ANALIZA I OBRADA PODATAKA

Nastavna skripta

Autor: Igor Nazor

Split, travanj 2012

Sadržaj

Uvod

Cilj ovog kolegija je dati studentima općenit uvid u metode obrade podataka i vrste informacijskih sustava koji te podatke generiraju. Naglasak će biti stavljen na poslovne informacijske sustave i obradu podataka koje oni generiraju. Biti će govora i o različitim načinima spremanja podataka ovisno o njihovoj namjeni.

Računala već davno dobro i brzo zbrajaju i oduzimaju, te dobivanje sumarnih podataka nije neka novost. Izazov je iskoristiti procesorsku snagu računala, te primjenjujući napredne statističke i programske metode iz sirovih informacija pokušati izvući neke složenije zaključke, tj. dobiti znanje.

Primjer jednostavne obrade podataka iz transakcijskog sustava je ispis izvještaja o mjesečnoj prodaji na temelju svih računa koji su izdani kupcima, njihovih iznosa i datuma. Složenija obrada i analiza bi mogla dati odgovor na pitanje kada je optimalno vrijeme za pokrenuto pogon za proizvodnju sladoleda ove godine, na temelju svih podataka o prodaji i proizvodnji sladoleda proteklih godina i podataka o trošku pokretanja i zaustavljanja proizvodnje? Ili, kojim kupcima u supermarketu ponuditi posebne uvjete ili poklone, kako ne bi počeli kupovati u konkurentnom prodajnom centru?

Kako bi se dalo odgovor na ovakva pitanja potrebno je uzeti u obzir velik broj parametara, od kojih su neke vrijednosti direktno spremljene u bazu podataka, kao što su iznosi prodaje ili podaci o prodanim artiklima. Neke vrijednosti se mogu izračunati koristeći jednostave postupke kao što su sumiranje ili grupiranje, dok postoje parametri čije vrijednosti se mogu samo procijeniti sa većom ili manjom preciznošću koristeći statističke metode.

U prvom dijelu se govori o informacijskom inženjeringu i razvoju informacijskih sustava, te će uz povijesni pregled biti ukazano na glavne probleme i izazove sa kojima se treba suočiti.

Drugi dio obrađuje načine spremanja podataka ovisno o namjeni, te metode višedimenzijske analize i rudarenja podataka.

Informacijski inženjering

Neke od definicija informacijskog inženjeringa su [1]:

- *Primjena međusobno povezanih formalnih tehnika za planiranje, analizu, dizajn i izradu informacijskih sustava na razini cijelog poduzeća ili u njegovom većem dijelu.*
- *Niz postupaka na razini poduzeća sa svrhom davanja pravih informacija pravim ljudima u pravo vrijeme*

U proteklih nekoliko desetljeća u većim poduzećima je razvoj informacijskih sustava pratio razvoj tehnologije upravljanja u spremanja podataka. U početku su se informacijski sustavi pojedinih odjela razvijali neovisno jedni od drugih tako da je, primjerice, odjel financija je imao odvojen informacijski sustav od kadrovske službe. Nastajali su takozvani "informacijski otoci", između kojih nije bio uspostavljen tok informacija. Ako tvrtka ima urede u više država, donedavno je bila praksa da svaka država ima odvojeni informacijski sustav, što je bilo potrebno zbog razlika u zakonodavstvu, lokalnim običajima i problemom sa udaljenom podrškom korisnicima. Takvi sustavi su najčešće imajli različite strukture podataka. Problem se je pojavljivao kod izvještavanja, jer nije postojao jednostavan način da se podatke iz raznorodnih informacijskih sustava objedini, kako bi se dobila slika stanja cijelog poduzeća.

Jedan od zadataka informacijskog inženjeringa je spajanje razdvojenih informacijskih sustava u jednu logičku cjelinu, iz koje se mogu dobivati objedinjeni podaci. Prvi korak u postupku objedinjavanja je izrada *modela poduzeća*. Jedan od postupaka izrade modela poduzeća, obrađen u kolegiju "Informacijski sustavi", sastoji se od slijedećih koraka: definiranje modela podataka, definiranje modela procesa, identificiranje sudionika, te određivanje tijeka informacija između sudionika i sustava (dijagram toka podataka).

Problem nedostupnosti informacija u praksi je veći nego što se to možda naizgled čini. Određene vrste poduzeća, naročito ona koja nisu profitno orijentirana, mogu funkcionirati na takav način. Međutim, neko veliko poduzeće, koje svoj glavni proizvod mjesec dana prodaje na tržištu po krivoj cijeni, zbog netočnih informacija koje je uprava dobila iz lošeg informacijskog sustava, će se sigurno naći u velikim problemima. Ovisnost organizacije o kvalitetnim informacijama iz poslovnog sustava raste sa njenom veličinom i geografskom dislociranosti njenih ureda. U manjim organizacijama je čest slučaj da "svatko poznaje

svakoga" i "svatko zna sve", pa je i neformalni tok informacija unutar poduzeća prilično efikasan. U ovakvom poduzeću će se teško dogoditi da se neka bitna kriva odluka donese zbog krivih podataka u softveru. Što je organizacija veća, i što se prostire na većem geografskom području, osobni kontakt među njenim zaposlenicima je manji ili uopće ne postoji, a bitne odluke na vrhu hijerarhije se u većoj mjeri donose na temelju izvještaja iz informacijskog sustava. Važnost raspolaganja točnim informacijama iz poslovanja za ovakva poduzeća je pitanje opstanka na tržištu.

Primjerice, jedna međunarodna firma koja ima predstavništvo i u Hrvatskoj, ima tvornice i predstavništva u preko 110 država u svijetu, ulaže nekoliko desetaka milijuna eura u uvođenje zajedničkog informacijskog sustava. U vrijeme dok je predstavništvo te firme u svakoj državi imalo svoj sustav, morao je postojati čitav niz ljudi čiji je jedini posao bio dopisivati se i telefonirati u sve države i bilježiti podatke o prodaji, troškovima, zaradi, marketinškim akcijama i slično. Tako prikupljeni podaci bi se objedinili i poslali upravi, koja bi na temelju njih donosila odluke. Prilikom "prepisivanja" podataka, kopiranja i slanja E-mailom, velika je vjerojatnost nastanka grešaka, koje se onda šire kroz sustav. Nadalje, plaćanje velikog broja zaposlenika čiji je praktički jedini posao da prikupljaju informacije i prenose ih nadređenima u lancu stvara velik trošak u poslovanju. Trošak je još i veći ako na temelju krivih informacija rukovodstvo takve tvrtke donese neku važnu stratešku odluku, primjerice o investiranju u novu tvornicu u nekoj državi.

Ako se u obzir uzme i konkurentno okruženje, tvrtke moraju pratiti konkurenciju, jer ako oni mogu brzo i točno donositi odluke, morate i vi.

Investicija u informatizaciju

Žurba za uvođenje novih tehnologija stavlja velik pritisak na odjele i osobe u organizaciji koji su zaduženi za funkcioniranje informacijskog sustava. Složene i kvalitetne aplikacije je potrebno implementirati u kratkom vremenu i uz prihvatljiv trošak, a moraju u potpunosti zadovoljavati potrebe krajnjih korisnika. Često je onima koji se bave informatizacijom teško dati valjane argumente za velike investicije u informatiku. Naime, takve investicije poduzeću ne donose direktnu zaradu, pa često i nisu visoko na listi prioriteta. Pravi razlog za kvalitetnu informatizaciju nažalost često postane jasan tek kada bude prekasno, tj. kada zbog lošeg funkcioniranja poslovnog sustava nastane poslovna šteta.

Kada se govori o koristima od informatizacije, pored direktnih, treba uzeti u obzir i one indirektne, koji nisu vidljivi na prvi pogled. Naime, iznimno skupe i kompleksne aplikacije koje se koriste u velikim korporacijama, često zahtijevaju dugotrajnu obuku krajnjih korisnika, koji opet velik dio posla moraju obavljati "pješke". S druge strane, loše postavljen informacijski sustav znači i velik gubitak vremena u poslovanju.

Jedan od primjera neusklađenog informacijskog sustava je rad odjela za dodjelu kredita. U jednom primjeru iz života, pri pokušaju stranke da dobije gotovinski kredit, službenica je u jednoj aplikaciji pogledala stanje tekućeg računa, zatim je u drugoj potražila obaveze po kreditima, a na papiru je bio popis svih kredita koje banka nudi. Kako ove aplikacije nisu međusobno povezane, podatke je iz jedne u drugu trebalo prepisati ručno. Službenici je trebalo oko pola sata da pruži informaciju o tome koliki iznos kredita, po kojim kamatama i uz koja jamstva smije odobriti.

Ako se ovo vrijeme pribroji utrošenom vremenu stranke, i pomnoži sa, nekoliko desetaka tisuća ovakvih zahtjeva godišnje, dolazi se do utroška radnog vremena koji bi opravdao čak i najveću investiciju u kvalitetan informacijski sustav.

Efikasno upravljanje informacijama je važan faktor uspjeha

Tvrtke shvaćaju da računala mogu učiniti puno više i automatizirati ono što se je prije radilo ručno. Računala mijenjaju način na koji tvrtke posluju, mijenjaju njihove odnose sa kupcima i dobavljačima, način na koji se donose odluke, čak i organizacijsku strukturu samih poduzeća, te ih povezuju sa novim kupcima i dobavljačima. U nekim slučajevima razvijaju se potpuno nove industrijske grane.

Mogućnosti i složenost aplikacija koje informacijski povezuju čitavo poduzeće također raste. Puno je teže informatizirati kompletni proizvodni proces, nego odvojeno njegove dijelove. Tvrtke se danas u poslovanju više oslanjaju na ovakve sustave, ali s druge strane su i više ovisne o njima.

Potpuna automatizacija svih poslovnih procesa je danas praktički standard u nekim industrijama. Primjeri su rezervacijski sustavi aviokompanija, ili sustavi proizvođača automobila gdje je direktno iz prodajnog salona moguće pokrenuti proizvodnju modela

automobila sa dodacima prema želji naručitelja. Industrija odjeće, primjerice, mora efikasno pratiti modne trendove (analizirati prodaju različitih modela odjeće po regijama), kako bi mogla brzo reagirati na promjene u navikama potrošača.

Alati za razvoj programskih rješenja

Aplikacije koje pokrivaju poslovanje cijelog poduzeća su iznimno kompleksne. Veze među njima također. Ako se struktura jedne od tih aplikacija promijeni, teško je predvidjeti kakav će to imati utjecaj na ostatak sustava.

Ove promjene ilustrira primjer iz nedavne prakse. Nedavno je uveden OIB, koji zamjenjuje matični broj poduzeća. Kako to najjednostavnije promijeniti u našem sustavu? Možemo u polje "MB" ukucati OIB, i eventualno izmijeniti nekoliko izvještaja, da ne prikazuju riječ "MB", nego "OIB". Međutim u aplikaciji se taj podatak koristi za izračun plaće, tako da se u tablici "plaće" zapisuje matični broj radnika. Sada kada je taj broj promijenjen, više ne možemo pregledati stare plaće, jer sustav nema vezu prema radniku. Problem riješimo tako da kreiramo program (SQL QUERY) koji će u svim tablicama izmijeniti MB u novi OIB. Na kraju godine ispisujemo poreznu karticu, te vidimo isti problem... Neke od tablica imaju za Matični broj predviđeno 8 mjesta, a neke više. Do sada nismo imali problema, jer podaci nisu bili dulji od 8 znakova. Ako propustimo izmijeniti te postavke na nekim tablicama, greška se može ukazati tek nakon nekoliko godina, kada se pojave 2 radnika sa istih prvih 8 znakova OIB-a....

Rješenje ovog problema je korištenje kvalitetne metodologije za razvoj softvera. Jedan od načina je i korištenje softvera koji u sebi ima informacije o strukturi našeg poduzeća, pa tako i o strukturi (modelu) podataka. Ako se izmjene rade na razini modela podataka, one se automatski propagiraju u sve dijelove aplikacije, gdje su potrebne. Automatski se mijenja izgled formi, izvještaja, pravila za unos i sl. Ovaj koncept, korištenje alata za modeliranje nekog sustava i automatsko generiranje aplikacija na temelju tog modela se zove "Computer - Aided Systems Engineering" - CASE. Iako već dugo postoje, u praksi se u nekim segmentima razvoja aplikacija koriste više, a nekima manje. Vrlo su korisni u dijelovima koji su strogo i formalno definirani, kao npr. generiranje relacijskog modela baze podataka na temelju modela. U drugim segmentima, kao što su generiranje obrazaca i izvještaja, se koriste kao pomoć dok finalna verzija u pravilu zahtijeva "ručne" dorade.

Jedan od rudimentalnih primjera generiranja objekata aplikacije pomoću modela se može vidjeti i u aplikaciji MS-Access, gdje se na temelju "čarobnjaka" i modela podataka pojednostavljeno generiraju obrasci i izvještaji.

Od podataka do shvaćanja - hijerarhija znanja

Dobro je razlučiti razliku između osnovnih "građevnih blokova" analize informacija ("DIKW hijerarhija") [W15]. U literaturi [6] postoje razna tumačenja ovih pojmova, koja se često razlikuju ovisno tome koje područje obrade informacija se želi objasniti (ljudsko učenje, strojna analiza podataka, itd.)

Podatak

Podatak može biti bilo kakav simbol, znak ili više njih. Sam za sebe nema neko značenje. Podaci mogu postojati u bilo kojem obliku, upotrebljivom ili neupotrebljivom. Primjer podatka je "kljhlkjhl", "25", "sunčano". Podatak je ulazna veličina za proces donošenja odluke.

Informacija

Informacija je podatak kojem je dodan neki značaj relacijskom vezom. Da bi podatak postao informacija potrebno ga je *interpretirati*. Ako se neki brojčani podatak nalazi u polju tablice koje ima naslov "Starost", onda imamo informaciju. Još uvijek ne znamo tko je star 25 godina, ili zašto je to bitno, ali znamo na što se podatak odnosi.

Primjerice, podatak "opiu23ypj08y73" ljudima nema nikakvo značenje, dok "sunčano vrijeme u Splitu", "visoka cijena hrane" imaju. Ovakvi podaci nemaju za svakoga identično značenje, već ono ovisi o *znanju* koje *interpretator* podataka posjeduje. Pomoću informacije odgovaramo na pitanje "tko","što", "kada", "gdje".

Znanje

Sa stajališta računalne analize podataka možemo reći da znanje predstavlja određena količina prikupljenih informacija na neku temu, kao i pravila za generiranje novih informacija (traženje uzoraka u podacima). Pomoću znanja se **podaci** transformiraju u **informacije**. (interpretacija podataka).

Na primjer, podatak "sunčano vrijeme" ćemo, ovisno o našim prethodnim iskustvima o sunčanom vremenu u određenom kontekstu, sebi predočiti na određen način. Ta informacija će za nas sigurno imati potpuno drukčije značenje ako živimo u Africi, nego ako živimo na sjeveru Evrope.

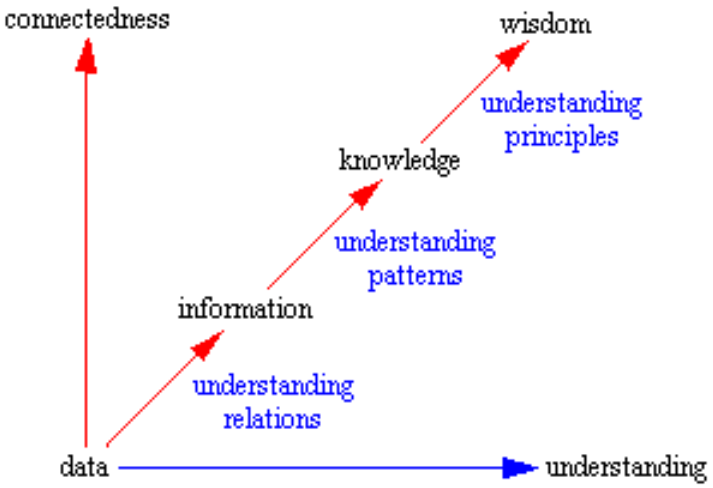
Pomoću znanja u određenim okolnostima je moguće **generiranje novog znanja** (učenje). Primjerice, djeca znaju napamet tablicu množenja do 10, ali to znanje ne mogu primijeniti da izračunaju 234 x 213, jer to znanje nisu stekli. Međutim, ako nauče način na koji se množe brojevi, to će biti u stanju.

Računalni sustavi spremanju znanje na različite načine. Sustavi koji spremaju znanje koje je definirano nizom pravila za obradu podataka zovu se **ekspertni sustavi.** Ekspertni sustav može, na temelju postavljenih pitanja i danih odgovora, zaključiti da, primjerice ako automobil ne pali a rade prednja svjetla, nije problem u akumulatoru.

Osim prikazivanja znanja pomoću formalnih pravila moguće je memorirati različite instance, te ih kasnije u najboljoj mogućoj mjeri ponavljati. Ovakav pristup spremanju znanja imaju sustavi pod nazivom *neuronske mreže*. Pomoću znanja se odgovara na pitanje "kako", te se u nekim slučajevima mogu predviđati događaji u budućnosti.

Mudrost

Mudrost bi se mogla definirati kao primjenjivanje shvaćanja, znanja i informacija u širem kontekstu, traženje novih istina, i primjena postojećih u kontekstu nekih moralnih i etičkih vrijednosti. Mudrost je specifična za ljudska bića i nema ekvivalent u računalnom svijetu.



Hijerarhija znanja

Jedna od uobičajenih krivih pretpostavki kod obrade informacije je da, ako dobro upravljamo detaljima, krajnji rezultat će automatski biti dobar [3]. To je u pravilu netočno, i slično je tvrdnji da, ako dobro zamiješamo cement, automatski imamo dobru autocestu.

U kontekstu obrade podataka nastale u poslovno informacijskim sustavima, kvalitetno prikupljanje što većeg broja detalja o poslovnim transakcijama je neophodno kao temelj za kasnije donošenje poslovnih odluka. Međutim, to je tek prvi korak. Iz prikupljenih podataka je potrebno dobiti kvalitetne informacije, i generirati novo znanje.

Sustavi za pohranu podataka

U počecima ozbiljnijeg razvoja elektroničke obrade podataka, u 60-im godinama 20. stoljeća, podaci su se spremali na magnetne trake. Problem kod ovakvog načina spremanja podataka je i tome što, da bi se došlo do nekog podatka koji je spremljen na sredini trake, potrebno je pročitati sve podatke koji su na traci spremljeni prije njega. Ovaj način spremanja podataka se zove sekvencijalno spremanje podataka. Prosječno vrijeme dohvata informacije sa trake je 20-30 minuta.

Drugi problem je što su sekvencijalno spremljeni podaci nisu međudobni logički povezani. Primjerice, ako su na jednoj traci podaci o svim klijentima banke, a na drugoj podaci o neplaćenim računima, potrebno je dosta vremena da se ti podaci povežu u smisleni izvještaj. Polovinom 60-ih, napretkom tehnologije i pojeftinjenjem medija, spremanje na trake doseže vrhunac primjene.

U sedamdesetim godinama dvadesetog stoljeća u upotrebu dolaze diskovi, čija je glavna prednost direktan pristup podacima, a vrijeme dohvata informacija se mjeri u milisekundama. Međusobno povezivanje računala u kombinaciji sa brzim pristupom podacima omogućilo je pristup većeg broja korisnika jednom računalu, koje je imalo dovoljno procesorske snage za paralelnu obradu više zahtjeva odjednom, što je temelj današnjih on-line usluga.

U 80-ima već postoje baze podataka iste strukture kao i današnje ali, zbog njihovog malog kapaciteta i visoke cijene, te nepostojanja Interneta, i dalje funkcioniraju kao "informacijski otoci" gdje na svakoj lokaciji neke kompanije postoji zasebna baza podataka. Da bi se dobili konsolidirani izvještaji potrebni podaci se pomoću tzv. *extract programa* izvlače u podebne baze podataka koje se koriste za izvještavanje. Problem je u tome što je baza podataka dinamičan sustav, a izvlačenjem podataka iz nje dobiva se njena "slika" u jednom trenutku. Već u narednom trenutku originalna baza podataka se može izmijeniti, i njena kopija neće odgovarati stvarnom stanju. Glavni problem dupliciranja podataka je u tome što ne postoji garancija da original i duplikati imaju isti sadržaj, zbog čega informacije dobivene iz duplikata nisu vjerodostojne. Podaci dobiveni ekstrahiranjem nemaju vremensku dimenziju, tj, ne može se sa sigurnošću reći na koji trenutak u vremenu se odnose.

Problem sa informacijskim otocima je i u tome što se isti podaci u različitim izvorima drukčije označavaju. Primjerice, artibut koji sadrži stanje skladišta u jednoj datoteci može zvati STSKL, a u drugoj NIVOSKL. Da bi se izradio točan izvještaj potrebno je iz svih datoteka identificirati polja koja označavaju stanje skladišta, i izraditi program koji sve te vrijednosti kopira u isto polje. Međutim, kako su oba "informacijska otoka" dinamični sustavi, postoji vjerojatnost da će se kroz nekoliko mjeseci ili godina će se pojaviti nove datoteke sa novim podacima, pa će i proces objedinjavanja trebati raditi ispočetka.

Načini sporemanja podataka

Kod suvremenog pristupa upravljanju podacima se mogu prepoznati četiri različita principa čuvanja podataka:

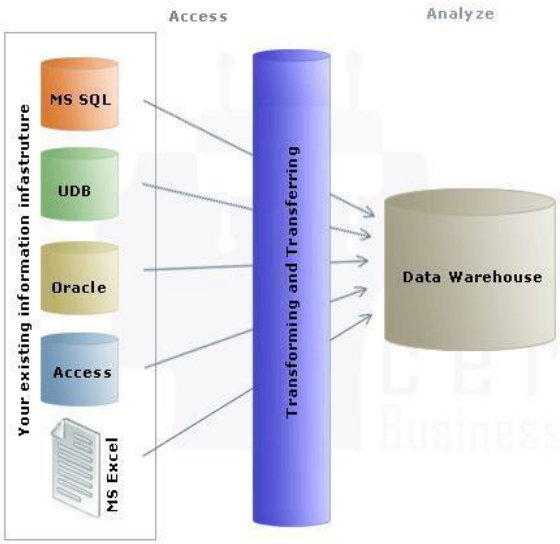
- **Operacijski (transakcijski) nivo** - podaci iz transakcija koji se upisuju onim redoslijedom kako nastaju
- **Spremište podataka** - (data Warehouse) spremaju se integrirani podaci iz povijesti poslovanja, složeni prema unaprijed određenim temama. Služi za brži pristup velikoj količini podataka koja je nastala kroz povijest.
- **Pod-skladište podataka** (Data mart) - spremaju se sumarni podaci, ovisno o njihovoj namjeni. Iz istih podataka o prodaji će, primjerice, odjel financija imati podatke o zaduženjima i bilanci pojedinog kupca, a odjel marketinga o prodaji pojedinih vrsta proizvoda po regijama. Podaci u Data Mart-ovima su denormalizirani, sumirani i oblikovani prema potrebama pojedinih korisnika. Primjer podataka u Data Mart-u je mjesečni popis prodaje po kupcima.
- **Individualni (ad-hoc) podaci** - Podaci koji sluše najčešće za individualnu upotrebu, i generiraju se po potrebi iz svoh dostupnih izvora. Primjer je kada voditelj prodaje za velike kupce može, primjerice, izvući popis svih kupaca od 2000. godine koji su kupovali u iznosu većem od 100.000 kn mjesečno i plaćali na vrijeme. Podataka ove vrste ima relativno malo, nastaju po potrebi i u svaki put mogu imati drukčiju strukturu.

Spremište podataka

Spremište podataka (engl. data warehouse) se formira na temelju podataka iz transakcijskog sustava. Podaci u njemu su integrirani i često prikupljeni iz više izvora. Glavni izazov kod formiranja spremišta je prikupljanje podataka iz različitih izvora, njihovo uspoređivanje i dovođenje u standardizirani oblik. Namjena spremišta podataka je čuvanje podataka na dučji vremenski rok (arhiviranje), te kao izvor povijesnih podataka za daljnju analizu.

Podaci u spremištu podataka su spremljeni na način da su:

- organizirani prema području interesa
- integrirani
- nepromjenjivi
- sadrže vremensku dimenziju



Arhitektura spremišta podataka

Organizacija prema području interesa znači da su, primjerice, podaci proizvođača automobila fokusirani na dijelove, prodaju distributerima, rad dobavljača, sirovine, sastavnice i sl. Područja interesa osiguravajuće kuće mogu biti isplata šteta za police zdravstvenog, automobilskog ili životnog osiguranja.

Integriranost podataka je iznimno bitna kod izrade skladišta podataka. Podaci u spremište dolaze iz različitih izvora u različitim oblicima, koji su često neusporedivi. Neki od čestih primjera iz prakse su:

- Podaci o spolu osobe u jednom sustavu mogu biti zapisani kao riječ (muško, žensko), u drugom binarno (1,0) , a u trećem kao jedan znak (M, Ž), ili engleskom varijantom (M, F).
- Numerički podaci se mogu spremati u različitim mjernim jedinicama (m, cm, inč).
- Opisi mogu biti jednom sadržani u jednom polju, a u nekom drugom sustavu u više polja. U praksi je čest je slučaj da se, primjerice, ime artikla zapisuje u dva polja. U jednom polju se zapisuje skraćeni naziv, koji se pojavljuje na računu, a u drugom detaljniji opis ("Strukto 25", i "cement Dalmacijacement strukto 25kg vreća"). Maksimalna duljina teksta z jednom sustavu može biti npr. 35, a u drugom 100. Prilikom prebacivanja podataka u ovom slučaju može doći do gubitka bitnih informacija.
- Šifre mogu u jednom sustavu mogu biti u formatu Char, u drugom sustavu u formatu Integer.
- Datumi mogu biti u različitim formatima. Sve te podatke je prilikom prebacivanja u jedan sustav je potrebno ujednačeno formatirati.

Ako imamo *m* sustava iz kojih vadimo podatke u spremište, i ako prebacujemo vrijednost *n* polja, moramo definirati *m***n* pravila za konverziju podataka, i napraviti *m* potprograma za prebacivanje podataka.

Nepromjenjivost podataka znači da podaci koji su ušli u skladište podataka se ne smiju mijenjati. Podaci u skladištima podataka su namijenjeni čitanju, i moraju biti organizirani na takav način da ih nije potrebno mijenjati, već isključivo dodavati nove. Naime, ne smije se dogoditi da ako isti izvještaj pokrenemo više puta, dobijemo drugačije podatke.

Vremenska dimenzija. Sve poslovne transakcije (prodaja, kupnja, rezervacija, proizvodnja nečega) se događaju u jednom trenutku u vremenu, koji je potrebno zabilježiti. Različiti sustavi čuvaju transakcije u različitim vremenskim periodima. Primjerice, rezervacijski sustav aviokompanije čuva podatke od posljednjih 12 mjeseci, telekomunikacijski operater čuva podatke o pozivima 6 mjeseci. Banka čuva podatke o transakcijama 24 mjeseca. Stariji

podaci se brišu iz osnovnog sustava, kako mu ne bi usporavali rad, i arhiviraju se. Ako u analizama i izvještajima koristimo podatke iz dalje povijesti, spremiti ćemo njihovu "sliku" u skladište podataka. Tu "sliku" možemo napraviti po volji često. Primjerice, u skladište podataka ne moramo spremati podatke o svakoj avio-rezervaciji, već možemo spremiti pregled stanja svakog dana, ili svaki sat. Naravno, kako je stanje u jednom vremenskom periodu nepromjenjivo, ovdje se automatski poštuje prethodno pravilo o nepromjenjivosti podataka u skladištu podataka. Period čuvanja tako pripremljenih informacija u skladištu podataka može biti 10, 20 ili više godina.

Punjenje spremišta podataka

Podaci iz transakcijskih baza se u spremište prebacuju na slijedeće načine:

- **Totalno punjenje**, kada se u određenim vremenskim trenucima, spremište podataka isprazni, pa zatim ponovo napuni svim podacima iz transakcijskog IS.
- **Inkrementalno punjenje**, gdje se prilikom punjenja u spremište prenose samo novi podaci koji su nastali u transakcijskom IS nakon prethodnog punjenja.

Važno svojstvo podataka u spremištu je njihova **granularnost**. Govori o tome koliki nivo detalja spremamo u bazu. Primjerice, ako se prate posjete web stranicama, i ako svaki korisnički klik generira zapis u bazi, čuvanje tih podataka u spremištu podataka nema smisla, jer će spremište ubrzo toliko narasti, da će postati neupotrebljivo. Ovo je primjer niskog nivoa detalja.

Prilikom prikupljanja podataka iz puno različitih izvora imamo često suprotan slučaj, podaci su u previsokom nivou granularnosti, i potrebno ih je "razdvajati" na detalje. Idealno bi bilo da u spremištu čuvamo podatak o svakoj transakciji, jer bi to omogućavalo najtočnije analize i istraživanje podataka u neograničenu dubinu.

Složenost modela podataka i količina informacija koje su spremljena u bazu imaju direktan utjecaj na brzinu dohvata i spremanja podataka. Složeni upiti, koji pristupaju velikom broju međusobno povezanih tablica koje imaju indeksirane podatke, i međusobno su povezane relacijskim vezama zahtijevaju veliku procesorsku snagu. Stoga baza podataka sa prevelikom količinom informacija u sebi, i složenom relacijskom strukturom postaje neupotrebljiva za kompleksne analize podataka.

Tehnološki zahtjevi za spremište podataka

Sustav koji se sastoji od sustava za upravljanje bazama podataka (DBMS) i hardvera na kojem se aplikacije izvršavaju, mora zadovoljavati određene kriterije:

- Biti u stanju upravljati velikom količinom podataka
- Upravljati podacima na različitim medijima
- Jednostavno indeksirati i provjeravat konzistentnost podataka
- Biti spojiv sa širokom paletom database tehnologija (u svrhu što jednostavnijeg prikupljanja podataka)
- Dopustiti da se korekcije mogu vršiti programski direktno u bazu
- Paralelno pristupati podacima i snimati ih
- Biti u stanju brzo napraviti obavljanje (restore) podataka sa medija za pohranu.

Prilikom odabira tehnologije za izradu spremišta podataka moramo razlikovati DBMS sustave koji su prvenstveno namijenjeni obradi transakcija, od onih koji služe za rad sa skladištima podataka. Sustav namijenjen obradi transakcija je optimiran za brz upis transakcija, i izmjenu podataka. Drugi je optimiran za efikasno izvršavanje upita, te za efikasnu raspodjelu resursa pri izvršavanju većeg broja zahtjevnih upita.

U tablici je dana usporedba glavnih svojstava transakcijskog i analitičkog informacijskog sustava [W12].

Transakcijski IS	*Strateški IS*
Trenutni podaci stanja i prometa	*Povijesni podaci stanja i prometa*
Dinamički podaci	*Uglavnom statički podaci*
Ponavljajuća obrada	*Ad hoc obrada po zahtjevu*
Visok nivo transakcijske aktivnosti	*Nizak / srednji nivo transakcijske aktivnosti*
Predvidljiv način korištenja	*Nepredvidljiv način korištenja*
Transakcijski orijentiran	*Analitički orijentiran*
Aplikativno orijentiran	*Orijentiran prema području*
Podrška svakodnevnom odlučivanju	*Podrška strateškom odlučivanju*
Poslužuje veliki broj korisnika	*Poslužuje manji broj korisnika*
Sadrži detaljne podatke	*Sadrži grupirane (manje detaljne) podatke*

Sustavi za upravljanje višedimenzionalnim podacima

Uvod - Višekriterijska analiza

Pretpostavimo da smo vlasnik malog poduzeća, koje želi rasti na tržištu. Kako imamo ograničene resurse, moramo dobro razmisliti odlučiti gdje ulagati energiju, novac i vrijeme kako bi postigli što veći rast.

Tražimo od voditelja financija iznos prodaje za prethodnu godinu, i dobijemo broj 133000.

Taj broj sam za sebe govori puno, ali nas interesira kako se je kretala prodaja kroz vrijeme, pa tražimo analizu po mjesecima, i dobijemo:

- siječanj 25000
- veljača 32000
- ožujak 36000
- travanj 40000

Sada smo dobili bolju sliku o tome kako ide prodaja, i možemo već uočiti nekakav trend. Međutim, kako prodajemo više proizvoda, zanima nas koliko smo kojeg proizvoda prodavali koji mjesec:

	siječanj	veljača	ožujak	travanj
kava		4000	6000	8000
čokolada	10000	12000	13000	14000
sokovi	15000	16000	17000	18000

Sada smo dobili strukturu veliku 4 (mjeseca) * 3 (proizvoda) = 12 brojeva. To je za to jer rezultate promatramo preko dvije međusobno nezavisne dimenzije, od kojih jedna ima 3 a druga 4 vrijednosti

Dalje, zanima nas koliko smo prodali po regijama, kako bi odlučili gdje ćemo investirati u marketing.

Regija	Proizvod	Siječanj	Veljača	Ožujak	Travanj
sjever	kava		3000	4000	5000
	čokolada	7000	10000	10000	9000
	sokovi	10000	11000	12000	13000
Dalmacija	kava		1000	2000	3000
	čokolada	3000	2000	3000	5000
	sokovi	5000	5000	5000	5000

Ovaj izvještaj ima 4 (mjeseca) * 3 (proizvoda) * 2 (regije) = 24 broja. Sastoji se od 3 nezavisna popisa vrijednosti, i iako na prikazu izgleda kao dvodimenzionalna tablica, zapravo je trodimenzionalan.

Ako nas sada zanima koliko smo komada prodali po mjesecu, proizvodu i regiji, imati ćemo ukupno 48 podataka, a struktura će imati četiri dimenzije i sadržavati podatke po 4 liste vrijednosti koje su neovisne jedna o drugoj. Pojam "dimenzija" se u slučaju ovih struktura podataka koristi za označavanje "međusobno nezavisnih lista vrijednosti"

Iz ovakve tablice ne možemo na prvi pogled izvući neki zaključak, već je potreban alat koji će te podatke grupirati i prikazati na način da ih se lakše može međusobno uspoređivati.

Tu namjenu imaju sustavi za upravljanje više-dimenzijskim bazama podataka.

Data Mart

Sustavi za upravljanje više-dimenzijskim podacima (Data Marts, OLAP sustavi) strukturiraju informacije na način da ih se može pregledavati iz različitih perspektiva, te dinamički istraživati međuovisnost različitih varijabli, kao i međuovisnost sumarnih podataka i detalja. Kvalitetno oblikovani podaci u skladištu podataka služe kao dobar, brz, konzistentan i pouzdan izvod podatka za više-dimenzijske DBMS. Problem nastaje ako imamo više različitih izvora podataka, i više OLAP kocki koje treba puniti. Tada broj veza i sučelja za prijenos podataka raste eksponencijalno, i teško ga je držati pod kontrolom.

Više-dimenzijski DBMS sustavi mogu imati različite strukture baze podataka. Može se koristiti relacijski model, kao i kod skladišta podataka i transakcijskog sustava, a moguće je podatke spremati u tzv. "OLAP kocke", (engl. on-line analytical processing) model koji je optimiran za pretraživanje kroz različite dimenzije.

OLAP kocka je struktura podataka organizirana tako da ju je moguće brzo pretraživati i analizirati kroz više dimenzija. Transakcijski sustavi, a i spremišta podataka nisu pogodni za brz pristup velikoj količini podataka kroz različite dimenzije, već su pogodnije za obradu pojedinačnih transakcija. Iako je većina poslovnih informacijskih sustava je zasnovana na relacijskim bazama podataka, jer im je primarni posao obrada transakcija, izrada sumarnih izvještaja iz ovakvih baza podataka, kada treba pročitati zapise u gotovo cijeloj bazi, dugo traje. Problem je veći ako treba pokrenuti više različitih analiza, gdje je iste podatke potrebno više puta pročitati, uz drugačije grupiranje.

OLAP sustav se sastoji od OLAP baze, (engine-a) i OLAP sučelja. Glavni proizvođači OLAP sustava su danas Cognos, Business Objects, MicroStrategy. Microsoft SQL Server ima aplikaciju Analysis Services, koja služi kao OLAP Engine, a Excel se koristi kao sučelje za pregled podataka.

OLAP kocku možemo promatrati kao višedimenzionalnu Excel tablicu, gdje možemo imati proizvoljan broj dimenzija. OLAP kocka se "puni" periodički, kada se u nju upisuju podaci iz spremišta podataka, ili direktno iz transakcijske baze. Pri svakom "punjenju" prethodni sadržaj OLAP kocke se briše, i upisuje se novi. Ove operacije se obično rade u vrijeme kada je manje opterećenje na bazu, tijekom noći. Podaci u OLAP kocki "kasne" za realnim

podacima neko kratko vrijeme (npr. 1 dan), ali to ne predstavlja problem, jer služe za analize koje se rade kroz dulje vrijeme, i za uočavanje trendova za koje dnevni podaci ne predstavljaju bitnu veličinu.

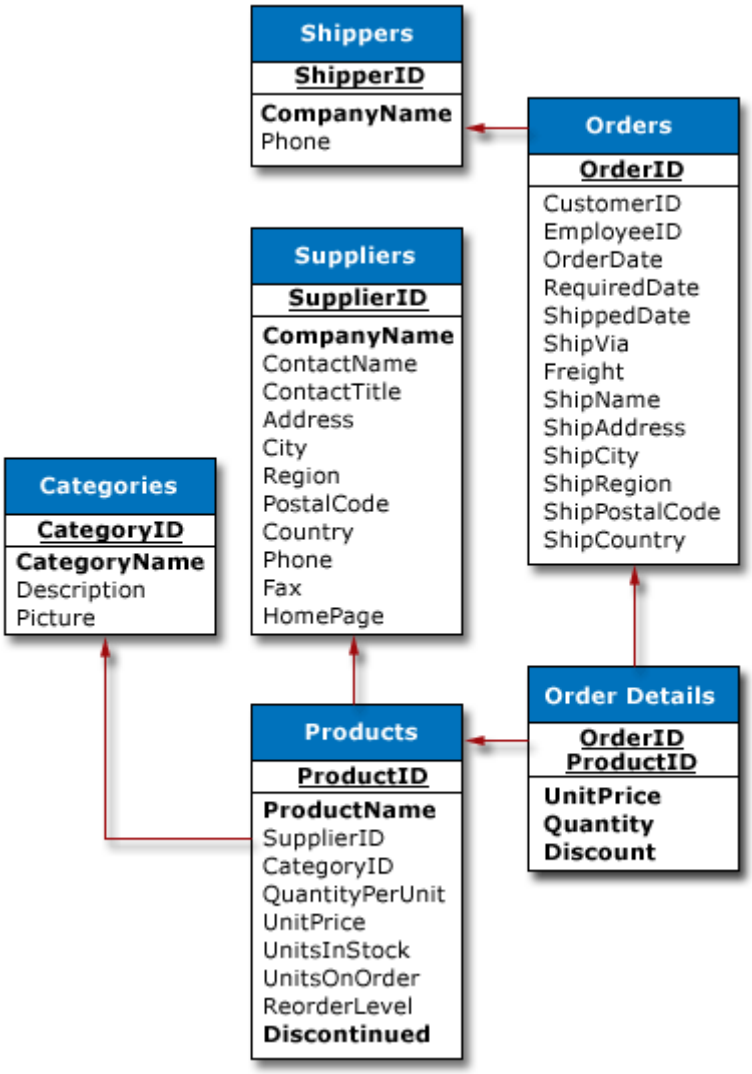
Preporučljivo je da se OLAP kocke pune iz spremišta podataka, međutim ukoliko transakcijski sustav radi na jedinstvenoj platformi, i ukoliko ima na raspolaganju sve potrebne podatke, moguće ih je puniti direktno.

Primjer korištenja OLAP kocke je analiza podataka o prodaji po regijama, gradovima, familijama proizvoda i sl. Različiti odjeli unutar poduzeća mogu imati različitu strukturu svojih OLAP kocki, ovisno o tome na koji način analiziraju podatke. Iste "sirove" podatke, koji se odnose na prodaju, recimo odjel marketinga može analizirati prema dobi kupaca i godišnjem dobu.

Osnovna mjera kod OLAP kocke je broj njenih dimenzija, i o njima direktno ovisi i njena veličina. Pojam "kocka" se odnosi na trodimenzionalni objekt, ali u OLAP terminologiji može označavati objekt sa bilo koliko dimenzija. Kako računamo volumen kocke? Pomnožimo sve 3 dimenzije. Volumen n-dimenzionalne "kocke" se računa na isti način.

Struktura podataka u transakcijskom informacijskom sustavu

Prvobitna namjena transakcijskog sustava je obrada poslovnih transakcija. Njihova struktura podataka je takva da uz što manje redundancije omogući što očitiji i jednostavniji unos transakcija kako nastaju. Primjerice, ako se u poslovanju događaju transakcije "Unos novog računa", logično je da u sustavu postoji tablica "Račun" u sa onim poljima koja unosimo u sustav kada unosimo novi račun. Bazu podataka bi se moglo organizirati i drukčije, pa podatke o jednom računu unositi u više tablica, što bi dovelo do redundancije i manje efikasnog korištenja sustava. Ovako modelirana baza podataka lako i brzo izvuče podatke za izvještaj koji se zasnivaju na pretraživanje po glavnim tablicama, npr. "Popis računa po kupcu", "prodaja po kupcu".

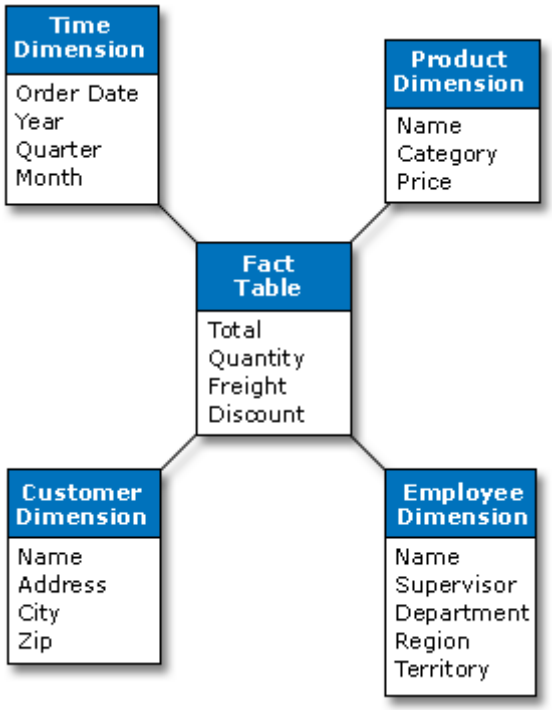


Model podataka transakcijskog sustava.

Podaci se u transakcijskim sustavima ne čuvaju predugo, kako ne bi došlo do usporenja rada zbog prevelike baze podataka. Ovisno o veličini poduzeća i kapacitetu DBMS-a, podaci se obično čuvaju 1-2 godine. Stariji podaci se arhiviraju, ili premještaju u spremišta podataka.

Struktura podataka u OLAP sustavu

OLAP ima drukčiju namjenu od transakcijskog. OLAP služi za analizu podataka kroz povijest, i to na različite načine, i kroz različite dimenzije. Zbog toga je primjerena i drukčija struktura koja čuva te podatke. Najčešće korištena struktura je "zvjezdasta shema". [W13]



OLAP - Zvjezdasti model podataka

Ostale OLAP strukture koje se koriste su:

Pahuljasta - (engl. snowflake) i zvjezdasto-pahuljasta (engl. starflake)

Ako tablice dimenzija (u gornjem primjeru Employee, Customer, Time, Product) imaju dodatna grupiranja. (Grad ima županiju, postoji šifrant kategorije proizvoda i sl.) Ako su te nove tablice (tablice Informacija) u samo jednom dodatnom nivou onda se radi o pahuljastoj, a ako su u više nivoa (tablica Županija ima vezu prema šifrantu Države), onda se radi o "starflake" strukturi.

Konstelacija – (engl. constellation), kada postoje više glavnih tablica (engl. fact table), koje koriste neke zajedničke dimenzije.

Glavni elementi OLAP strukture su:

- Glavna tablica - sadrži glavne podatke nad kojima se vrši analiza
- Dimenzijska tablica (Dimension table)
- Tablica informacija (Information table)

Koristeći ovaj model podataka lako možemo odgovoriti na pitanje "U kojoj regiji se daje najviše popusta", ili ukupna isporučena količina po danu, mjesecu i kvartalu, "U kojem gradu se prodaju najskuplji proizvodi".

Formiranje OLAP kocke

Izgradnja modela OLAP kocke se sastoji definiranja dimenzija i njihovih hijerarhija i nivoa. Dimenzija je, npr., vrijeme, nivoi su mjesec, tjedan, dan u tjednu. Jedna dimenzija može imati više hijerarhija. Dimenzija *vrijeme* može npr. imati hijerarhiju:

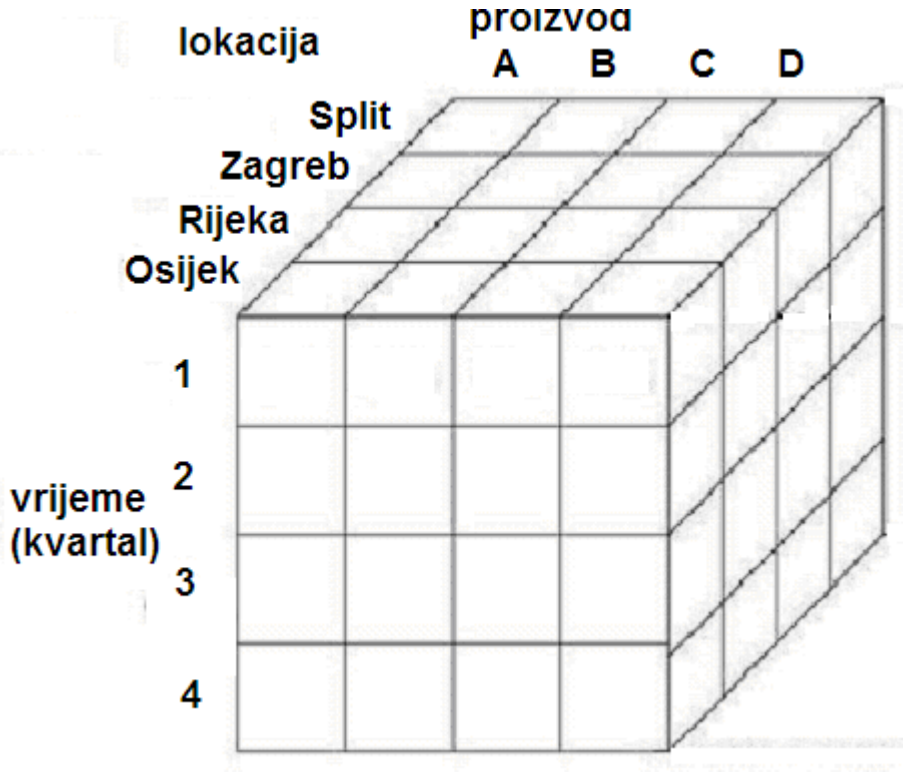
- Godina
- Kvartal
- Mjesec u kvartalu
- Tjedan u mjesecu-
- Dan u tjednu

U nekoj drugoj situaciji dimenziju *vrijeme* se može promatrati kroz hijerarhiju:

- Godina
- Mjesec
- Dan u mjesecu

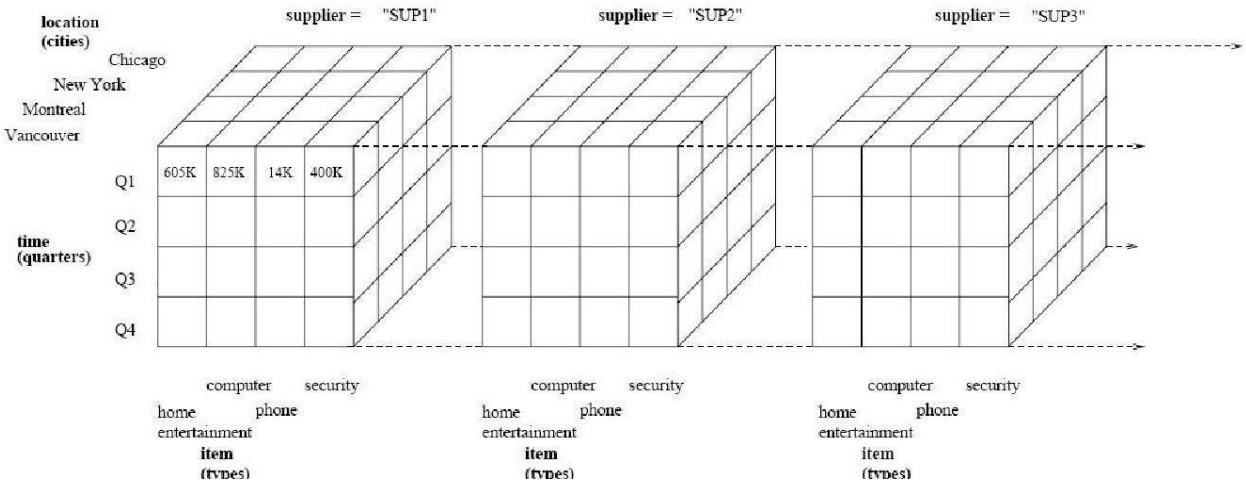
Strukturu OLAP kocke se gradi u ovisnosti o tome kakav tip odgovora se traži, tj. koje "dimenzije" se promatra. Fizička veličina strukture podataka je određena brojem dimenzija koji u OLAP kockama i brojem nivoa i hijerarhija u koje su podijeljeni podaci. Pritom treba razmotriti mogućnost kreiranja više kocki sa manjim brojem dimenzija, obzirom da veličina kocke raste proporcionalno umnošku veličine svih njenih dimenzija (geometrijskom progresijom).

Struktura podataka koja sadrži tri dimenzije se može zamisliti kao kocka, na čijem svakom bridu je jedna dimenzija podataka (na slici).



OLAP kocka sa 3 dimenzije

Struktura podataka sa četiri dimenzije se može prikazati kao niz 3D kocki, tako da se svaka kocka odnosi na jednu vrijednost četvrte dimenzije.



OLAP kocka sa 4 dimenzije

Operacije transformacije nad OLAP strukturom

Operacije transformacije podataka koje su moguće sa OLAP kockama:

- **SLICE** - izdvajanje podataka u grupi podataka za dati uvjet po jednoj dimenziji.
- **DICE** - izdvajanje podataka u grupi podataka za date uvjete po dvije ili više dimenzija.
- **PIVOT** - Operacija vizualizacije koja okreće dimenzijske osi radi alternativnog prikaza podataka.
- **ROLL UP** - Grupiranje podataka u OLAP kocke, bilo penjanjem po hijerarhiji dimenzije bilo izostavljanjem jedne dimenzije.
- **DRILL DOWN** (propadanje) - Detaljna raščlamba OLAP kocke, bilo spuštanjem po hijerarhiji dimenzije bilo uvođenjem jedne nove dimenzije. Obrnuto od ROLL UP.

Dubinska analiza podataka

Još od davnih vremena ljudi nastoje predvidjeti buduće događaje na temelju onih koji su se već dogodili. Ako neko znanje, koje smo stekli iskustvom ili opažanjem neke pojave, želimo podijeliti sa većim brojem ljudi, potrebno ga je na neki način zapisati. Da bi se znanje zapisalo, potreban je opće prihvaćen način zapisivanja, sustav simbola koji svi razumiju na isti način. Simboli, najprije grafički, koji su kasnije evoluirali u različita pisma, postoje od samih početaka civilizacije. Daljnjim razvojem društava razvijale su se i metode prikaza znanja, te su se za prikaz specifičnih znanja počeli primjenjivati "specijalizirani" skupovi simbola (npr. geometrijski simboli, simboli za prikaz matematičkih i kemijskih formula).

Do znanja se može doći na različite načine. Možemo zamijetiti neku pojavu u prirodi jednom ili više puta, i staviti ju u kontekst već poznatih činjenica. Primjerice, pola sata nakon što smo vidjeli udar munje, vidjeli smo i požar u šumi. Iz ovoga zaključujemo da munje mogu prouzrokovati šumske požare.

Do formula se je u početku došlo nekom vrstom opažanja ili istraživanja. Nakon što je eksperiment ponovljen više puta, primijećene su neke međuovisnosti i zabilježene. (npr. mjereno je vrijeme pada novčića sa različitih visina, te se je zaključilo da ovisi o kvadratu visine). Sada smo se dogovorili kojim simbolima ćemo prikazivati brzinu, visinu, i gravitacijsku silu i njihove međuodnose, pa možemo zapisati formulu.

Znanje prikazano na ovaj način je razumljivo svima na Zemlji, i može služiti za generiranje novog znanja (iz matematičkih i fizikalnih formula izvodimo nove). Formule, u najširem smislu, prikazuju odnose među poznatim vrijednostima (fizikalnim veličinama, matematičkim varijablama, ili npr. kemijskim elementima). Znanje prikazano formulama je najčešće egzaktno, jer uzima u obzir sve parametre o kojima ovisi krajnji rezultat (visina, gravitacijska konstanta).

Osnovna karakteristika ovako zabilježenog znanja jest da se može opet primijeniti, u datum okolnostima, uvijek će dati identičan rezultat (eksperiment sa novčićem se može ponoviti bilo kada i bilo gdje, kao i neka kemijska reakcija).

Često se međutim dogodi da nije lako doći do svih parametara o kojima ovisi ishod nekog eksperimenta, ili neka pojava. Npr. ako promatramo uspjeh studenata na studiju i uspoređujemo ga sa vremenom provedenim učeći u danu, na velikom broju studenata ćemo sigurno vidjeti neku povezanost, ali ju nećemo moći prikazati formulom tipa: *br_sati_učenja_dnevno * broj_dana * konstanta = postotak_uspješnosti_na_ispitu*.

Naime, o krajnjem uspjehu na ispitu ovisi velik broj parametara od kojih sve niti ne poznajemo, pa ih ne možemo ni izmjeriti.

Ako izvršimo istraživanje na uzorku od, primjerice, 1000 studenata na način da doznamo koliko su sati učili dnevno, i koliko su bodova osvojili na ispitu, možemo izvesti slijedeći zaključak:

Ako kolegij X učiš 2 sata dnevno imaš 70% šanse za dobiti ocjenu 4 ili 5.

Za ostalih 30% ne znamo koji parametri na njih utiču. Ovako formulirano znanje nije idealno, jer nam ne daje jednoznačno rješenje, ali je bolje od ne-imanja nikakvih informacija. Kako većina eksperimentalno dobivenih podataka ne daje 100% odgovor na neko pitanje, razvijena je teorija koja se bavi manipuliranjem ovako prikupljenih informacija.

Današnje društvo generira golemu količinu podataka o svim aktivnostima. Svaka kupovina karticom, telefonski poziv, spajanje na računalo i na web-stranicu ili on-line komunikacija generira informacije koje se negdje čuvaju. Ovisno o prirodi tih informacije, mogu se čuvati neselektivno i kraće vrijeme (među-spremnik na mrežnim uređajima), ili sistematski, na siguran način i u vremenskom roku koji je određen zakonom (podaci o bankovnim transakcijama i telefonskim pozivima).

Takvi podaci su spremljeni na različitim mjestima, i njihov kontekst se može odrediti sa različitom sigurnošću. Naime, za neke podatke se točno zna tko ih je generirao, a za druge se to može po potrebi doznati, uz više ili manje truda, i sa većom ili manjom točnošću.

Velik dio napora oko analize podataka se svodi na podatke iz različitih poslovnih sustava, gdje su spremljene sve transakcije koje neko poduzeće ima sa svojim kupcima i dobavljačima. Iz transakcijskih baza je lako izvući određene informacije, kao na primjer koji

je najbolji kupac ili koliko moramo uplatiti poreza u tekućem mjesecu. To je zato jer su "dimenzije" prema kojima te podatke tražimo već na taj način spremljene u sustavu. Podatak o porezu se upisuje u polje "porez", a podatak o prodaji u polje "prodaja", te ih je potrebno samo filtrirati i zbrojiti.

Međutim, u ovoj gomili "sirovih "podataka se krije puno veća količina znanja od one koja nam je očita. Primjerice, ako se podaci o prodaji u nekom vremenskom periodu povežu sa geografskom lokacijom kupaca, a ti podaci spoje sa prognozom ekonomskog rasta pojedinih županija koje redovito objavljuje Zavod za statistiku, može se doći do zaključka u koje regije je isplativo ulagati više sredstava u marketing. Prikupljanjem i analizom ovakvih podataka se bavi područje koje se naziva **Poslovna Inteligencija** (engl. Business Intelligence - BI)**.**

Motivatori za dubinsku analizu podataka

Jedan od najvećih korisnika Data Mining metoda su tvrtke koje se bave proizvodnjom ili prodajom robe ili usluga "široke potrošnje". U uvjetima sve veće globalne konkurencije i proizvođači i prodavači se bore za kupce. Jedan od načina na koji ih se želi pridobiti je postići da se "osjećaju kao kod kuće", tj. nastojati predvidjeti njihove želje i očekivanja vezano uz kupovinu, te im ciljano nuditi proizvode i usluge.

Naravno opće poznat faktor odabira je cijena, međutim na svakom tržištu se ubrzo postigne konsenzus oko toga da se ne isplati "ratovati" cijenama, jer to smanjuje zaradu svim konkurentima. Većina "ratova" cijenama kojima smo svjedoci u velikim trgovačkim lancima je obično ciljana na manji broj proizvoda, sa ciljem da se kupce navede da kupuju i druge, skuplje proizvode. Koje proizvode pritom ponuditi i po kojim cijenama, se nastoji zaključiti iz povijesnih podataka o prodaji.

Takav osjećaj kupca, "kao kod kuće", je uobičajen kod manjih "kvartovskih" dućana, gdje prodavač poznaje gotovo svakog kupca, te mu može ponuditi ono što ga zanima. Veliki trgovački lanci nedostatak osobnog poznanstva kompenziraju analitičkim praćenjem potrošnje, i svrstavanjem kupaca u različite "kategorije", prema kojima se kasnije radi ponuda.

Veliki trgovački lanci su dobar primjer za analizu poslovnih podataka u svrhu poboljšanja prodaje. Zbog same prirode poslovanja, rada sa velikim brojem kupaca, i velikog broja kupljenih proizvoda, raspolažu poslovnim informacijama koje su idealne za statističku analizu.

Naime, za uočavanje bilo kakvih trendova potreban je velik broj podataka za analizu. Primjerice, ako proizvodimo 2 broda godišnje, teško ćemo na temelju podataka iz prodaje uspjeti profilirati kupce.

Kod velikog broja transakcija mogu se prepoznati sličnosti po različitim kriterijima, npr. vrijeme kupovine, koliko komada je zajedno kupljeno. Ako smo u stanju i identificirati pojedinog kupca (npr. prema kartici vjernosti), na raspolaganju imamo i demografske podatke (dob, spol, eventualno obrazovanje, radno mjesto...).

Situacija sa proizvođačima robe široke potrošnje (FMCG - Fast Moving Consumer Goods) je još složenija, jer, osim što moraju konkurirati kod kupaca, moraju se izboriti i za mjesto u prodajnim lancima. U takvoj situaciji proizvođači često nemaju veliku moć pri postavljanju trendova, već im preostaje da pažljivo prate što se događa na tržištu i stalno prilagođavaju ponudu, mijenjanju i unaprjeđuju proizvode, te traže specijalizirana tržišta, tzv. "tržišne niše".

Jedna od strategija prodaje je i traženje **tržišnih niša**, tj. naći grupu kupaca koje zanima specifičan proizvod, ili posebna varijanta nekog proizvoda, kojeg nema puno na tržištu. (primjerice, skupi satovi, ili dizajnerski mobiteli). Kupci su takve proizvode voljni više platiti, što znači bolju zaradu za proizvođača. Obično je kod takvih proizvoda i manja konkurencija.

Na koju vrstu pitanja se traže odgovori rudarenjem podataka?

Rudarenje podataka, nalaženje skrivenog znanja iz velikih baza podataka, je tehnologija koja omogućuje uvid u velike količine poslovnih podataka koje svako veće poduzeće već posjeduje. Ovi alati primjenjuju jednu ili više metoda analize podataka na skup podataka koji želimo analizirati, sa ciljem otkrivanja skrivenih obrazaca, trendova i sl. Transakcijski

sustavi, a i alati za više-dimenzijsku analizu nude pretežno pogled u prošlost, dok data mining tehnike omogućuju predviđanje, tj. pogled u budućnost.

Nadalje, tradicionalni alati daju one informacije koje tražimo, (kakva je raspodjela troškova po mjesecima), a analiza pomoću naprednih statističkih metoda koje koriste alati za Data-Mining može otkriti povezanost između, primjerice, dobi prodavača i vrste proizvoda koji najbolje prodaje. Najčešće su svi podaci već tu, samo ih treba promatrati kroz "pravu prizmu".

Naravno, kod rudarenja podataka veliku ulogu igraju statističke metode, i procesorska snaga računala koja moraju pročitati goleme količine podataka. Statistika kao znanost postoji već odavno. Data mining koristi već poznate statističke metode, međutim u puno većem opsegu i na većoj količini podataka nego što je donedavno bilo moguće. Iako je za različite specifične primjene razvijeno na desetke različitih metoda, a čak i one najnovije postoje već barem desetak godina, "revolucija" u korištenju Data Mining-a je nastala otkada su računala postala dovoljno jaka da te metode primijene na ogromnom skupu podataka, a i kada su ti podaci postali dostupni, zbog korištenja spremišta podataka i više-dimenzijskih baza podataka.

Uvjet opstanka na tržištu je redovito praćenje svega što se zbiva u našem okruženju, što radi konkurencija, koliko smo i u čemu bolji od drugih, što moramo popravljati, itd. Neka od tipičnih pitanja na koja stalno tražimo odgovore u poslovanju su:

- Koji radnici imaju najbolji omjer rada i učinka?
- Koji profil kupaca ne kupuje naše proizvode?
- Kako prepoznati kupce koji više ne namjeravaju kupovati u trgovačkom lancu, i kako ih zadržati?
- Koji su najuspješniji prodavači?
- Koji naši korisnici će najvjerojatnije promijeniti operatera u slijedećih mjesec dana?
- Koja je najrizičnija skupina korisnika zdravstvenog osiguranja?
- Kako ide prodaja kave po vrsti, regiji i godišnjem dobu?
- Gdje najviše prodajemo kave?
- Koji proizvodi su manje **cjenovno osjetljivi** (neće im drastično pasti prodaja ako poskupe) ?

Koji sustav za upravljanje podacima daje odgovore na kakav tip pitanja?

Podaci iz tekućeg poslovanja se čuvaju u transakcijskom sustavu. Tamo su spremljeni podaci najnižeg nivoa, o svakom poslovnom događaju (prodaja, kupnja, skladištenje, kontakt...). Ovi sustavi su fokusirani na brzo i sigurno spremanje velikog broja transakcija, istovremeno posluživanje velikog broja korisnika, uz očuvanje integriteta podataka. Najčešće ne spremaju puno povijesnih podataka, a koriste se za tekuće izvještavanje, koje je manje zahtjevno što se tiče obrade podataka.

Spremišta podataka čuvaju povijesne podatke, koji mogu dolaziti iz više izvora i formatirani su na jedinstven način neovisno o tome otkuda dolaze i kada su spremljeni. Podaci u spremištu podatka su najčešće obrađeni na način da im je pojednostavljena struktura, uz određenu redundanciju i smanjenu količinu detalja u odnosu na relacijski sustav. Oni sadrže "vremenske snimke" - (engl. snapshot) poslovanja u nekom trenutku (npr. prodaja u dućanu u jednom danu).

Ovi sustavi će najbolje odgovoriti na pitanja koja zahtijevaju sumiranje i grupiranje podataka po postojećim kategorijama:

- Koji radnici imaju najbolji omjer rada i učinka?
- Gdje najviše prodajemo kave?

OLAP struktura sprema višedimenzionalne podatke, prema unaprijed određenim dimenzijama. OLAP struktura se izgrađuje svaki puta kada se pokreće analiza, te omogućuje da podatke promatramo svaki puta sa drugim dimenzijama - "iz druge perspektive". Pomoću OLAP sustava lako dobivamo odgovore na probleme koji uključuju analizu prema različitim kriterijima, naročito ako postojeće kategorije treba malo izmijeniti, ili prilagoditi (npr. vrijeme podijeliti jednom na semestre, a drugi put na dane u tjednu):

- Kako ide prodaja kave po vrsti, regiji i godišnjem dobu?
- Koji radnici imaju najbolji omjer rada i učinka?

Postoji čitav niz, obično kompleksnih i jako bitnih problema, na koji je potrebno promijeniti pristup pri traženju odgovora. Naime, ponekad pokušavamo vidjeti kakvi su parametri poslovanja međusobno povezani, a da ih nemamo unaprijed definirane. U tim slučajevima se koriste različite statističke metode i modeli kako bi se iz "sirovih", na prvi pogled nepovezanih podataka uočile neke zakonitosti, tj. pronašlo znanje. Pitanja na koja se može pokušati odgovoriti ovim metodama su:

- Koji profil kupaca ne kupuje naše proizvode?
- Koji naši korisnici će najvjerojatnije promijeniti operatera u slijedećih mjesec dana?
- Koja je najrizičnija skupina korisnika zdravstvenog osiguranja?
- Kako prepoznati kupce koji više ne namjeravaju kupovati u trgovačkom lancu, i kako ih zadržati?

Postoje pitanja kod kojih je potrebno najprije definirati varijable (dimenzije), pa tek onda provesti analizu po njima. Na takva pitanja se ne može jednostavno odgovoriti pomoću klasičnih sustava za upravljanje bazama podataka i klasičnih alata za izvještavanje.

Primjerice, da bi zaključili koji će nas kupci najvjerojatnije uskoro napustiti, možemo pokušati analizirati ponašanje kupaca koji su nas već napustili, te pronaći neke zajedničke karakteristike njihovog ponašanja. Analizom njihovog ponašanja prije odlaska može se npr. vidjeti da su u posljednjih mjesec dana više puta nazivali službu za korisnike, ili da su kupovali sve manje proizvoda, ili da su se rjeđe pojavljivali u dućanu.

Tek sada, kada smo definirali kriterije po kojima možemo tražiti kupce koji će možda otići (definirali smo dimenzije) možemo analizirati povijesne podatke i prepoznati takve kupce. Sada na prepoznatu skupinu kupaca možemo usmjeriti novu marketinšku kampanju. Primjerice, služba za korisnike može dobiti zadatak da takve kupce nazove i pokuša otkriti razlog njihovog nezadovoljstva.

Za razliku od višedimenzionalne analize pomoću OLAP kocki i izvještaja iz transakcijskih sustava, metode rudarenja podacima često ne daju egzaktne odgovore, već određenim mogućim ishodima (varijantama) pridružuju vjerojatnosti.

Data Mining je korištenje statističkih i programskih metoda za analizu podataka, gdje se postojeći povijesni podaci obrađuju pomoću različite statističkih modela.

Iako postoji velik broj metoda dubinske analize podataka, može ih se podijeliti u četiri osnovne grupe:

- *Klasifikacijske*, gdje se definiraju pravila za stavljanje podataka u pojedine grupe
- *Asocijacijske* metode traže veze (asocijacije) među raznim svojstvima.
- *Klasteriranje* je postupak grupiranja podataka koji su po nečemu slični.
- Numeričko predviđanje.

Rezultat takvih analiza su određene veze među podacima, prepoznavanje do sada neprimijećenih veza, prepoznavanje trendova i ponašanja, ili predviđanje budućnosti na temelju uočenih trendova.

Današnji poslovni sustavi najčešće već sadržavaju ogromnu količinu podataka iz svakodnevnog poslovanja. Svi podaci o narudžbama, ponudama, upitima, kupcima, dobavljačima, proizvodima, itd. su već negdje spremljeni. Međutim klasični informacijski sustavi ih mogu samo ograničeno obrađivati. Takvi sustavi pretežno pružaju uvid u događaje iz prošlosti i eventualno sadašnjosti. Odgovore na pitanja koja nas stvarno zanimaju, a to je kako se ponašati u budućnosti, ovakvi sustavi ne pružaju.

Nadalje, kada se provodi neko istraživanje ili anketa, barata se sa ograničenim skupom podataka, pa uvijek ostaje mjesta sumnji u to koliko su odabrani skup ispitanika reprezentativan te koliko prikupljeni podaci odražavaju stvarno stanje. Kod analize podataka iz transakcijskog sustava tih problema nema, jer se radi o potpunom skupu svih poslovnih događaja koji su se dogodili u nekom periodu.

Radi jasnoće, usporediti ćemo način rada OLAP-a i Data Mining-a, te kako se upotpunjuju.

OLAP ima unaprijed određene dimenzije, tj. unaprijed su poznati kriteriji prema kojima će se promatrati podaci. Pokretanjem OLAP analize, tj. kreiranjem OLAP kocke, podaci se mogu promatrati i uspoređivati kroz te dimenzije. (Npr. Koliko studenata je prošlo ispit iz prvog pokušaja, kolika je stopa odustajanja studenata nakon prve godine, itd.).

Ovakav pristup je dobar i daje rezultate kada je poznato kroz koje dimenzije se želi promatrati podatke.

Međutim, kako odgovoriti na pitanje: Koji studenti će vjerojatno napustiti studij ove godine?

Među postojećim podacima iz nastave (ocjene, dolasci na nastavu, broj polaganja...) može se pokušati zaključiti koje činjenice se mogu dovesti uvezu sa kasnijim napuštanjem studija.

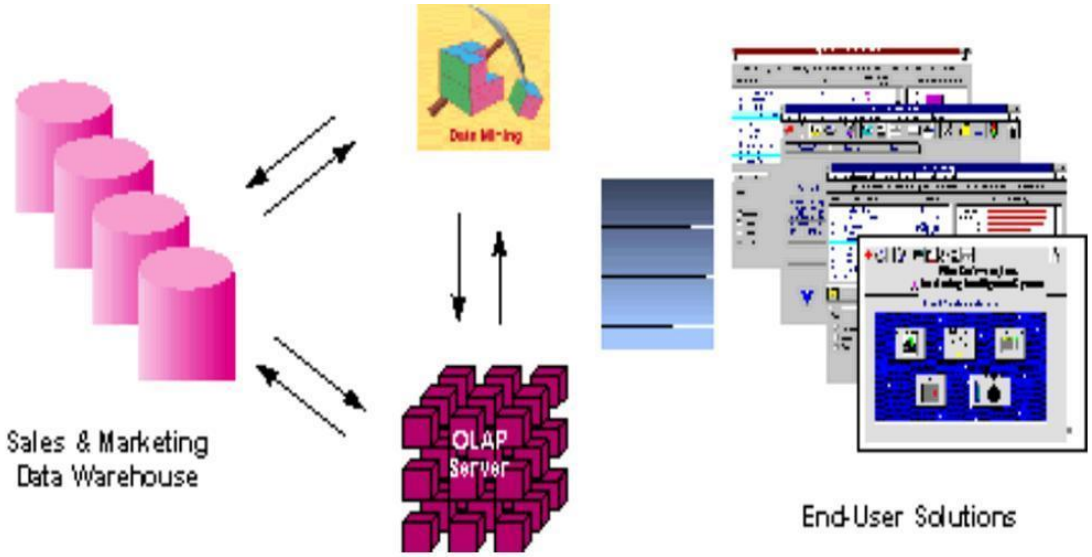
Nakon analize pomoću Data Mining-a može se npr. vidjeti da su studenti koji su prošle godine napustili studij, zadnjih nekoliko mjeseci često izostajali sa predavanja.

Sada se u OLAP-u definira dimenzija: Postotak pohađanja predavanja, pomoću koje se identificiraju studenti za koje postoji šansa da napuste studij u slijedećem semestru.

Arhitektura Data Mining-a

Glavni preduvjet za početak procesa rudarenja podataka je da postoje podaci za analizu.

Idealno je ako se ti podaci nalaze u spremištu podataka, unaprijed normalizirani i pripremljeni za analizu. Međutim, u praksi se često događa da se koriste podaci iz različitih izvora, (Data warehouse, OLAP, transakcijski sustav, odvojeni informacijski sustavi, rezultati analiza u tabličnom obliku, itd.), kao što se vidi na slici.



Arhitektura Data Mining - a

Pored same aplikacije za dubinsku analizu podataka, koja vrši obradu, postoji niz aplikacija za prezentiranje rezultata. Postupak Data Mining-a je često iterativan, tj. nakon prve obrade se dobiju određeni rezultati, na temelju njih se promijene parametri obrade, i dobiveni podaci se koriste kao ulazni ni podaci za neku drugu statističku metodu, pa se obrada pokreće ponovo.

Postupak rudarenja podataka

Iako točan redoslijed radnji ovisi o metodi (jednoj ili više) Data Mining-a koja se koristi, općeniti slijed koraka je:

- Pred-obrada - pripremanje skupa podataka
- Obrada podataka
- Interpretiranje rezultata
- Testiranje uzorka

Pred-obrada

Potrebno je odrediti skup podataka za analizu. Najčešće nije moguće analizirati sve podatke kojima raspolažemo, jer bi to zahtijevalo previše vremena. Tada se odabire dio podataka, za kojeg smatramo da dobro predstavlja sve podatke (da je reprezentativan). Pri ovom odabiru treba biti iznimno pažljiv, potrebno je biti siguran da odabrani skup podataka sadrži trendove koje analiziramo ili tražimo.

PRIMJER: Analiziramo prodaju različitih vrsta sladoleda po ljeti. Kako ukupan skup podataka iz transakcijskog sustava sadrži nekoliko desetaka milijuna stavki (podaci o svakom proizvedenom i isporučenom pakovanju sladoleda), odlučujemo se za dio podataka. Ako smo podatke tako odabrali da smo eliminirali čitavu jednu vrstu sladoleda, nećemo moći izvući dobar zaključak o tome koji se sladoled najbolje prodaje. Ili, ako smo eliminirali čitav jedan vremenski period, nećemo moći dobiti dobru vremensku analizu.

Pred- obrada (engl. pre-proccessing) je važan korak u dubinskoj analizi podataka.

Ako se provodi nad podacima iz spremišta podataka, velika je šansa da su podaci već kvalitetno pripremljeni, Međutim, i tada je moguć problem granularnosti, gdje su podaci iz različitih perioda spremljeni sa različitim nivoom detalja.

Kako se dubinska analiza često provodi nad podacima iz više izvora, moguće je da sve relacijske veze među podacima nisu uspostavljene, da u nekim skupovima podataka ne postoje sve vrijednosti. U tom slučaju analitičar odlučuje da li nedostajuće vrijednosti

umetnuti pomoću neke od statističkih metoda, ili ignorirati one podatke koji nemaju odgovarajuće vrijednosti.

U nekim bazama podataka, češće tamo gdje se podaci unose bez formalne provjere, se mogu nalaziti krivo uneseni podaci. Ako se radi o podacima koji imaju tipične vrijednosti, onda je lako uočiti one koje odstupaju, te odlučiti radi li se o greški ili ne. Naravno, na ovaj način se ne rješava problem vrijednosti koje jesu unutar uobičajenih, ali su krivo unesene.

Ako se koriste podaci iz različitih izvora, ili kroz dulji vremenski period, može se dogoditi da su spremljeni na različitim nivoima granularnosti. Na primjer, u jednom sustavu se sprema prodaja po dućanu i danu, a u drugom po dućanu u tjednu. Da bi te dvije vrijednosti bile direktno usporedive, potrebno ih je svesti na zajedničku "mjernu jedinicu". To možemo napraviti tako da tjednu vrijednost podijelimo sa 7, i dopunimo 7 dana u bazu, ili da dnevne vrijednosti iz druge baze zbrojimo i sve prikazujemo na tjednoj razini. U oba slučaja radimo kompromise: u prvom slučaju imamo netočne podatke o raspodjeli prodaje po danima u tjednu, a u drugom gubimo detalje o dnevnoj prodaji, a time i informaciju o tome u kojem dijelu tjedna prodaja ide bolje.

Općenito, problem sa različitim nivoima granularnosti, a i sa prevelikim brojem podataka može se umanjiti **kategoriziranjem**, gdje se određene grupe vrijednosti smještaju u jednu kategoriju. Tako se sve visine manje od 160 se smještaju u kategoriju "Nizak", ostale u kategoriju "Visok". Pri kategoriziranju je potrebno voditi računa o tome gube li se tim postupkom bitne informacije za analizu koja se provodi.

Potrebno je odrediti skup podataka za analizu. Ponekad je u transakcijskom sustavu previše podataka za analizu, pa je za neke korake potrebno smanjiti njihovu količinu. Najčešće nije moguće analizirati sve podatke kojima raspolažemo, jer bi to zahtijevalo previše vremena. Tada se odabire dio podataka, za kojeg se smatra da dobro predstavlja sve ostale podatke, da je **reprezentativan**. Pri ovom odabiru treba biti iznimno pažljiv, potrebno je biti siguran da odabrani skup podataka sadrži trendove koje analiziramo ili tražimo.

Primjerice, ako se analizira prodaja različitih vrsta sladoleda ljeti, a kako ukupan skup podataka iz transakcijskog sustava tvornice sadrži nekoliko desetaka milijuna stavki (podaci o svakom proizvedenom i isporučenom pakovanju sladoleda), odlučujemo se za dio

podataka. Ako se skup podataka smanji na način da se eliminiraju svi podaci o proizvodnji jedne vrste sladoleda, neće biti moguće izvući dobar zaključak o tome koji se sladoled najbolje prodaje. Ili, ako su eliminirani podaci o čitavom jednom vremenskom periodu, neće biti moguće uraditi kvalitetnu analizu trendova u vremenu.

Krajnji rezultat pred-obrade podataka je najčešće **jedna tablica sa podacima** koji se u daljnjim koracima analiziraju.

Obrada podataka

Obrada podataka se sastoji od odabira odgovarajuće metode, njenog parametriranja i pokretanja. Neke od najvažnijih metoda obrade podataka su:

- *Neuronske mreže* - Nelinearni modeli za predviđanje koji oponašaju prirodne neurone, i "uče na primjerima" - može ih se trenirati.
- *Stabla odluke* - Strukture u obliku stabla čije grane predstavljaju skupove odluka. "grananje" se vrši prema unaprijed određenim pravilima klasifikacije skupa podataka.
- *Genetički algoritmi* - Tehnike za optimiranje koje nad skupom podataka programski simuliraju evolucijske procese kao što su genetsko kombiniranje, mutiranje ili prirodni odabir.
- *Metoda najbližeg susjeda* - Klasificiranje podatka na temelju informacija o tome kako su klasificirani podaci najsličniji njemu.
- *Indukcija prema pravilima* - Definiranje pravila za klasifikaciju podataka, u obliku if-then sekvenci, i klasificiranje podataka prema njima. (ništa, malo, srednje, puno narudžbi)

Interpretiranje rezultata - testiranje uzoraka

Teško je pravilno odabrati uzorak podataka, jer se često unaprijed ne zna kakve rezultate je moguće očekivati.

Rezultati obrade podataka mogu biti više ili manje očiti, ovisno o kategorijama koje su odabrane za početak rada.

Često se obrada podataka ne radi na cijelom skupu podataka, jer bi predugo trajala. Umjesto toga, odabire se manji skup podataka te se na njemu vrši obrada. Pritom postoji rizik da rezultat koji smo dobili ne bude primjenjiv na cijeli skup podataka, jer nije odabran dovoljno reprezentativan uzorak. U nekim slučajevima je teško je pravilno odabrati uzorak podataka, jer se često unaprijed ne zna kakve rezultate je moguće očekivati, tj. među kakvim varijablama će se pojaviti povezanost.

Stoga je često nužno dobivene rezultate testirati u praksi. Za primjer uzmimo istraživanje o lojalnosti korisnika mobilne mreže, gdje je zaključeno da *"kupci mlađi od 25 godina iz Osijeka imaju najveću šansu promijeniti operatera"*.

Tezu se može testirati na slijedeći način: Formiraju se dvije grupe korisnika iz Osijeka mlađih od 25 godina. Prema jednoj grupi se usmjeri promotivnu akciju, npr. ponude se nagrade ako ne promijene operatera u slijedećih 6 mjeseci, i uspoređuje se njihov ostanak sa drugom, kontrolnom grupom. Ukoliko se pokaže velika razlika u broju odlazaka u dvije grupe, teza se može smatrati potvrđenom, i akcija se može primijeniti na cijeli skup korisnika.

Metode dubinske analize podataka

Osnovni pojmovi iz teorije vjerojatnosti i statističke analize

Za pravilno korištenje i odabir odgovarajućih metoda potrebno je poznavanje statistike i teorije vjerojatnosti.

Vjerojatnost:

Vjerojatnost odvijanja nekog događaja D u N ponavljanja nekog eksperimenta je

$$P_D = N_D * N$$

gdje su: P_D - vjerojatnost događaja D (0 - 1), N_D - broj odvijanja događaja D.

Na primjer, ako bacamo kocku 100 puta i pratimo pojavljivanje broja 6, za očekivati je da ćemo u otprilike ⅙ bacanja dobiti broj 6. Prema tome, vjerojatnost pojavljivanja broja 6 je ⅙.

Vjerojatnost da će se dogoditi dva uzastopna nezavisna događaja je umnožak njihovih pojedinačnih vjerojatnosti:

$$P_{AB} = P_A * P_B$$

Iz ovoga proizlazi da je vjerojatnost da će se dogoditi neki nezavisni događaj ako se je drugi nezavisni događaj već dogodio jednaka vjerojatnosti drugog događaja. Naime, ako su A i B nezavisni događaji, a A se je već dogodio, vjerojatnost njegovog događanja je 1, pa se može primijeniti formula za kombinaciju dvaju nezavisnih događaja.

PRIMJERI:

1. Kolika je vjerojatnost da ćemo iz 2 bacanja kocke dobiti brojeve 3 i 3?

$P_{33} = 1/6 * 1/6 = 1/36$

2. Kolika je vjerojatnost da ćemo iz 2 bacanja kocke dobiti brojeve 1 i 2, u bilo kojoj kombinaciji?

Imamo dva događaja koji zadovoljavaju traženi kriterij, kombinaciju 12 i 21. Vjerojatnost je, prema tome zbroj vjerojatnosti oba zadovoljavajuća događaja 2/36.

Varijable u statističkoj analizi

U statističkoj analizi varijablama nazivamo svojstva koja posjeduje predmet našeg ispitivanja. To mogu biti svojstva čije podatke imamo, ili svojstva čiju vrijednost tražimo. Primjerice, ako u nekom istraživanju ispitujemo utjecaj količine gnojiva na rast biljke onda su nam zanimljive varijable tjedna količina gnojiva, visina biljke ili količina sunca u danu.

Nezavisna varijabla je svojstvo kojim manipuliramo ili čije vrijednosti poznajemo u nekom istraživanju. U navedenom primjeru to je količina gnojiva koje se daje biljkama. Nezavisne varijable još se zovu i *prediktori,* ili *ulazne varijable*.

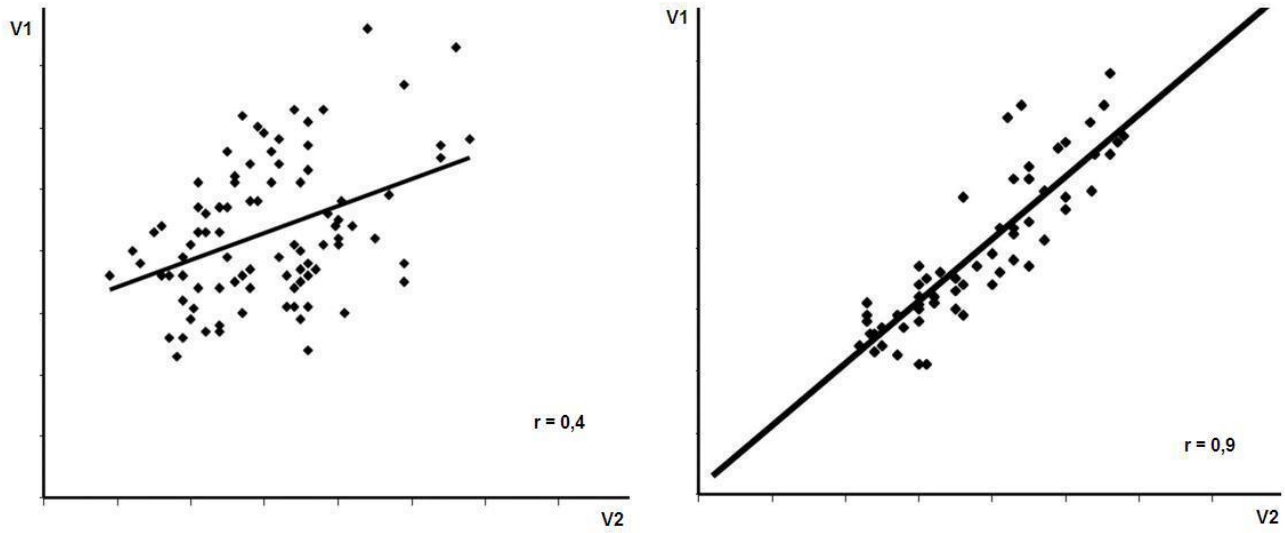
Zavisna varijabla (*izlazna varijabla)* je ona varijabla čiju se vrijednost mjeri, u ovom slučaju to mogu biti visina biljke i debljina stabljike. Cilj statističkog eksperimenta je dokazati utiču li promjene nezavisne varijable na zavisnu varijablu i u kojoj mjeri.

Kontrolirane varijable su one za koje se zna da mogu uticati na ishod eksperimenta, pa ih je potrebno držati pod kontrolom tijekom trajanja eksperimenta, kako bi se eliminirao njihov utjecaj na ishod eksperimenta. U ovom primjeru to su npr. količina sunca kojima su izložene biljke, temperatura zraka, i sl. Kontroliranim varijablama ne moramo znati vrijednost, jedino je potrebno osigurati da se sva ponavljanja eksperimenta odvijaju u istim uvjetima. U opisanom primjeru nije nužno mjeriti jakost sunca, ako su mu sve biljke nad kojima se provodi eksperiment izložene na isti način i dobivaju ga u istoj mjeri.

Statistički pokazatelji

Koeficijent korelacije

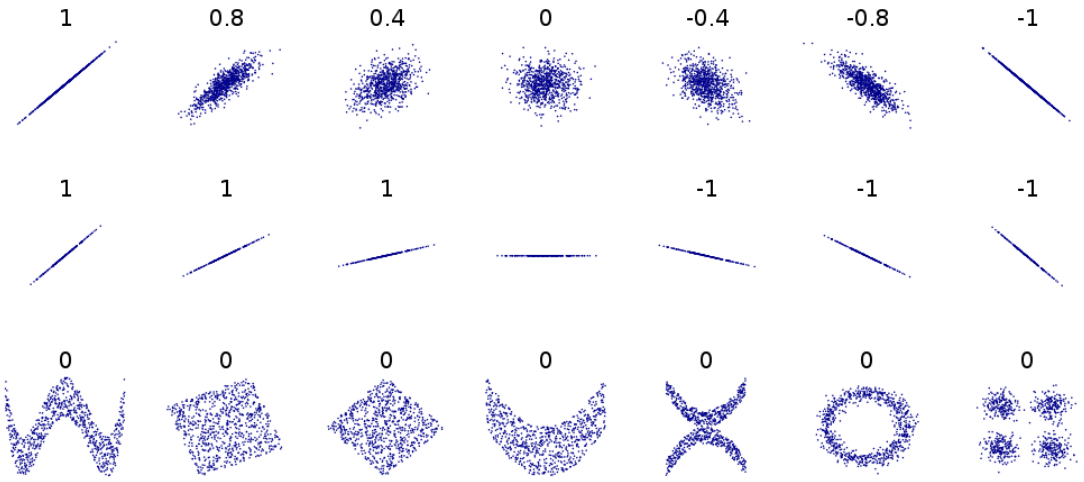
Koeficijent korelacije (Pearsonov koeficijent - *r*) je mjera linearne ovisnosti jedne statističke varijable o drugoj. Poprima vrijednosti u području [-1,1]. Kada je koeficijent korelacije između dvije varijable 1 ili -1 tada su one funkcionalno zavisne, tj. za svaku vrijednost jedne varijable druga varijabla može poprimiti samo jednu vrijednost. r = 1 znači da vrijednost jedne varijable potpuno ovisi o vrijednosti druge. Drugim riječima, ako znamo jednu varijablu, možemo sa 100% sigurnosti reći kolika je vrijednost druge. Za r = 0 ne postoji ovisnost među varijablama, tj. ako poznajemo vrijednost jedne varijable, to nam ništa ne govori o vrijednosti druge. U statističkim istraživanjima većina parova varijabli ima koeficijent korelacije veći od 0, ponekad čak i varijable koje smatramo nezavisnima.



Usporedba statističkih varijabli sa r=0,4 i r=0,9

Na slici se vidi da su vrijednosti varijabli čiji je r=0,9 manje "raštrkane" oko središta, nego u primjeru gdje r=0,4.

Na slijedećoj slici [W14] se vide primjeri različitih koeficijenata korelacije.



Grafički prikaz i dviju varijabli sa različitim koeficijentima korelacije

Regresijska analiza

Regresijska analiza je najstarija i najpoznatija statistička tehnika koja se koristi za analizu podataka. Regresija uzima skup numeričkih podataka i traži matematičku formulu koja opisuje njihovu međuovisnost.

Ograničenje ove tehnike je da dobro radi isključivo s kontinuiranim kvantitativnim podacima (kao što su težina, brzina ili starost), a za obradu podataka u kategorijama koji nisu u nizu (kao što je boja, ime ili spol) koriste se druge tehnike.

Pregled važnijih metoda dubinske analize podataka

Klasteriranje - grupiranje

Klasteriranje (eng. clustering) je metoda formiranja grupa podataka koji su po nekom kriteriju slični. Jedan od algoritama procjenjuje udaljenost među podacima, te ih prema unaprijed određenom kriteriju stavlja u grupe.

Algoritmi za grupiranje mogu biti hijerarhijski ili particionalni. Hijerarhijski algoritam u više iteracija formira stablo na čijem početku su svi elementi koje treba grupirati. U slijedećim iteracijama se formiraju grupe od više elemenata na temelju njihove međusobne udaljenosti, sve do tražene razine grupiranja.

Grupiranje se primjenjuje kada se u skupu podataka traže oni sa sličnim svojstvima.

Particionalno grupiranje (k-means)

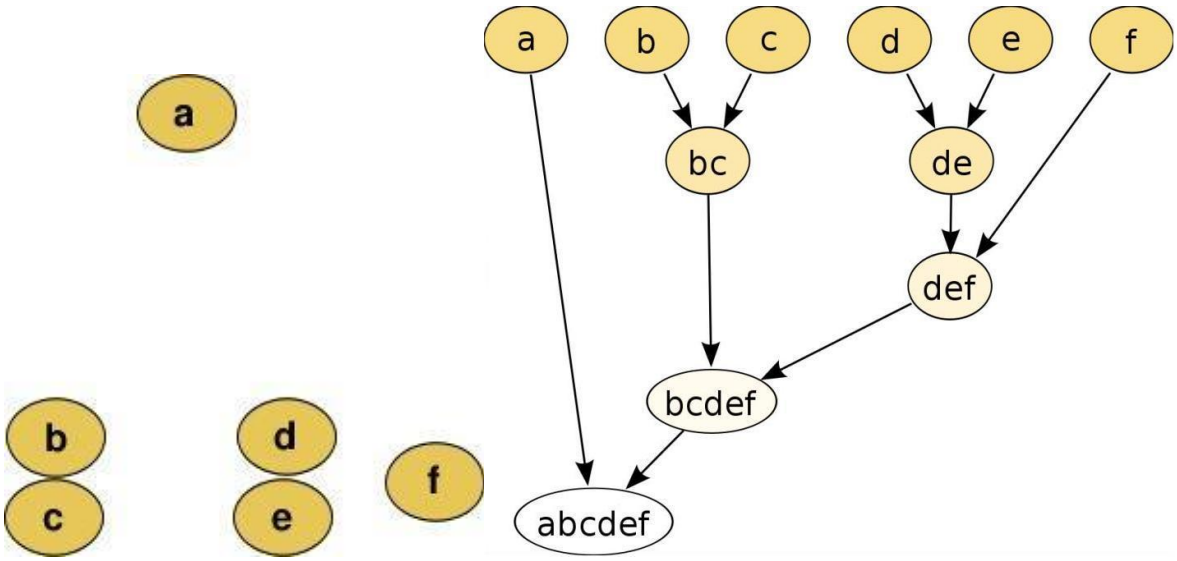
U ovom tipu grupiranja broj grupa se odredi unaprijed, ovisno o traženim rezultatima, ili koristeći neku od metoda za tu namjenu. Svakoj točki se odredi lokacija u n-dimenzionalnom prostoru rezultata. Te točke se nazivaju centroidi.

Sada se izračunava udaljenost svake točke-rezultata mjerenja od centroida, i stavlja ju se u onu grupu čiji joj je centroid najbliži. Ova metoda je pogodna za programsku primjenu jer se u iteracijama relativno jednostavno mijenjati koordinate centroida dok se ne dobiju optimalne grupe.

Kratak pregled ostalih metoda grupiranja može se naći u [W4]

Hijerarhijsko grupiranje

Pretpostavimo da imamo raspodjelu međuovisnosti dobi studenta i broja polaganja ispita, prikazanu na slici, gdje je na y-osi dob studenta, a na x- osi broj polaganja ispita:



Distribucija rezultata mjerenja

Nakon što se pomoću algoritma odrede udaljenosti među rezultatima, u prvoj iteraciji se mogu formirati grupe međusobno najbližih: a, (b,c), (d,e), f . U slijedećoj iteraciji se dobivene grupe mogu dalje grupirati, pa se dobiva: a, (bc), (def).

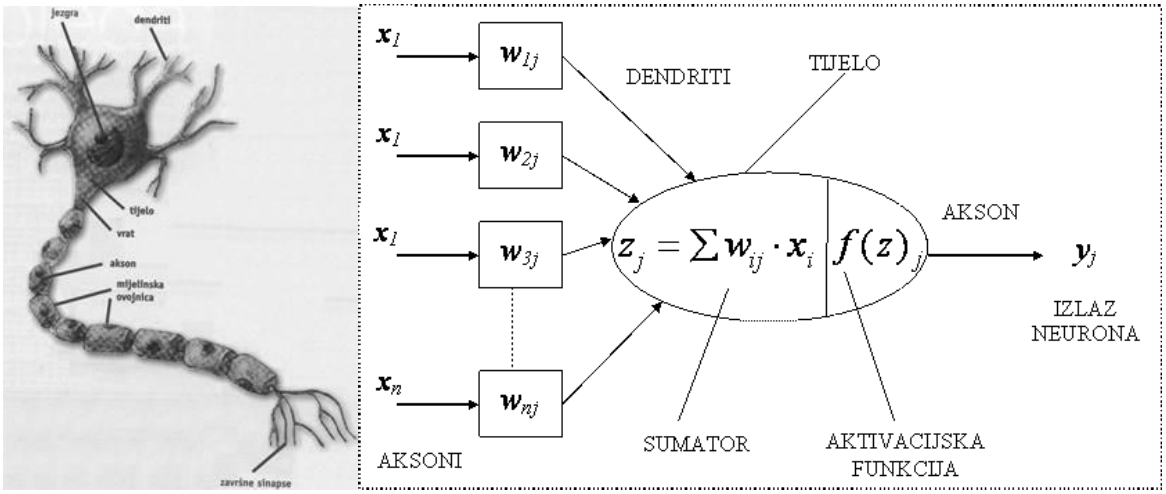
Način na koji se ova operacije može provoditi je da se formira *matrica udaljenosti*, sa i redova i j stupaca, čiji elementi predstavljaju udaljenosti između i-tog i j-tog elementa. U svakoj iteraciji grupiranja se spajaju redovi i stupci matrice. U ovako dobivenoj matrici elementi predstavljaju grupe koje su nastale spajanjem. Sada je potrebno u novoj matrici ponovo izračunati udaljenosti među elemenata (grupa elemenata). Udaljenost između dvije grupe, ili između grupe i elementa se može izračunati na više načina: kao najmanja udaljenost među dva elementa grupe, kao najveća ili kao srednja udaljenost između elemenata grupa. Ovaj postupak se može po volji ponavljati.

Priprema podataka za klasteriranje (grupiranje)

Kako se grupiranje bazira na izračunavanju udaljenosti među mjerenjima, podaci moraju biti u numeričkom obliku. Kako se može dogoditi da su mjere podataka u različitim skalama, potrebno je podatke normirati. Primjerice, ako je dob u rasponu 0-80, a zarada u rasponu 1.000 - 20.000, obje vrijednosti se mogu pomnožiti s konstantom tako da budu u rasponu 0-10. Sve nenumeričke varijable je potrebno prikazati numerički, npr. županije pomoću njihovih pozivnih brojeva, ili nazive regija pomoću šifri.

Neuronske mreže - algoritmi za učenje

Neuronske mreže se zasnivaju na načinu funkcioniranja živčanog sustava živih bića, koji se sastoji od neurona. Neuron je stanica koja reagira na električne podražaje, obrađuje ih i prosljeđuje drugim neuronima. Dendriti jednog neurona su povezani sa *sinapsama* drugog. Neuroni "uče" na način da, ako kroz neke dendrite signali ulaze češće i većeg su intenziteta, prolaz im se olakšava, i lakše se šalje izlazni signal kroz akson.



Biološki neuro i programska simulacija neurona

Umjetni neuron strukturom oponaša biološki, gdje aksone simulira skup ulaznih varijabli. Jezgru neurona koja obrađuje signale simulira funkcija koja svakom ulaznom signalu dodjeljuje određeni težinski faktor (w_{ij}), sve vrijednosti sumira, te ovisno o ukupnom iznosu (rezultatu aktivacijske funkcije) šalje ili ne šalje signal na izlaz.

U umjetnoj neuronskoj mreži neuroni (čvorovi) su organizirani u slojeve. U pravilu postoje jedan ulazni i jedan izlazni sloj, te jedan ili više skrivenih slojeva. Čvorovi ulaznog sloja primaju podatke, dijelom ih obrađuju, prosljeđuju skrivenim slojevima, koji ih nakon obrade prosljeđuju na izlaz.

Učenje se simulira na način da se podešavaju težinski faktori (w_{ij}), ovisno o snazi, učestalosti ili nekom drugom svojstvu ulaznog signala. veći težinski faktor znači da će taj signal više sudjelovati u sumi svih ulaznih signala, i lakše će aktivirati aktivacijsku funkciju.

Ideja učenja je slična kao i u prirodi; nakon što neuronska mreža nešto nauči, moći će to primjenjivati kada slijedeći put dođe u istu situaciju (primi iste ili slične informacije na ulazu).

Bayesove mreže

Bayesove mreže su model pomoću kojeg se procjenjuje vjerojatnost nekog događaja uz uvjet da su određeni uvjeti ispunjeni.

Ulazni podaci moraju biti u kategoričkim ili diskretnim varijablama. Varijable sa kontinuiranim vrijednostima potrebno je kategorizirati. primjerice, ukoliko imamo varijablu "vrijeme" koja može poprimiti neograničen broj vrijednosti, potrebno ju je prikazati npr. kao "datum" ili "mjesec", sa konačnim brojem vrijednosti. Također, varijable koje imaju velik broj diskretnih vrijednosti, a sve ne utiču na kvalitetu analize, mogu se kategorizirati. Tako se, primjerice varijabla "dob" može prikazati pomoću kategorija : < 18, 18-25, 25-35, 35, 45, 55 >.

Sastavni dio Bayesove mreže je i Tablica uvjetnih vjerojatnosti, koja sadrži vjerojatnosti da se neki elementarni događaj dogodi ako su se dogodili neki drugi događaji. Tu tablicu se može kreirati na temelju podataka iz informacijskog sustava, na temelju subjektivnih procjena ili rezultata istraživanja.

Prije početka obrade potrebno je definirati i uvjetovane događaje, (engl. *evidence),* tj. one događaje koji su se već dogodili.

Bayesove mreže mogu dati odgovor na slijedeća pitanja:

- Utiče li na iznos kupovine tehničke robe činjenica što kupac dolazi u dućan samo u jutarnjim satima, i što plaća gotovinom?
- Je li starija populacija sklonija kupovini mobitela na pretplatu od mlađe populacije?

Postupak rada sa Bayesovim mrežama u drugom primjeru je slijedeći: Najprije se kao uvjetovani događa (evidence) odredi da se radi o starijoj populaciji. Dakle promatraju se situacije kada su pripadnici starije populacije kupili mobitel na pretplatu, i kada nisu.

U drugom koraku se kao evidence definira mlađa populacija, i traže se iste situacije (kada su kupljeni mobiteli na pretplatu, a kada nisu).

Kao izlazni podatak se dobivaju vjerojatnosti kupovine mobitela na pretplatu kod starije i kod mlađe populacije.

Survival modeli

Survival modeli služe za analizu faktora koji utiču na duljinu trajanja nekog stanja, zapravo na njegov prekid ("smrt"). "Smrt", ovisno o području istraživanja može biti smrt pacijenta tijekom liječenja, kvar nekog uređaja, prekid pretplatničkog odnosa i sl. Ovo područje se još naziva i "modeliranja trajanja", ili "analiza trajanja".

U ovom modelu definira se ***funkcija preživljavanja***, kao vjerojatnost da će prekid (smrt, raskid ugovora i sl.) nastupiti iza vremena t, argumenta funkcije. Definira se i ***funkcija života***, kao komplementarna funkcije preživljavanja. ***Gustoća događaja*** je derivacija funkcije života, i govori koliko se događaja (prekida) događa u jedinici vremena.

Model preživljavanja se koristi za analizu prekida poslovnih odnosa, raskida ugovora, promjene operatera i sl.

Na izlazu model daje statističku mjeru povezanosti raznih prediktivnih varijabli, kao mogućih uzroka prekida poslovnog odnosa. Ta mjera povezanosti može biti npr. koeficijent korelacije. Na temelju dobivenih rezultata možemo iščitati da, npr. *najveću vjerojatnost odlaska imaju manji ljudi koji su tri puta nazivali pozivni centar u posljednja tri mjeseca*.

Priprema podataka za obradu u modelima preživljavanja

Ulazni podaci u model preživljavanja su kategorizirani podaci o ponašanju koje želimo analizirati, sa prediktivnim varijablama koje želimo istražiti.

Podaci moraju sadržavati varijablu koja označava status života (poslovni odnos je prekinut - 1, ili nije prekinut - 0) .

Asocijativni modeli

Asocijativni modeli odgovaraju na pitanje koliko se neki događaji pojavljuju zajedno, i koliko se pojavljuju u promatranom skupu događaja. Primjer asocijativne analize je tzv. "potrošaka košarica", gdje se istražuje sklonost potrošača da neke artikle kupuju zajedno, i u kojoj količini.

Primjerice, rezultat asocijativne analize ponašanja potrošača u nekom trgovačkom lancu može biti:

- 30% kupnji gdje je kupljeno pivo, kupljen je i čips
- 10% svih kupnji sadrži pivo i čips

Kod asocijativne analize najprije se definiraju ***pravila***, izjave čiju valjanost želimo izmjeriti. Pravilo bi u ovom slučaju bilo: *Kada kupac kupi pivo, kupi i čips.* Pravilo je prikazano u obliku "događaj A implicira događaj B". U našem primjeru A je kupovina piva, a B je kupovina čipsa.

Mjera pouzdanosti pravila je omjer pojavljivanja pravila (A implicira B) u ukupnom broju događaja A.

Mjera podrške pravila je omjer pojavljivanja pravila u ukupnom skupu podataka

Asocijativna pravila mogu biti iskazana i u negacijskoj formi. Tako se mogu istraživati situacije u kojima se uz kruh **ne** kupuje mlijeko.

Priprema podataka za obradu asocijativnim algoritmima

Kod obrade velike količine podataka važno je dobro definirati cilj istraživanja, smislena pravila, kako bi se optimiziralo vrijeme obrade. Naime kako se obrada vrši na velikom uzorku podataka, uspoređivanje sa velikim skupom pravila zahtijeva puno vremena. Nadalje, količina podataka se može smanjiti kategoriziranjem i grupiranjem. Tako se primjerice sve vrste mlijeka mogu promatrati kao jedna grupa proizvoda, što višestruko smanjuje skup ulaznih podataka, a možda ne utiče na rezultat analize.

Literatura

1 Panian, Ž., Klepac, G., Poslovna inteligencija, Masmedia, Zagreb, 2003.

2 Martin J.,Information Engineering, Book 1, Prentice Hall, NJ,

3 Inmon W. H., Building the data warehouse, Fourth Edition, Wiley Publishing, 2005

4 Jacobson R., Misner S., Microsoft SQL Server 2005 Analysis Services Step By Step, Microsoft Press, 2006

5 Witten I. H., Frank E., Data Mining Practice Machine Learning Tools and Techniques, Second Edition, Elsevier, 2005

6 Agnar Aamodt, Mads Nygård, Different roles and mutual dependencies of data, information,and knowledge - an AI perspective on their integration, Data and Knowledge Engineering, 1995

Web reference (datum zadnjeg pristupa: 20.05.2011)

1 http://www.scribd.com/doc/51726087/STATISTIKA-sredjeno-2

2 http://en.wikipedia.org/wiki/Dependent_and_independent_variables

3 http://www.l3a.com.hr/Poslovno-i-osobno.html

4 http://en.wikipedia.org/wiki/Cluster_analysis

5 http://hr.wikipedia.org/wiki/Grupiranje

6 http://en.wikipedia.org/wiki/Neuron

7 http://en.wikipedia.org/wiki/Survival_analysis

8 http://www.tsrb.hr/meha/index.php?option=com_content&task=view&id=14&Itemid=1

9 http://www.thearling.com/text/dmtechniques/dmtechniques.htm

10 http://www.thearling.com/text/dmwhite/dmwhite.htm

11 http://www.mayato.com/downloads/Summary_mayato_Data-Mining-Study_2009.pdf

12 http://rti.etf.bg.ac.rs/listarhiva/public/si3is1/2008/pdfT1Tj2c5060.pdf

13 http://www.dwreview.com/OLAP/Introduction_OLAP.html

14 http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

15 http://www.systems-thinking.org/dikw/dikw.htm

16 http://www.grad.hr/nastava/vis/vjerojatnost.pdf