v polju Detajl shranimo natančen podatek o vsebini, npr. opis izdelka ali naslov članka. Tako obseg dimenzije naraste z nekaj 100 tipov strani na nekaj 1.000 končnih strani spletnega mesta.

6.1.4. Obiskovalec

Atribute v dimenziji obiskovalcev bi lahko razdelili v več skupin, ki jih polnimo postopno, bolj ko spoznavamo naše obiskovalce. Najprej imamo opravka z neidentificiranimi uporabniki, ki jih nato prepoznamo (se nam predstavijo), te informacije pri kupcih nadgradimo z zbirnimi podatki o pogostosti nakupov, končni (v praksi zelo težko uresničljiv) cilj pa je segmentacija obiskovalcev glede na njihove eksplicitno izražene in implicitno prepoznane interese, dodajanje demografskih podatkov, torej zbiranje praktično vsega, kar vemo o strankah, na enem mestu. Dimenzija obiskovalcev bo v prihodnosti zelo pomemben vir informacij, saj bomo na njeni osnovi izvajali prilagajanje vsebin.

Atribut	Pomen
Obiskovalec_key	Neodvisen ključ, 1 .. N
Podatki o neznanem obiskovalcu	
Tip obiskovalca	Neznano, enkratni, poznani-piškotek, poznani-prijava, kupec-piškotek, kupec-prijava
VisitorID	Trajna oznaka (v piškotku)
Število obiskov	Skupno število obiskov tega obiskovalca
Zadnji obisk	Datum zadnjega obiska
Podatki o poznanem obiskovalcu	
Tip imena	Nepreverjeno, psevdonim, pravo ime
Ime, priimek, ...	Splošni atributi
Naslov, ulica, kraj, pošta, ...	Splošni atributi

Poštno območje	Neznano, 1 .. 9, izven Slovenije
Država	Ime države
Spol	Neznan, moški, ženski
Starost	Vnaprej definirana obdobja
E-pošta	Elektronski naslov
Naročnik e-novic	Prejema e-novice, Ni naročnik e-novic
Ostali kontaktni podatki ...	
Številka MKZ	Merkurjeva kartica zaupanja – lojalnostna kartica
Podatki o kupcu	
Število opravljenih nakupov	
Datum zadnjega nakupa	
Skupni znesek nakupov	
...	

V praksi (še) nimamo na voljo demografskih podatkov, podatkov o interesih, o nakupnih navadah in podobnih atributov, s katerimi bi lahko tabelo obiskovalcev dopolnili in nato izvajali poglobljene poizvedbe, ki bi nas pripeljale do tipičnih skupin obiskovalcev (segmentacija) in bi bile osnova za prilagojeno ponudbo.

Pri obiskovalcih se moramo spoprijeti s problemom počasi spreminjajoče se dimenzije. Podatke o obiskovalcih namreč dokaj pogosto dopolnjujemo in nadgrajujemo. Pri določenih spremembah lahko uporabimo tehniko prepisovanja (predvsem, ko prej neznane vrednosti atributov prepišemo z veljavnimi podatki, saj takrat ne vplivamo na točnost poizvedb). Ob večjih spremembah atributov, ki opisujejo obiskovalce, pa uporabimo tehniko podvajanja zapisov. Če npr. nek obiskovalec opravi prvi nakup, podvojimo njegov zapis v dimenzijski tabeli, tako da stari podatki (o prejšnjih sejah, obiskih) kažejo na 'starega' obiskovalca – tistega, ki še ni opravil nakupa, vsa nova dejstva (seje, obiski, nakupi) pa so povezana z 'novim' obiskovalcem – kupcem.

6.1.5. Izdelek

V dimenziji izdelkov so poleg osnovnih podatkov o izdelku zanimivi atributi, ki opisujejo razporeditev v posamezne kategorije. Problem pojavljanja istega izdelka v več kategorijah rešimo s podvajanjem zapisov, saj le tako lahko spremljamo popularnost izdelkov v posameznih kategorijah. Hierarhija kategorij je opisana z dvema nivojema: v atributu Oddelek je shranjen osnovni oddelek znotraj področij Merkur Dom in Merkur Mojster, v drugem atributu (Polica) pa konkretna virtualna polica, na kateri je izdelek.

Atribut	Pomen
Izdelek_key	Neodvisen ključ, 1 .. N
Merkurjeva koda	Interna oznaka – format nnnnnn-nnnn
EAN koda	Proizvajalčeva oznaka, 12 mest
Kratek opis	Za prikazovanje na izpisih

Blagovna znamka	Npr. Electrolux
Vrsta izdelka	Generično ime, npr. hladilnik
Model	Npr. HL 512a
Opis	
Klasifikacija-oddelek	Npr. Dom-Bela tehnika, Mojster-Elektro, ...
Klasifikacija-polica	Npr. Hladilniki, Stikala, ...
...	

6.1.6. Akcija

Dimenzija prodajnih in promocijskih akcij nam bo pomagala razložiti vpliv akcij na obisk v spletni trgovini (vpliv na število obiskovalcev, kakovost obiskov) in na gledanost ter prodajo posameznih izdelkov.

Akcije lahko pomensko razdelimo na dva tipa. V prodajnih akcijah imajo izdelki znižano ceno, taka akcija vpliva direktno na število ogledov in prodajo posameznih izdelkov. Promocijska akcija (oglasne pasice, oglasi v drugih medijih, e-novice) pa vplivajo na obisk in na strukturo obiskovalcev. Zato smo dimenzijo akcij razdelili na dve; prvo bomo uporabili v modelu za analizo obiskanosti strani, drugo pa v modelu za analizo uporabniških sej.

Dimenzija prodajnih akcij:

Atribut	Pomen
Akcija_key	Neodvisen ključ, 1 .. N
Opis akcije	
Tip akcije	Ni akcije, Prodajna-posezonsko znižanje, prodajna-stare zaloge, prodajna-vroče cene, prodajna-ostalo

Dimenzija promocijskih akcij:

Atribut	Pomen
Akcija_key	Neodvisen ključ, 1 .. N
Opis akcije	
Tip akcije	Ni akcije, interna promocijska akcija, zunanja promocijska akcija
Tisk	Opis uporabljenih tiskanih oglasnih sredstev
Radio	Opis uporabe medija
TV	Opis uporabe medija
Internet	Opis uporabe medija
Drugo	opis uporabe drugih medijev

Dimenzija akcij je t.i. vzročna dimenzija *(causal dimension)*, saj pomaga razložiti vzroke za dogajanja v našem sistemu.

6.1.7. Članek

S stališča analize je dimenzija člankov zelo podobna dimenziji izdelkov – imamo hierarhično urejeno množico vsebin, za katere nas zanima gledanost. Članki so urejeni v področja in po potrebi podpodročja.

Atribut	Pomen
Članek_key	Neodvisen ključ, 1 .. N
Naslov	
Avtor	Ni podatka / Ni pomembno / Ime ...
Grafična oprema	Slike / Brez slik
Področje	Neznano, Pomoč, Novice, Svetovanje, ...
Podpodročje	Dom, vrt, delavnica (znotraj Svetovanje)
...	

6.1.8. Viri

V dimenziji virov so natančneje opisana različna spletna mesta, s katerih prihajajo k nam obiskovalci. Poizvedbe, povezane z dimenzijo virov, nam dajo koristne informacije o učinkih povezovanja z drugimi spletnimi mesti, spletnega oglaševanja, vpisovanja v imenike in pošiljanja obvestil po elektronski pošti.

Atribut	Pomen
Vir_key	Neodvisen ključ, 1 .. N
Opis	
Tip	Nepoznano / Iskalnik / Oglasna pasica / Plačana povezava / Izmenjava povezav / E-novica ...
URL vira	Npr. http://www.domena.com/povezave.html
Domena vira	www.domena.com

6.2. Dimenzijski podatkovni model za analizo uporabniških sej

Prvi proces, ki ga bomo opisali z dimenzijskim podatkovnim modelom, je namenjen analiziranju uporabniških sej. Pomagal nam bo odgovoriti na vprašanja, kot so:

- splošna vprašanja o obisku spletne trgovine (število obiskovalcev skozi čas, dolžina obiska, število pogledanih strani, osnovni nameni obiska, ...),
- vpliv promocijskih aktivnosti na obisk in »kakovost« obiskovalcev, ki jih pripeljejo različni zunanji viri,
- vpliv različnih faktorjev znotraj seje na končni uspeh – izvedbo nakupa.

Osnovni element tabele dejstev

Osnovni element v tabeli dejstev bo ena zaključena uporabniška seja (obisk). Opisali jo bomo z naslednjimi dimenzijami: datum in ura začetka seje, obiskovalec, vir, vstopna stran, zadnja stran, namen seje, promocijska akcija. Dejstva, ki posamezno sejo opisujejo, so trajanje, število obiskanih strani, število pogledanih izdelkov, število opravljenih iskanj, število pogledanih nasvetov in drugih člankov, znesek opravljenih nakupov in šifra opravljenega nakupa (številka računa). Zaradi enostavnejših poizvedb (seštevanja namesto štetja z SQL stavki) je uvedeno tudi polje s konstantno vrednostjo 1.

Večino dimenzijskih tabel smo že spoznali, v nadaljevanju je opisana še dimenzija namenov seje.

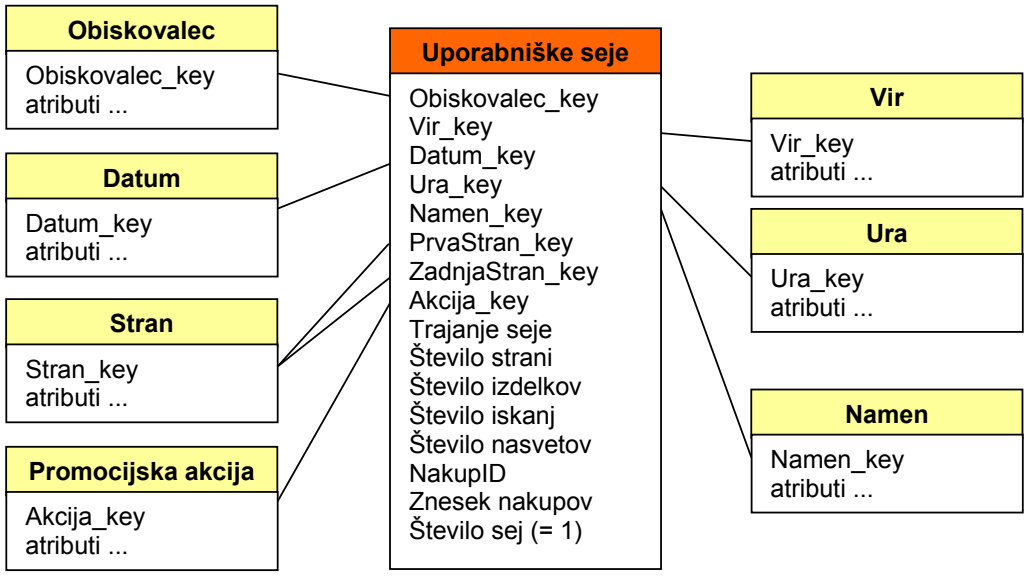
Dimenzijska tabela namenov seje

Namen seje bomo opisali glede na strani in opravila, ki jih je obiskovalec znotraj seje pogledal ali izvedel. Dimenzijska tabela vsebuje zapise za vse potrebne kombinacije izbranih namenov:

Atribut	Pomen
Namen_key	Neodvisen ključ, 1 .. N
Opis	Združen opis namena seje Npr. »Info-izdelki, iskanje, nakup, zabava«
Info-izdelki	Ali je pregledoval informacije o izdelkih?
Info-akcije	Ali je obiskal strani prodajnih akcij?
Info-novice	Ali je pregledoval novice?
Info-splošno	Ali ga zanimajo splošne informacije?
Iskalnik	Ali je uporabil iskalnik? Brez iskanja / eno iskanje / več iskanj
Nakup	Ali se je lotil nakupa - kako daleč je prišel? Brez nakupa / voziček / blagajna / opravil nakup
Zabava	Ali je obiskal dele spletne trgovine, namenjene zabavi?

Dimenzijski model

Nastali model ima tipično zvezdasto strukturo:



Slika 6.1. Dimenzijski model podatkovnega skladišča za analizo uporabniških sej

Viri podatkov

Osnovni vir podatkov, s katerim bomo napolnili zgornji podatkovni model, je dnevnik spletnega strežnika. Program, ki bo procesiral obsežne zapise v dnevniku in jih združeval v seje, mora znati prepoznati ponovne obiskovalce, predvsem pa iz razdrobljenih podatkov ugotoviti tipe strani in namen celotne seje. Ob tem uporabi vnaprej pripravljene podatke o namenih posameznih strani in promocijskih akcijah.

6.3. Dimenzijski podatkovni model za analizo obiska posameznih strani

Z modelom za analizo obiska posameznih strani bomo dobili odgovoriti na vprašanja, kot so:

- Katere so najbolj priljubljene strani, izdelki in članki (na nivoju spletnega mesta, v posameznih kategorijah, pri izbranih vrstah obiskovalcev, skozi čas, ...)?
- Kako se obiskovalci gibljejo po spletnem mestu? Kateri so najpogostejši prehodi med stranmi?
- Koliko časa si obiskovalci vzamejo za ogled posameznih strani?

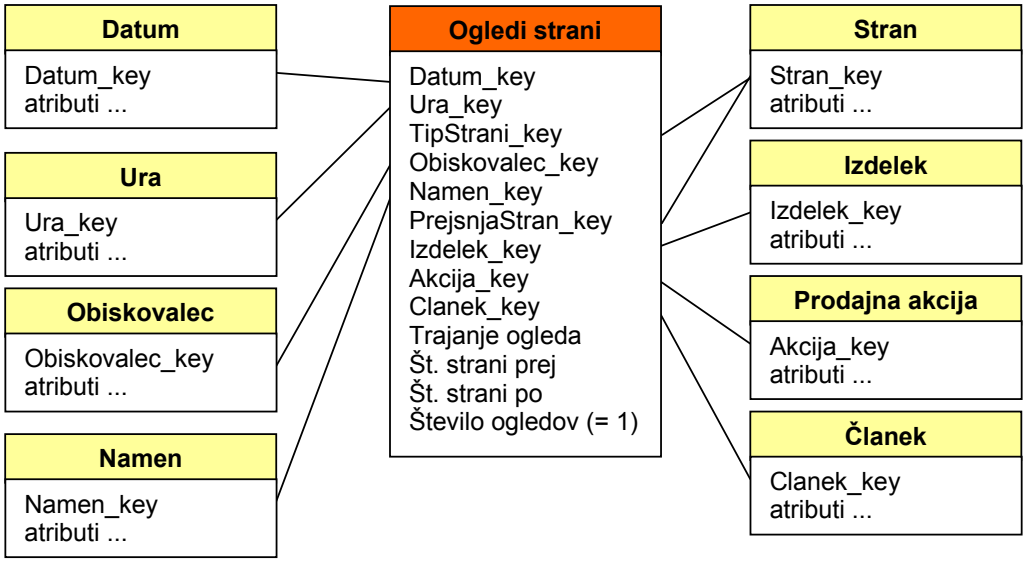
- Katere so najpogostejše kombinacije strani, po katerih obiskovalci spletno mesto zapustijo?

Tabela dejstev

Osnovni element v tabeli dejstev bo ogled ene strani, ki ga izvede nek obiskovalec. En ogled bomo opisali z naslednjimi dimenzijami: datum in ura, tip strani, izdelek, prodajna akcija, članek, obiskovalec, namen celotne seje. V tabelo dejstev bomo poleg ključev navedenih dimenzij zapisali še trajanje ogleda strani, prejšnjo stran in število ogledov strani pred in po tej strani znotraj aktivne seje.

Dimenzijski model

Ker smo v predhodnih odstavkih že spoznali vse dimenzijske tabele, si oglejmo zvezdno shemo dimenzonalnega modela.



Slika 6.2. Dimenzijski model podatkovnega skladišča za analizo ogledov strani

Viri podatkov

Tabelo dejstev v zgornjem podatkovnem modelu bomo polnili na osnovi podatkov v dnevniku spletnega strežnika. Dimenzijske tabele (predvsem strani, izdelke, članke, akcije, obiskovalce) moramo pripraviti pred tem ali vzporedno s pomočjo drugih virov. Zaradi velikega obsega in združevanja podatkov iz velikega števila virov je polnjenje tega modela še posebej zahtevno.

6.4. Dimenzijski podatkovni model za analizo nakupov

Čeprav prodaja ni edini (pogosto niti najpomembnejši) cilj spletne trgovine, želimo tudi na osnovi opravljenih nakupov pridobiti spoznanja o navadah in zahtevah obiskovalcev našega spletnega mesta. Poleg najbolj očitnega vprašanja, kateri izdelki se najbolje prodajajo, trgovca običajno zanimajo tudi odgovori na naslednja vprašanja:

- Katere skupine izdelkov so bolj in katere manj primerne prodajo po internetu? Kako se to spreminja skozi čas?
- Kakšna je povezava med številom ogledov in prodajo izdelkov?
- Kako na prodajo vplivajo prodajne in promocijske akcije?
- Ali nekateri viri prinašajo bolj verjetne kupce od drugih?
- Kako na prodajo vplivajo demografske značilnosti kupcev?

Tabela dejstev

Osnovni element tabele dejstev bo v tem dimenzijskem modelu nakup ene vrste izdelka, ki ga opravi znotraj ene seje en obiskovalec (v praksi je to običajno ena vrstica na računu). Nakupe bomo opisali z naslednjimi dimenzijami: datum in ura, izdelek, prodajna akcija, promocijska akcija, obiskovalec, vir, tip reklamacije. V tabeli dejstev bodo še podatki o številu kupljenih izdelkov in ceni ter številka nakupa, na osnovi katere lahko povežemo vse izdelke enega naročila.

Dimenzijska tabela reklamacij

Ker je reševanje reklamacij za spletnega trgovca zelo drago opravilo, bomo podatkom o nakupih v našem podatkovnem skladišču dodali tudi zaznamek o morebitni reklamaciji. Kasneje bomo lahko s pomočjo analize odkrili faktorje, ki vplivajo na število reklamacij, in se jim pri nadaljnjem poslovanju poizkusili izogniti.

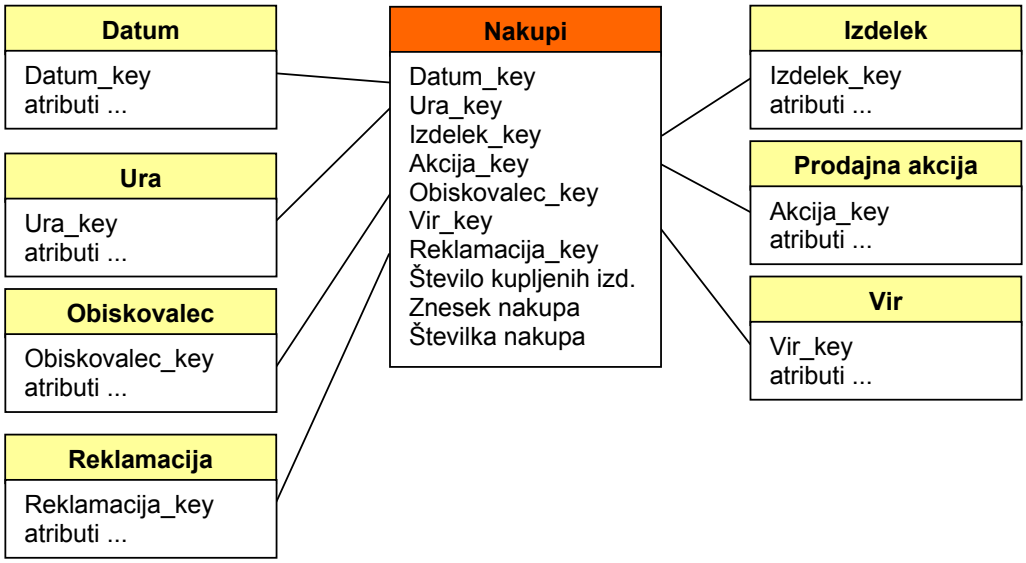
Reklamacije so vezane na kupljene izdelke; opisali jih bomo z dvema atributoma: z razlogom reklamacije in načinom rešitve problema.

Atribut	Pomen
Reklamacija_key	Neodvisen ključ, 1 .. N
Opis	Tekstovni opis tipa reklamacije
Razlog zavrnitve	Ni reklamacije / neznan / ni naročil / izdelek ni zadovoljil pričakovanj / izdelka ne potrebuje / izdelek poškodovan pri transportu / izdelek ni deloval / drugo
Način rešitve	Neznan / dostavljen nov izdelek / dostavljen nadomestni izdelek / vrnjen denar / drugo

Nastala dimenzijska tabela je zelo majhna, vključuje le smiselne kombinacije naštetih razlogov zavrnitve in rešitve reklamacij.

Dimenzijski model

Na spodnji sliki je zvezdna shema dimenzijskega podatkovnega modela:



Slika 6.3. Dimenzijski model podatkovnega skladišča za analizo nakupov

Viri podatkov

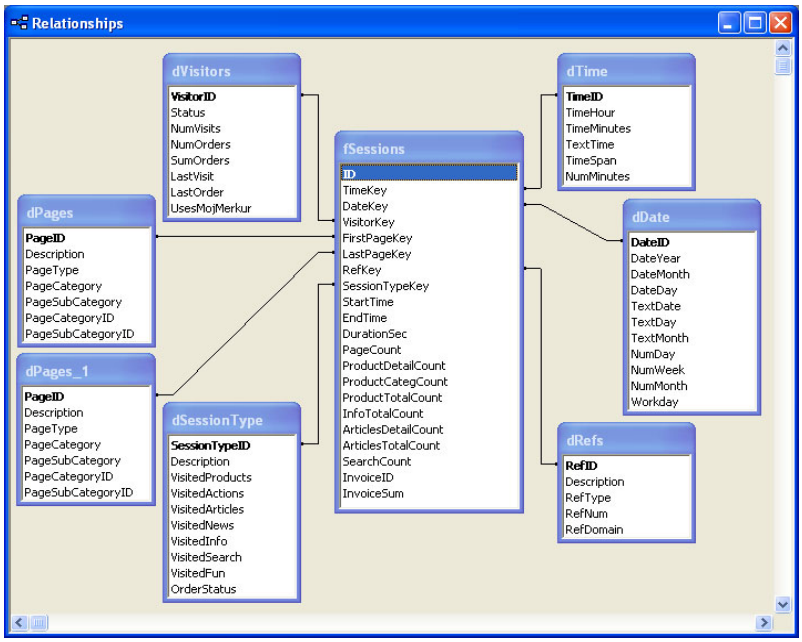
Podatke o nakupih v Merkurjevi spletni trgovini so na voljo v poročilu o nakupih v obliki Excel datoteke, ravno tako so podatki za polnjenje tabele dimenzij izdelkov in obiskovalcev zaenkrat na voljo v tej obliki. Določene dimenzije lahko napolnimo iz dnevnika spletnega strežnika (vir, obiskovalec), pred objavo jih po potrebi le še dopolnimo z dodatnimi atributi (npr. natančnejši opis virov). Podatki o prodajnih akcijah so na voljo ob izdelkih. Dimenzija reklamacij v praksi še ni podprta v nobenem viru. Prikazana je predvsem kot možna nadgradnja podatkovnega modela v prihodnosti.

7. Praktičen preizkus modela

V prejšnjem poglavju smo zasnovali tri dimenzijske podatkovne modele, v katerih lahko shranimo podatke o uporabniških sejah, obisku posameznih strani in nakupih v spletni trgovini. Prvega od modelov, namenjenega analizi uporabniških sej, bomo preizkusili tudi v praksi: napolnili ga bomo s podatki iz Merkurjeve spletne trgovine (za polletno obdobje od februarja do julija 2002) in jih analizirali s pomočjo orodja Microsoft SQL Server 2000 – Analysis Services.

7.1. Prenos podatkov v dimenzijski model

Na osnovi temeljite analize razpoložljivih podatkov, opisanih v 5. poglavju, smo prilagodili dimenzijski model, opisan v poglavju 6.2. Končna zvezdna shema je prikazana na sliki 7.1.



Slika 7.1. Tabela dejstev in dimenzijske tabele podatkovnega skladišča za analizo sej

Tabela dejstev eno sejo opisuje z naslednjimi podatki: trajanje obiska, število pogledanih strani (skupno in natančneje opredeljeno po skupinah strani), znesek

opravljenega nakupa in številka računa. Seje so opisane s sledečimi dimenzijami: Ura, Datum, Obiskovalec, Vir, Vstopna stran, Izhodna stran in Tip seje. (Podatki o promocijskih akcijah niso na voljo v urejeni obliki, zato jih v naš model nismo vključili.)

Glavni vir podatkov za tabelo dejstev in za dimenzijske tabele je dnevnik spletnega strežnika. Določene podrobnosti (opisne podatke o posameznih kategorijah strani, podatke o nakupih) bomo vključili iz drugih virov, na koncu pa v nekaterih dimenzijah tudi ročno preverili, prečistili in dopolnili zbrane podatke, da bomo dosegli enega glavnih ciljev podatkovnega skladišča – razumljivost informacij za končnega uporabnika.

Od dnevnika spletnega strežnika do podatkov v zvezdni shemi bomo prišli v več korakih:

- izbor ustreznih zapisov izmed vseh zapisov v dnevniku spletnega strežnika (iščemo samo podatke o ogledih strani, ki nas v analizi zanimajo),
- združevanje podatkov o obiskih strani v posamezne seje,
- kreiranje dimenzijskih tabel, finalizacija zvezdne sheme.

Vmesni rezultat med prvim in drugim korakom bodo prečiščeni dnevniki (še vedno v tekstovni obliki), med drugim in tretjim korakom pa tabela v podatkovni bazi z vsemi podatki o eni seji.

7.1.1. Izbor zapisov o ogledih strani

Kot smo omenili v uvodu tega poglavja, bomo v nalogi praktično analizirali podatke o obisku Merkurjeve spletne trgovine v polletnem obdobju od februarja do julija 2002. Naša vhodna zbirka podatkov (dnevniki spletnega strežnika) obsega čez 24 milijonov zapisov, shranjena je v obliki tekstovnih datotek in je velika okoli 6,67 Gb.

Prvi korak je relativno enostaven: za vsak zapis iz dnevnika preverimo, ali v polju URI (*Uniform Resource Identifier* – opisuje element, ki ga uporabnik zahteva od strežnika) vsebuje klic strani – te prepoznamo po končnici datoteke: .html, .htm ali .asp. Ustrezne zapise prepišemo v vmesno datoteko. Aplikacija je bila narejena v orodju Borland Delphi.

Obseg podatkov smo v prvem koraku zmanjšali na desetino: vmesna datoteka vsebuje okoli 2,4 milijona zapisov o ogledih strani in je velika 667 Mb.

7.1.2. Združevanje podatkov o posameznih sejah

Vsebinsko veliko bolj zahteven je drugi korak pretvorbe, saj kakovost združevanja podatkov o straneh v podatke o sejah zelo vpliva na končni rezultat – pravilnost podatkov v podatkovnem skladišču.

Najpomembnejši dejavniki pri združevanju podatkov o sejah so:

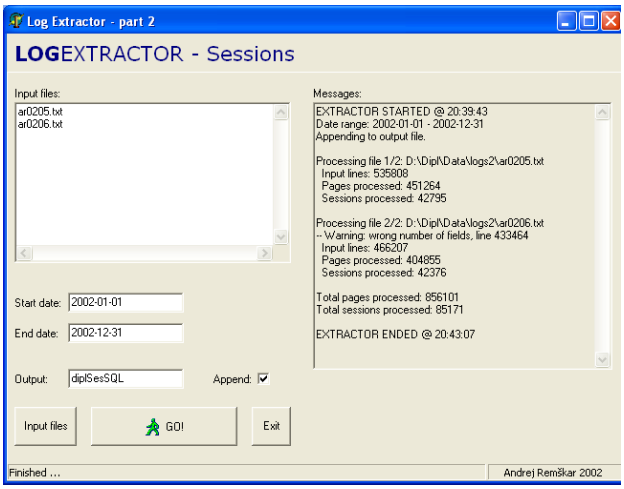
- identifikacija seje,
- identifikacija zaključka seje,
- izločanje 'smeti' v podatkih.

Osnovno orodje za identifikacijo seje je IP naslov obiskovalca. S precejšnjo natančnostjo lahko predpostavimo, da zahteve, ki v določenem časovnem obdobju prihajajo z enega IP naslova, tvorijo eno sejo. (Zavedamo se problema požarnih zidov, zaradi katerih imajo vsi obiskovalci iz enega podjetja lahko isti IP naslov.) V našem primeru nam dodatno pomagata še dve oznaki: sam spletni strežnik vsako sejo označi s 24-mestnim nizom (najdemo ga zapisanega v piškotku v spremenljivki *ASPSessionID)*, pa tudi sama aplikacija Merkurjeve spletne trgovine vsaki seji priredi 20-mestno kodo in z njo opremi vse zahteve. Ta koda (z oznako *MySes)* se prenaša med ostalimi parametri zahtev v ukaznem nizu (t.i. *Query String*).

Smatramo, da je seja zaključena, ko nek obiskovalec več kot 20 minut ne sproži nove zahteve. Vsako kasnejšo zahtevo (tudi tako, ki vsebuje iste oznake, npr. IP številko ali oznako seje) obravnavamo kot začetek nove seje. Trajanje seje določa razlika v času med prvo in zadnjo zahtevo, ki ji prištejemo 10 sekund za ogled zadnje strani.

Določene zahteve izločimo iz baze, če prepoznamo, da nam glede na namen podatkovnega skladišča ne bodo koristile. Med drugimi so take zahteve po prikazu neveljavnih strani, zahteve za prikaz strani za urednike spletne trgovine in za poslovne partnerje, pa tudi vse zahteve, ki jih sprožajo t.i. roboti (programi, ki samodejno prebirajo strani na internetu in jih analizirajo za potrebe iskalnikov). Neustrezne zahteve prepoznamo na več načinov: na osnovi samega teksta zahteve (določene strani, ki nas v analizi ne zanimajo, se tako nahajajo v imenikih */urednik/* ali */metalurgija/*), na osnovi kode rezultata (neveljavne strani, npr. koda 404 ob zahtevi za stran, ki ne obstaja) ali oznake uporabniškega agenta, v kateri se roboti prepoznavno podpišejo (tako se program-robot iskalnika najdi.si podpiše s *»TridentSpider«*).

Pripravili smo podatkovno bazo (zaenkrat z eno samo tabelo, v kateri bodo vsi podatki o sejah) in aplikacijo v orodju Borland Delphi. Aplikacija za vsak zapis preveri, h kateri seji sodi, in beleži osnovne podatke o sejah (podatkovna struktura je prikazana na sliki 7.3). Shranjene podatke ob preteku seje zapiše v tabelo v podatkovno bazo.



Slika 7.2. Program za združevanje podatkov o sejah

```
type
  sessionType=record
    // osnovni podatki o seji
     ssStartTime, ssEndTime: TDateTime;
     ssDurationSec: integer;
    // oznake
     ssAspSes: string [24];
     ssMySes, ssVisitorID, ssIP: string [20];
    // prva in zadnja zahteva, vir
     ssFirstUri, ssFirstQS, ssLastUri, ssLastQS, ssRef: string;
    // števci ogledov strani po področjih
     ssPageCount, ssSearchCount, ssPrDetCount, ssPrCatCount, ssPrTotCount,
     ssInfoTotCount, ssArtDetCount, ssArtTotCount: integer;
    // oznake o ogledu posameznih področij
     ssLProducts, ssLActions, ssLNews, ssLArticles, ssLInfo, ssLInfoPrivacy,
     ssLInfoDelivery, ssLInfoHelp, ssLInfoPayment, ssLFun, ssLMM: boolean;
    // podatki o nakupu
     ssBuyStatus: (buyNone, buyCart, buyCheck, buyDone);
     ssInvoiceID: string;
     ssInvoiceSum: integer;
    // kazalec na naslednji zapis v seznamu
     next: PSessionType
  end;
```

Slika 7.3. Podatkovna struktura, v kateri se zbirajo podatki o sejah

Rezultat drugega koraka postopka pretvorbe je končnih 206.000 zapisov o posameznih sejah. Te bomo v naslednjem koraku pretvorili v dimenzijski model, primeren za obdelavo v podatkovnem skladišču.

7.1.3. Kreiranje dimenzijskih tabel, finalizacija sheme

Zadnji korak bomo v celoti izvedli v okolju podatkovne zbirke MS SQL Server 2000. S pomočjo SQL poizvedb in ukazov bomo iz pripravljenih podatkov izluščili

podatke o sejah, ki bodo v naši končni zvezdni shemi shranjeni v dimenzijskih tabelah, in predelali tabelo sej, da bo primerna za tabelo dejstev. Postopek je za vse dimenzije podoben:

- s pomočjo ukaza `SELECT DISTINCT` iz tabele sej izločimo različne vrednosti ključev,
- zapise v tabeli dejstev opremimo s številko ključa,
- dimenzijsko tabelo dopolnimo z opisnimi podatki, ki bodo pripomogli k lažji uporabi podatkovnega skladišča.

Na koncu iz tabele dejstev pobrišemo še odvečna polja – podatke, ki so sedaj shranjeni v dimenzijskih tabelah.

Na primeru dimenzije Čas je obdelava enostavne dimenzijske tabele prikazana na spodnji sliki.

```
/* KREIRANJE DIMENZIJSKE TABELE dTIME */
USE diplSes

-- pobrišimo staro tabelo ------------------------------------------------------
TRUNCATE TABLE dTime

-- določimo enolične ključe (HH:MM) in jih prepišimo v tabelo ----------------------
INSERT INTO dTime
  SELECT DISTINCT
    convert (int, substring(convert(char(5), fSe.ssStartTime, 108), 1, 2)) AS tHour,
    convert (int, substring(convert(char(5), fSe.ssStartTime, 108), 4, 2)) AS tMin,
    convert (char(5), fSe.ssStartTime, 108) AS tText,
    0 AS NumMinutes
  FROM fSessions AS fSe
  ORDER BY tText

-- zapisom v tabeli dejstev vpiši prave ključe dimenzije dTime ----------------------
UPDATE fSessions
  SET TimeKey = dTime.TimeID
  FROM dTime INNER JOIN fSessions
    ON dTime.TextTime = convert (char(5), fSessions.ssStartTime, 108)


-- dodelaj dimenzijsko tabelo dTime -------------------------------------------------
-- - določi zaporedno številko minute
-- - vpiši opis - obdobje dneva,
UPDATE dTime
  SET NumMinutes = (60 * TimeHour) + TimeMinutes,
      TimeSpan = CASE
        WHEN (TimeHour >= 1 AND TimeHour <= 6) THEN 'Ponoči (01:00 - 06:59)'
        WHEN (TimeHour >= 7 AND TimeHour <=12) THEN 'Dopoldne (07:00 - 12:59)'
        WHEN (TimeHour >= 1 AND TimeHour <=18) THEN 'Popoldne (13:00 - 18:59)'
        ELSE 'Zvečer (19:00 - 00:59)'
      END
```

Slika 7.4. SQL ukazi za kreiranje in ureditev dimenzijske tabele dTime

Zaradi neurejenih podatkov je bilo v praksi malce težje obdelati dimenzijo Vir, zaradi obsega pa tudi dimenzijo Stran, ki se uporablja za opis vstopnih in zadnjih strani seje. V slednji smo opise strani dopolnili z imeni kategorij, kjer je to smiselno (pri izdelkih in nasvetih). Tabela obiskovalcev zaenkrat ne vsebuje nobenih osebnih ali demografskih podatkov, primernih za analizo.

Izkazalo se je, da je pretvorba podatkov zelo zahtevno opravilo, ki mora biti izvedeno dosledno, saj neposredno vpliva na kakovost podatkov v podatkovnem skladišču. Nekatera manjša dopolnilna dela (npr. urejanje opisov) bo verjetno še precej časa težko popolnoma avtomatizirati in bi v praksi ostala v domeni urednikov spletnega mesta.

7.2. Microsoft SQL Server 2000 Analysis Services

Microsoft SQL Server 2000 Analysis Services (v nadaljevanju Analysis Services) je strežnik za sprotno analizo *(OLAP – online analytical processing)* in odkrivanje zakonitosti v podatkih *(data mining)*. V njem lahko zgradimo in urejamo večdimenzionalne podatkovne kocke, katerih glavni namen je omogočiti končnim uporabnikom in drugim aplikacijam hiter dostop do podatkov v podatkovnih skladiščih. Strežnik sam dopolni podatke z zbirnimi vrednostmi, ki bistveno pohitrijo dostop in analize. Ustrezno zna obdelati in upoštevati tudi nove ali spremenjene podatke. Čeprav je Analysis Services del podatkovne zbirke Microsoft SQL Server 2000, lahko kot vir uporabi tudi podatkovno skladišče v drugih podatkovnih zbirkah (npr. Oracle, DB2 in katerakoli baza, dostopna preko standardnega vmesnika ODBC).

Delo s strežnikom Analysis Services poteka preko orodja Analysis Manager. V njem kreiramo in urejamo analitične podatkovne zbirke ter izvajamo poizvedbe. Najprej določimo vir podatkov, nato definiramo podatkovne kocke. Vsaki kocki določimo tabelo dejstev in definiramo dimenzije. Z vgrajenim pregledovalnikom lahko podatke v podatkovnih kockah poljubno pregledujemo, jih režemo in omejujemo glede na različne dimenzije *(slice and dice)*, pa tudi raziskujemo v globino *(drill down)*. Poleg osnovnih zbirnih funkcij, kot sta seštevanje in štetje, lahko z vgrajenim jezikom MDX *(Multidimensional Expressions Language)* na osnovi podatkov v tabeli dejstev definiramo tudi dodatna, izračunana dejstva. [13, 14]

7.3. Rezultati analize

V zaključku praktičnega dela naloge bomo predstavili nekaj konkretnih analiz, ki jih bomo izvedli nad pripravljenim dimenzijskim podatkovnim modelom s programom Analysis Services. Na Merkurjevo željo v rezultatih niso izpostavljene konkretne prodajne številke.

Za posamezne analize bomo v programu definirali podatkovne kocke z osnovnimi dejstvi o sejah (število sej, skupno trajanje sej in skupno število pogledanih strani) ter izračunanimi dejstvi, kot so povprečno trajanje seje in povprečno število strani

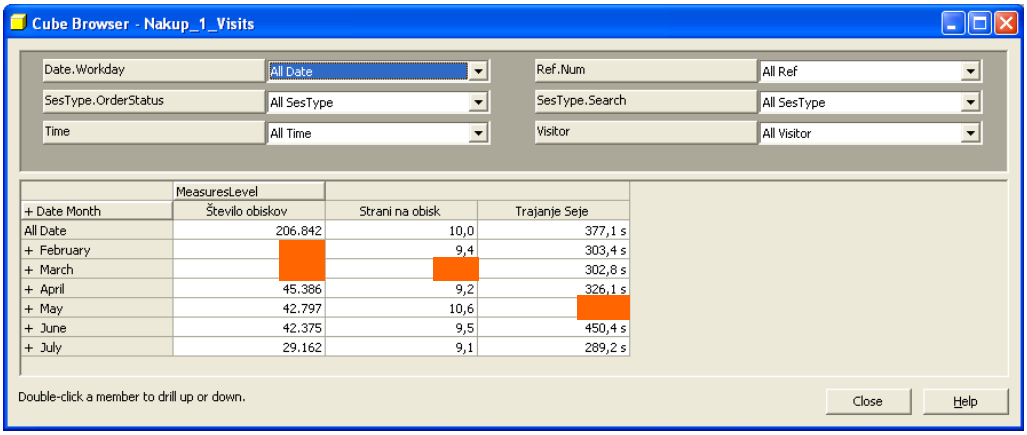
na sejo. Kocki bomo dodali dimenzije, ki so v konkretnem primeru pomembne (čas, datum, tip seje, obiskovalci, viri obiskovalcev, vstopne in izstopne strani).

Za lažje razumevanje izrazov, ki se pojavljajo med rezultati, so v spodnji tabeli navedena dejstva, ki so na voljo ali smo jih izračunali.

Dejstvo	Izračun
Število obiskov	Podatek
Trajanje seje	Podatek
Število strani	Podatek
Število nakupov	Podatek
Znesek nakupa	Podatek
Povprečno število strani	Skupno število strani / število sej
Povprečno trajanje obiska	Vsota trajanj sej / število sej
Verjetnost nakupa	Število nakupov / število obiskov
Povprečna vrednost nakupa	Vsota nakupov / število nakupov
Povprečna vrednost obiska	Vsota nakupov / število obiskov

7.3.1. Analiza časovne dimenzije

V prvem primeru se osredotočimo samo na datumsko in časovno dimenzijo obiskov. Iz osnovnih dejstev (vsota trajanja sej, skupno število pogledanih strani) izračunamo povprečno trajanje seje in povprečno število strani v eni seji ter ju spremljamo skozi obdobje.



Slika 7.5. Gibanje števila obiskovalcev, povprečnega števila pogledanih strani in trajanja seje

Poleg pričakovanih ugotovitev (pozitivni učinki rednih prodajnih akcij in pošiljanja e-novic po elektronski pošti) smo v podatkih našli zanimiva dejstva o vplivu

nagradnih iger. Opazimo namreč bistven skok obiska v marcu (od sredine marca dalje je dnevno število obiskovalcev praktično podvojeno) in podaljšanje trajanja seje v maju.

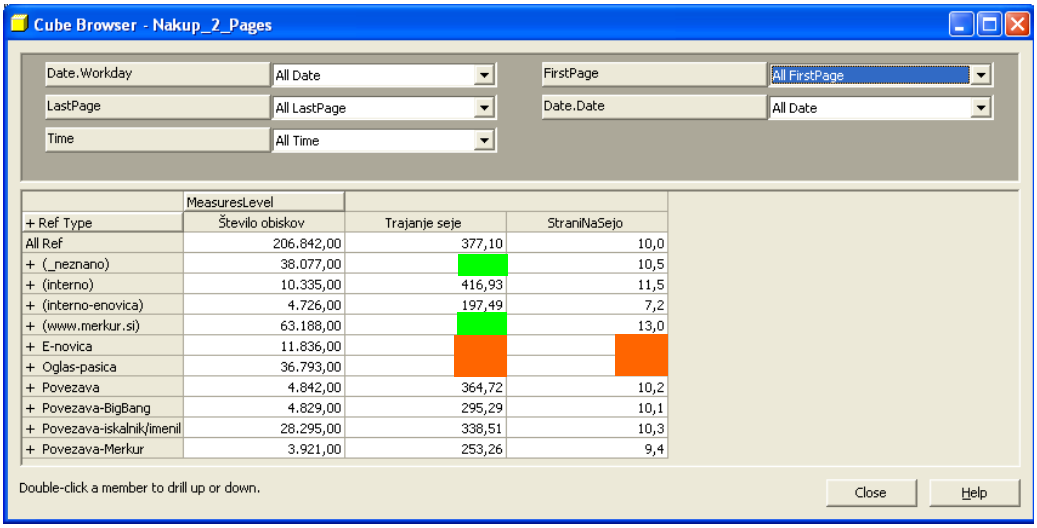
Merkur je imel v analiziranem obdobju dve večji nagradni igri – velikonočno iskanje pirhov in nogometno igro. Značilnost prve igre je bila, da so morali obiskovalci po spletnem mestu poiskati določeno število naključno razporejenih pirhov; med stranmi so očitno prehajali precej hitreje kot običajno, saj se je kljub za tretjino večjemu številu pogledanih strani v eni seji povprečno trajanje seje celo zmanjšalo. Nasprotno pa se je v času trajanja nogometne nagradne igre (enostavno streljanje na gol, vpis na internetno lestvico in na koncu nagrade za najboljše) trajanje obiska zelo podaljšalo, saj so se obiskovalci očitno zelo radi igrali. V maju in juniju se je ohranilo visoko število obiskovalcev, ki je v juliju zaradi zaključka nagradne igre in dopustov pričakovano upadlo, a ostalo na bistveno višjem nivoju kot pred začetkom nagradnih iger.

Pregled sej po urah je pokazal, da je velika večina obiskov v dopoldanskih in zgodnjih popoldanskih urah (iz službe).

Ko smo v model vključili še podatke o nakupih, smo ugotovili, da je povprečna vrednost nakupov od februarja do julija narasla skoraj za polovico. Ob delavnikih vrednost nakupa sicer za slabih 10 % presega tisto ob vikendih in praznikih, a je zaradi veliko večje verjetnosti nakupa (skoraj 40 % boljše razmerje med nakupi in obiski) ob dela prostih dnevih preračunana vrednost enega obiska (prihodek, preračunan na en obisk) kar za 22 % večji. Znotraj dneva je število nakupov največje zgodaj dopoldne (ob začetku službe), zgodaj popoldne (po odmoru za kosilo) ter zgodaj zvečer (uporabniki se od doma na internet priključijo po pocenitvi impulza ob 19. uri).

7.3.2. Analiza virov in vstopnih strani

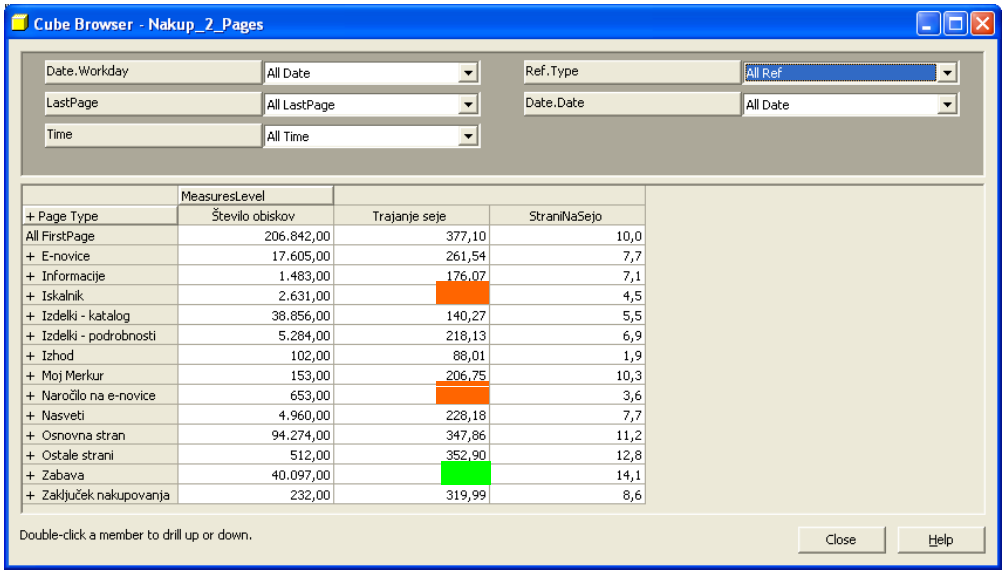
Dimenzija virov vsebuje podatke o tipu vira (v grobem gre povezave iz Merkurjevih novic, ki jih naročnikom pošiljajo po elektronski pošti, povezave z oglasnih pasic, povezave v imenikih in iskalnikih ter druge povezave). Ugotoviti želimo, ali je obnašanje obiskovalcev odvisno do tega, od kod prihajajo.



Slika 7.6. Pregled parametrov seje v povezavi z virom

Hitro lahko ugotovimo, da so najboljši tisti obiskovalci, ki so se na spletno mesto podali sami – v brskalnik so neposredno vpisali Merkurjev spletni naslov *nakup.merkur.si* (oznaka vira (_neznano)) ali *www.merkur.si* (oznaka vira (www.merkur.si)). Dejstvo je lahko razložiti – naslov vpišemo v spletni brskalnik le takrat, ko nas stran resnično zanima. Tisti obiskovalci, ki prihajajo prek elektronskih sporočil (e-novice), še posebej pa tisti, ki so jih privabile Merkurjeve oglasne pasice, na spletnem mestu porabijo bistveno manj časa od povprečja. Povprečen obisk traja dobrih šest minut; obiskovalci, ki jih je privabila e-novica, se na spletnem mestu zadržijo le slabe štiri minute, tisti, ki so kliknili na spletni oglas, pa celo manj kot dve minuti – torej trikrat manj od povprečja. Slednji so tudi obupno slabi kupci – verjetnost za odločitev za nakup je pri obiskovalcih, ki pridejo prek oglasne pasice, kar 10-krat manjša od povprečne; pri e-novici bistvenega odstopanja od povprečja ni. Na drugi strani se obiskovalci, ki se sami odločijo za obisk (vir je neznan), na spletnem mestu zadržijo v povprečju skoraj devet minut, tudi možnost odločitve za nakup je pri njih nadpovprečno visoka.

Z vrtanjem v globino lahko v tej podatkovni kocki v dimenziji virov poiščemo bolj ali manj učinkovite spletne pasice ter elektronska obvestila, ki so pripeljala več obiskovalcev in kupcev (ali daljše seje) od ostalih, ravno tako lahko spremljamo učinkovitost posameznih virov skozi čas.



Slika 7.7. Pregled parametrov seje v povezavi s tipom vstopne strani

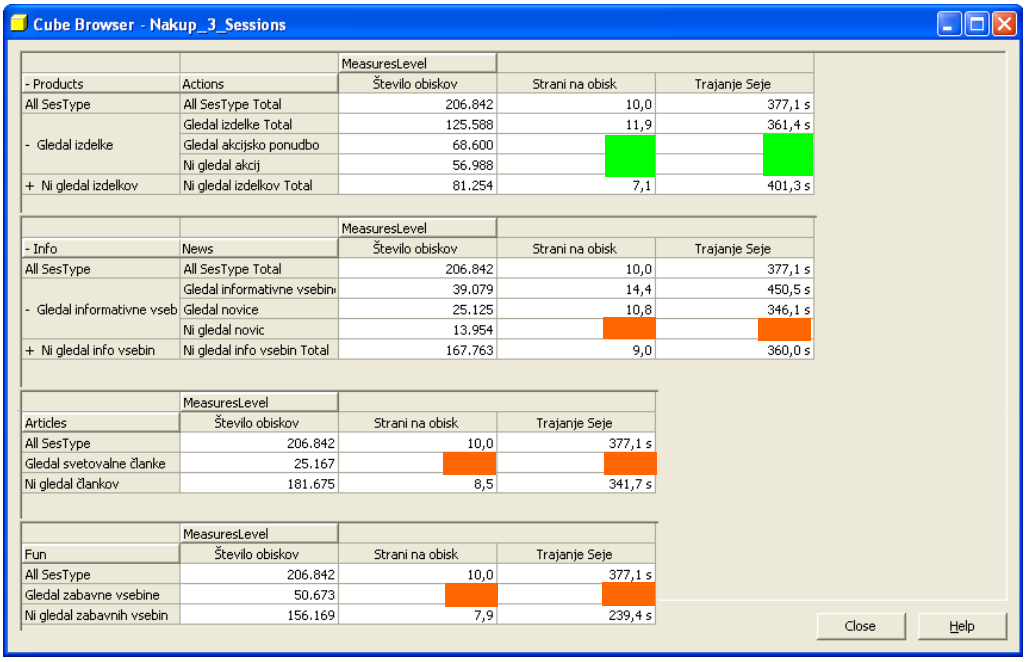
Na zgornji sliki so prikazani parametri seje v odvisnosti od tipa vstopne strani. Če izločimo manj pomembne tipe strani z minimalnimi obiski (Moj Merkur, Izhodna stran, Zaključek nakupovanja) in na hitro ugotovimo, da preko osnovne strani vstopajo zelo »povprečni« obiskovalci, najbolj opazimo kratke seje tistih, ki najprej pridejo na iskalnik ali naročijo e-novice, ter dolge seje tistih, ki se na spletnem mestu zabavajo. Kar se tiče iskalnika, ki sploh ni tipična vstopna stran spletnega mesta: kratke seje so najverjetneje posledica oglasne pasice z vključenim iskalnim poljem; uporabniki so lahko že kar v samo pasico vpisali iskani pojem, a se je izkazalo, da so pogosto iskali precej splošne pojme, ki niso povezani z Merkurjevo spletno trgovino, in so zato razočarani zelo hitro zapustili strani.

Tisti, ki so se prišli na spletno mesto zabavat, sicer niso dobri kupci, a moramo upoštevati dejstvo, da so spletne igre namenjene predvsem drugim ciljem: vzpostavljanju vednosti o spletnem mestu ter pridobivanju elektronskih naslovov za pošiljanje e-novic.

7.3.3. Analiza tipa seje

Vsaka uporabniška seja ima v dimenziji Tip seje označene lastnosti, vezane predvsem na obisk posameznih področij spletnega mesta. Tako v tej dimenziji vidimo, ali je obiskovalec gledal izdelke, ali je gledal akcijske izdelke, ali je bral svetovalne vsebine in informativne vsebine, ali je obiskal področja, namenjena zabavi in končno tudi, ali je opravil nakup. Poleg tega imamo označeno, če nakupa ni opravil, pa je prej dodajal izdelke v nakupovalni voziček in tudi, ali je poskusil

zaključiti nakup, pa mu to ni uspelo. Z analizo obiskov glede na dimenzijo tipa sej bomo dobili odgovore o vplivu posameznih področij na parametre seje.



Slika 7.8. Pregled parametrov seje v povezavi s tipom seje oz. obiskanimi vsebinami

Kar se tiče ogleda izdelkov in znotraj tega akcijskih izdelkov (podatki v prvi tabeli, označeni z zeleno), je zanimiva ugotovitev, da se obiskovalci, ki gledajo predvsem izdelke v akcijski ponudbi, v trajanju obiska in možnosti odločitve za nakup ne razlikujejo od tistih, ki predvsem gledajo izdelke na rednih prodajnih policah.

Pri analizi opazimo tudi, da obiskovalci, ki ne berejo informativnih vsebin, na spletnem mestu ostanejo manj časa (in se tudi redkeje odločijo za nakup); na drugi strani je verjetnost nakupne odločitve pri tistih, ki te vsebine obiščejo, skoraj trikrat večja od povprečja. Še posebej pozitivno izstopajo tisti uporabniki, ki berejo predvsem informacije o samem postopku nakupa, dostavi, vračilu in podobno, ne pa aktualnih novic.

Na dolžino obiska in tudi na nakupno odločitev pa vsekakor vplivata dva druga dejavnika: obisk zabavnih vsebin (dolžina obiska je bistveno nadpovprečna, nakupna odločitev pa zelo malo verjetna) in obisk svetovalnih vsebin (dolžina obiska nad deset minut ali skoraj dvakrat več od povprečja, rahlo povečana verjetnost odločitve za nakup).

Za spletnega trgovca je zelo pomemben tudi podatek, koliko obiskovalcev dejansko zaključi začet nakup. V Merkurjevi trgovini lahko ugotovimo, da nakup do konca

izvede le dobra desetina obiskovalcev, ki so izdelke naložili v voziček; to zaradi »raziskovalne« naravnanosti obiskovalcev v fazi, ko spletno nakupovanje še ni tako razvito, ni tako presenetljivo. Izkaže se tudi, da nakup do konca izvede tretjina obiskovalcev, ki postopek nakupa nadaljujejo preko nakupovalnega vozička do zaslona, na katerem morajo vpisati osebne podatke in podatke o načinu plačila.

V tem poglavju smo navedli le nekaj ugotovitev, ki so na voljo v podatkovnem skladišču; zaradi možnosti skoraj takojšnjega kreiranja poljubnih vpogledov, dodajanja novih izračunanih dejstev in razrezovanja ter vrtanja po dimenzijah je količina koristnih informacij, ki so dejansko na voljo in lahko zelo pripomorejo pri odločitvah urednikov, povezanih z nadaljnjim razvojem spletne trgovine, praktično neomejena.

8. Zaključek

V diplomski nalogi smo s tehniko dimenzijskega modeliranja zasnovali podatkovno skladišče, namenjeno analizi dogajanja v spletni trgovini. Glavni vir podatkov je dnevnik spletnega strežnika, v katerem so zapisana vsa dejanja obiskovalcev. Na osnovi poznavanja razpoložljivih podatkov in zahtev končnih uporabnikov smo izdelali tri dimenzijske modele, ki opisujejo posamezne uporabniške seje, oglede strani in nakupe izdelkov. Z analizo podatkov v teh modelih lahko uredniki spletne trgovine dobijo zelo podrobne odgovore in pojasnila o dejavnikih, ki vplivajo na vedenje obiskovalcev in posledično na uspešnost trgovine. Sledi obiskovalcev so med redkimi viri podatkov, v katerih je vidno tudi vedenje obiskovalcev, ki niso sklenili posla (nakupa).

Ob praktičnem preizkusu modela za analizo uporabniških sej na podatkih Merkurjeve spletne trgovine smo ugotovili, da je z začetno pripravo podatkov (predvsem urejanjem dimenzij) veliko dela, ki ga ni mogoče enostavno avtomatizirati. Vhodni podatki namreč še niso v celoti v obliki, ki bi omogočala pretežno avtomatsko obdelavo in prenos v podatkovno skladišče; z manjšimi dodelavami spletne aplikacije bi ta problem lahko omilili. Program Microsoft SQL Server 2000 Analysis Services se je izkazal kot zelo ustrezen pripomoček, s katerim lahko končni uporabnik v dobro pripravljenih podatkih odkrije marsikatero koristno zakonitost; z analizo Merkurjevih podatkov smo prišli do izredno zanimivih ugotovitev, ki jih lahko uredniki neposredno uporabijo pri odločitvah o razvoju in oglaševanju spletne trgovine.

Predlagano podatkovno skladišče lahko uporabimo ne glede na vrsto spletnega mesta, ki ga analiziramo; poskrbeti bi morali le za pretvorbo razpoložljivih podatkov v predlagani dimenzijski model.

V nalogi je nakazano, s katerimi viri bi lahko dopolnili informacije, da bi dobili še podrobnejše rezultate. V naslednji fazi bi lahko združili tudi podatke o prodaji v fizičnih trgovinah in obiskih na spletnem mestu ter preučili medsebojni vpliv, o katerem trgovci zaenkrat predvsem ugibajo. Obiskovalce bi lahko na osnovi gledanih vsebin razvrstili v nekaj skupin, za katere bi pripravili prilagojeno ponudbo. V končni fazi želimo na osnovi obiskovalčevih sledi v realnem času zaznati posamezne vedenjske vzorce in nanje takoj ustrezno odgovoriti.

Literatura

[1] R. Kimball, *The Data Warehouse Toolkit*, John Wiley & Sons, 1996

[2] R. Kimball, R. Merz, *The Data Webhouse Toolkit*, John Wiley & Sons, 2000

[3] R. Kimball, *„A Dimensional Modeling Manifesto"*, DBMS Magazine, August 1997

[4] Robert Calliau, Dan Connolly, *„A Little History of the World Wide Web",* http://www.w3.org/History.html

[5] *Raziskava RIS – Raba interneta v Sloveniji*, junij 2002, http://www.ris.org

[6] M. Porter, *„Strategy and the Internet"*, Harvard Business Review, March 2001

[7] Google Inc., http://www.google.com, 2001

[8] Forrester Research, *„Consumer Technographics Benchmark 2001",* 2001

[9] Internet Software Consortium, *„Internet Domain Survey Host Count",* julij 2002, http://www.isc.org/ds/host-count-history.html

[10] Harris Interactive, *„Shopping Spending and Satisfaction Up",* Nua Internet Surveys, januar 2002, http://www.nua.com/surveys

[11] Jupiter Communications, *„Report: Online Research Drives Offline Spending",* E-Commerce Times, maj 2000, http://www.ecommercetimes.com

[12] M. Berry, G. Linoff, *Mastering Data Mining – The Art and Science of Customer Relationship Management*, John Wiley & Sons, 2000

[13] Datapro, Gartner Group, *Microsoft Corp. SQL Server 2000 Analysis Services,* november 2000, http://www.microsoft.com

[14] Microsoft, *Microsoft SQL Server 2000 Books Online,* 1999

[15] Interni podatki podjetja Merkur, d. d.

[16] Predstavitev podjetja Merkur, www.merkur.si, avgust 2002

Izjava o samostojnosti dela

Izjavljam, da sem diplomsko delo izdelal samostojno pod vodstvom mentorja prof. dr. Viljana Mahniča. Izkazano pomoč drugih sodelavcev sem v celoti navedel v zahvali.