© dr.sc. Goran Klepac
www.goranklepac.com
e-mail: goran**AT**goranklepac.com

Povezivanje na izvore podataka

Često puta, osim što metodama rudarenja podataka zahvaćamo veliku masu podataka, ti podaci mogu biti sadržani u različitim izvorima podataka. Osim što ponekad jedinstven izvor transakcijskih podataka spremljen u vidu relacijskog modela u nekoj od baza poput Oracle-a, DB2, MySQL-a, potrebni podaci za analizu mogu se istovremeno nalaziti u više spomenutih baza istovremeno.
Isto tako podaci ne moraju nužno biti spremljeni u bazi podataka, to mogu biti i ASCII datoteke, Excel datoteke, DBF datoteke i slično.
S obzirom na cilj analize, potrebno je konektirati se na sve relevantne izvore podataka, i izvući podatke relevantne za analitički proces.
Generalno gledajući pristup podacima bismo mogli podijeliti na nekoliko najznačajnijih situacija

- Postoji skladište podataka i svi relevantni atributi za analizu nalaze se u skladištu podataka
- Postoji skladište podataka i svi relevantni atributi za analizu se ne nalaze u skladištu podataka
- Postoji baza, ili niz bazi podataka kojima možemo pristupiti putem ODBC-a, ili putem direktne konekcije, a svi relevantni atributi analize nalaze se unutar baza podataka
- Postoji baza, ili niz bazi podataka kojima možemo pristupiti putem ODBC-a, ili putem direktne konekcije, a svi relevantni atributi analize se ne nalaze unutar baza podataka te je potrebno pristupiti eksternim izvorima podataka (skladište podataka, Excel tablice, ASCII datoteke, DBF datoteke)
- Kombinacije proizašle iz prethodno nabrojanih kategorija. Na primjer, transakcijski podaci se nalaze u MySQL bazi i DB2 bazi, podaci sa rezultatima telefonske ankete se nalaze u Excel tablici, sa postojanjem jasnog identifikacijskog ključa koji je veza sa bazama podataka, podaci o tečaju na današnji dan nalaze se u ASCII formatu.

Idealna je situacija kada postoji skladište podataka, te se u njemu nalaze svi relevantni atributi koje namjeravamo koristiti u analizama.
U praksi to najčešće nije slučaj, već smo prisiljeni integrirati podatke iz različitih izvora s ciljem dobivanja pretprocesirane tablice, na koju možemo primijeniti metode rudarenja podataka.

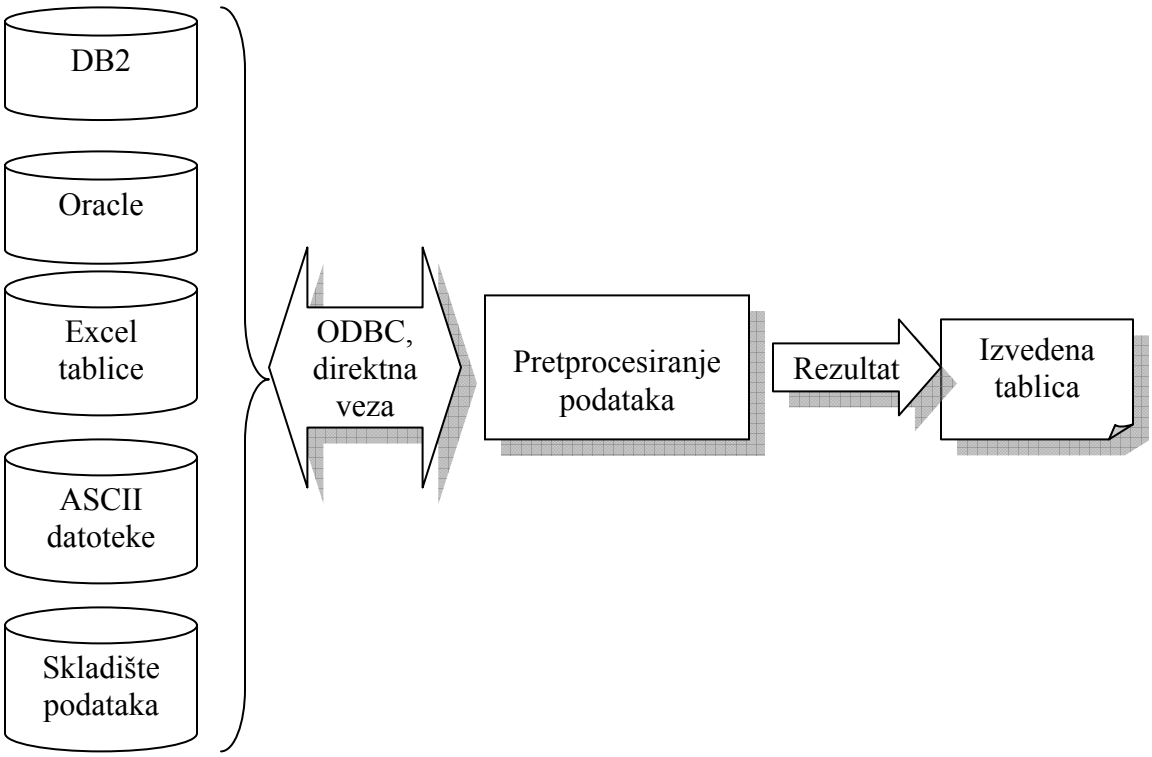
Osoba koja se bavi rudarenjem podataka, osim što mora biti vješta u različitim disciplinama koje imaju upliv na ovo područje, te mora poznavati strukturu algoritama za rudarenje podataka i biti upoznata sa komercijalnim alatima za rudarenje podataka i SQL-om, mora poznavati barem jedan programski jezik.
Poznavanje programskog jezika, omogućava kreiranje objekata za konekciju na različite izvore podataka, bez obzira na vrstu veze (ODBC, direktna veza, datoteka određenog formata). Isto tako poznavanje programskih jezika omogućava programiranje novih analitičkih metoda za rudarenje podataka, ako i “mutiranje” postojećih algoritama. Uz poznavanje programskog jezika analitičar koji se bavi rudarenjem podataka mora

poznavati osnovne karakteristike baza podataka, relacijske modele, formiranje “storage” procedura i mogućnosti manipulacije bazama podataka.
Iako komercijalni alati za rudarenje podataka u sebi sadrže module za direktnu konekciju na bazu, praksa pokazuje da se kod takvih pristupa podacima često puta pojavljuju određene poteškoće prilikom transfera.
Drugi glavni argument u korist poznavanje programskih jezika i bazi podataka, leži u činjenici da se dobar dio pretprocesiranja može obaviti direktno na izvornoj bazi, što može rezultirati “povlačenjem ” iz baze čiste pretprocesirane tablice koja je u potpunosti pripremljena za analizu.
Zamislimo situaciju u kojoj želimo analizirati kvartalnu prodaju u nekom maloprodajnom lancu prema grupama artikala.
Cilj nam je otkriti na razini deskriptivne statistike da li postoji preferencija kupnje određene grupe proizvoda s obzirom na kvartal.
U bazi podataka, kao interesantan i iskoristiv podatak nalaze se transakcije na razini pojedinog računa u posljednjih 6 godina.
Ako pretpostavimo da maloprodajni lanac u prosjeku ima 10 000 računa dnevno sa prosječno 10 stavaka računa to je 100 000 slogova dnevno. Ako to pomnožimo aproksimativno sa 30 dana mjesečno i multipliciramo sa 6 godina dobijemo brojku od impresivnih 216 000 000 slogova.
Iako je ovo relativno mali broj slogova koji se inače pojavljuju kod rudarenja podataka, lako je za pretpostaviti kakve bi tehničke probleme prouzrokovalo povlačenje ove mase podataka preko ODBC-a.
Prihvatljivija strategija je formirati “storage” procedure za pretprocesiranje podataka, ili izvršiti pretprocesiranje podataka putem SQL upita vezanog uz konkretnu bazu na kojoj se nalaze podaci.
SQL upiti odnosno “storage” procedure za pretprocesiranje podataka u konkretnom slučaju imali bi funkcije :

- Dijagnostika nedostajućih vrijednosti
- Formiranje izvedenih atributa za grupe proizvoda (npr. mlijeko i mliječni proizvodi, pekarski proizvodi, bijela tehnika ...)
- Grupiranje (sažimanje podataka) na razini kvartala

Grupiranjem podataka na razini kvartala dobili bismo i izvedene vrijednosti atributa kao što je prosječna prodaja određene grupe proizvoda po kvartalima, ili pak ukupna prodaja određene grupe proizvoda po kvartalima.
Nakon pretprocesiranja podataka broj slogova za transfer može se izračunati po formuli 24 (broj kvartala* broj grupa proizvoda).
Ovako pretprocesiranu tablicu mnogo je lakše transferirati lokalno, kako bi bila spremna za rudarenje.
Postoje varijante alata za rudarenje podataka koje analiziraju podatke direktno na serveru, no to je mnogo skuplja tehnologija, a dodatno komplicira problem kada koristite niz različitih baza za rudarenje podataka.
Ovo je jedna od mogućih metodologija pretprocesiranja podataka i rudarenja podataka.
Naravno, varijanta lokalnog povlačenja podataka iz baze ovisi prvenstveno o masi

pretprocesiranih podataka. U slučaju nesažimanja velike količine pretprocesiranja podataka, postoji varijanta provođenja direktne analize na serveru.
Nakon što pristupimo podacima, i izvršimo pretprocesiranje podataka, krajnji cilj svakog od ovih postupaka, je dobivanje jedinstvene dvodimenzionalne tablice, koja u sebi sadrži atribute relevantne za analizu.
To možemo prikazati slikom 1.



Slika 1. Cilj pristupa podacima i pretprocesiranja podatake je dobivanje jedinstvene dvodimenzionalne tablice

Nikako se ne smije smetnuti s uma važnost analize relevantnosti atributa, jer u samom procesu analize koja se recimo bazira na pravilima možemo odvojiti atribute sa slabim utjecajima na ciljnu varijablu, što nam značajno može pomoći kod izbjegavanja kombinatorne eksplozije u metodama kao što su to npr. Bayesove mreže.

U slučaju postojanja skladišta podataka poslovi vezani uz rudarenje podataka mogu biti znatno olakšani, jer skladišta već sadrže pretprocesirane podatke kroz ETL procese. Usprkos tome praksa pokazuje da prilikom rudarenja podataka moramo imati pristup

izvornim podacima na kojima se temelje skladišta podataka iz nekoliko osnovnih razloga:

- Podaci u skladištu nisu na zadovoljavajućem stupnju granulacije
- Podaci koji nam trebaju u procesu analize nisu obuhvaćeni u skladištu podataka
- Podaci u skladištu su transformirani na neodgovarajući način za rudarenje podataka

Idealnom se situacijom sa perspektive rudarenja podataka može smatrati slučaj kada se svi relevantni podaci potrebni za analizu nalaze unutar skladišta podataka. Takvi primjeri su relativno rijetki, posebice ako se uzme u obzir da se u procesu rudarenja podataka uglavnom koristi čitava paleta eksternih podataka i pomoćnih baza izvan skladišta podataka.

Osim klasičnih, spomenutih izvora podataka, rudarenje podataka može se primjenjivati i na Web-u i tekstualnim podacima. Metodologija pristupa takvim podacima razlikuje se s obzirom na strukturu i smještaj ovakvog tipa podataka.
Kada želimo rudariti podatke proizašle s Web-a, tada možemo rudariti "Web log" datoteke, na osnovu kojih možemo provoditi segmentaciju model ponašanja i interesa posjetitelja Web-a, njihove preferencije, preferirano vrijeme dolaska, ili pak možemo raditi analize povezanosti i sličnosti različitih Web stranica.
Pristup ovim podacima bitno se razlikuje od klasičnog pristupa bazama podataka i datotekama u različitim formatima podataka.
Zbog specifične strukture podataka, potreban je i različiti pristup pretprocesiranja podataka, koji u konačnici formira tablice i strukture nad kojima se provodi rudarenje podataka.

Rudarenje teksta kao disciplina, također ima svoje specifičnosti kako i prilikom pretprocesiranja podataka, tako i kod postupaka same analize.
Najčešći zadaci rudarenja teksta svode se na pronalaženje sličnih tekstova, pri čemu tekstovi mogu biti u različitim odvojenim datotekama, nemali broj puta i u različitim formatima (.doc, .rtf , .txt, .ws). Pristup tekstualnim podacima mora riješiti sve ove probleme, kako bi se tekstovi u kasnijoj etapi uspješno analizirali.

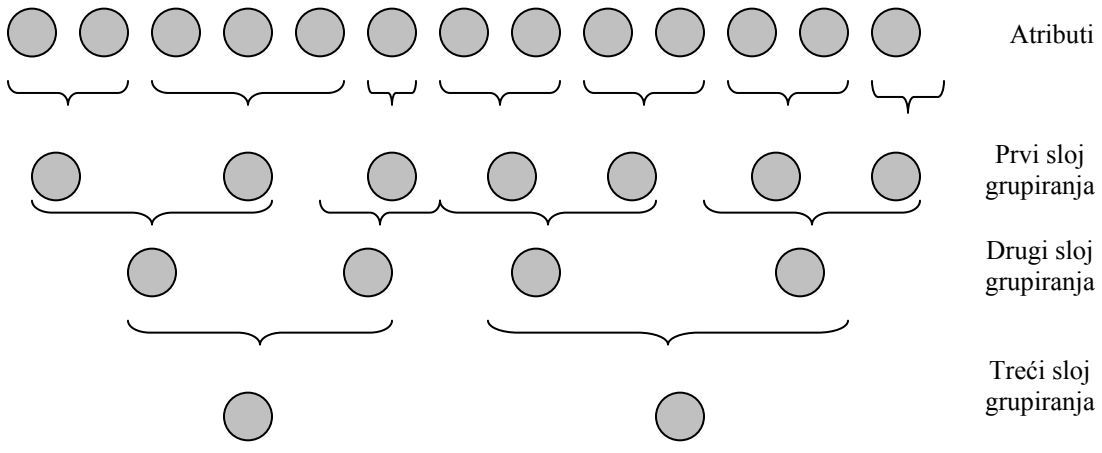
Problemi "kombinatorne eksplozije"

Prilikom čišćenja i pretprocesiranja podataka moramo voditi računa o problemima vezanu uz kombinatornu eksploziju koja može biti direktna posljedica primjene određenih algoritama rudarenja podataka.
Problemi "kombinatorne eksplozije" u uskoj su vezi sa pretprocesiranjem podataka. Naime, jedan od postupaka pretprocesiranja podataka je i postupak grupiranja. Ako primjerice grupiramo vrijednosti atributa po razredima, tada o broju razreda direktno ovisi broj mogućih kombinacija.

© dr.sc. Goran Klepac
www.goranklepac.com
e-mail: goran**AT**goranklepac.com

Ako formiramo premali broj razreda u samom procesu analize, postoji mogućnost gubitka dijela značajnih informacija, a kod prevelikog broja razreda aktualizira se problem “kombinatorne eksplozije”.
Ponekad je vrlo teško intuitivno odrediti raspone razreda, kada je riječ o specifičnim područjima koje zahtijevaju ekspertno znanje. Mnogo je lakše odrediti razrede za dobnu distribuciju, od primjerice koncentracija određenih tvari. Kod dobi možemo intuitivno odrediti dobne razrede, te imamo predodžbu o njihovom značenju, no laiku koncentracija ugljičnog monoksida u zraku ne znači previše bez dodatnog pojašnjenja eksperta.
Dakle, u pojedinim slučajevima prilikom pretprocesiranja podataka potrebna nam je pomoć eksperta iz određenog područja.
Komercijalni softverski proizvodi za rudarenje podataka s obzirom na konkretnu korištenu metodu prilikom obrade podataka nastoje posebno konstruiranim algoritmima reducirati vrijeme obrade i smanjiti utjecaj velikog broja kombinacija.
Kao primjer navodimo metodu potrošačke košarice, kod koje je posebno izražena problematika kombinatorne eksplozije.
U tradicionalnim verzijama algoritama potrošačke košarice “kombinatorna eksplozija” pokušava se reducirati uvođenjem prividnih varijabli, ili grupiranjem skupova proizvoda.
Na ovaj način djelomično se je rješavao problem “zagušenja” stroja prilikom obrade, ali je ostalo otvoreno pitanje gubitka informacija zbog reduciranja seta osnovne populacije.
Algoritam koji na efikasan način rješava problem “kombinantorne eksplozije” u metodi potrošačke košarice “stablo frekventnih uzoraka” (eng. frequent pattern tree) čuva temeljne informacije iz osnovne populacije podataka koja se analizira, a spomenute probleme rješava primjenama posebno konstruiranih algoritama, koji se baziraju na proračunima vjerojatnosti, kako a priori, tako i a posteriori vjerojatnosti.

Ovakav trend rješavanja problema ove prirode vidljiv je i u algoritmima za proračune uvjetne vjerojatnosti u tablicama uvjetnih vjerojatnosti kod konstrukcija Bayesovih mreža.



© dr.sc. Goran Klepac
www.goranklepac.com
e-mail: goran**AT**goranklepac.com

Slika 2 Grupiranje i kombinatorna eksplozija u pretprocesiranju podataka

Slika 2 prikazuje jednu od popularnih tehnika redukcije kombinatorne eksplozije grupiranjem varijabli.
Ako primjerice na prvi negrupirani sloj primijenimo primjerice algoritam potrošačke košarice, i to algoritam koji računa frekvencije pojavnosti metodom "svi na sve", vrlo je lako za uočiti da čak i na reduciranom setu podataka kao što prikazuje slika dolazi do velikog broja potencijalnih kombinacija. Postoje tehnike koje proračunavaju pojedinačnu frekvenciju pojavnosti pojedinog artikla kroz bazu, pa na osnovu zadanog parametra izuzimaju elemente koji ne prelaze prag.
Bez obzira na to u trgovačkim centrima koji imaju nekoliko desetaka tisuća artikala na policama, primjenom ovog principa još uvijek postoji veliki broj potencijalnih kandidata za analizu koji mogu doprinijeti "kombinatornoj eksploziji".
Jedno od popularnih rješenja je grupiranje artikala tako da recimo sve proizvode tipa mlijeko koji imaju različite udjele masnoće, grupiramo u kategoriju mlijeko.
Daljnji korak (drugi sloj grupiranja) može ići do razine grupiranje svih vrsta mlijeka bez obzira na proizvođača. Treći korak može grupirati sve mliječne proizvode svih proizvođača, pa bi tako mlijeko, kiselo vrhnje, šlag, jogurt, kiselo mlijeko i slični proizvodi ušli u istu kategoriju "mliječni proizvodi".
Na taj način smanjujemo kandidate kolji ulaze u analizu, ali gubimo dobar dio korisnih informacija.
Naravno, kategorije atributa i broj varijabli treba formirati na način da oni realno ocrtavaju problem, te da u skladu s tim na osnovu raspoloženih varijabli i njihovih kategorija možemo efikasno pronaći rješenje.
Ponekad je velik broj varijabli i velik broj kategorija unutar varijabli nužnost, što je uvaženo, te su trendovi premošćivanja ovih problema uglavnom usmjereni ka konstrukciji inteligentnijih algoritama koji se uspješno mogu nositi sa ovim vrlo izraženim problemom.

Umjesto trenda većeg sažimanja, prevladava trend razvoja efikasnijih algoritama koji svoje korijene vuku iz kombinatorike, teorije grafova, i specifičnih područja matematike i statistike.
Poneke metode su usprkos tome i dalje vrlo osjetljive na ovu problematiku, tako da ju je vrlo važno apsolvirati i uzeti u obzir prilikom razrade strategije analize i pretprocesiranja podataka.

Područja koja nam mogu biti od koristi prilikom procjene broja kombinacija, te planiranja upravljanja problemima ovakve vrste su kombinatorika i teorija vjerojatnosti.
Problematika kombinatorne eksplozije često puta se neopravdano zaobilazi kao predmet koji je zahtijeva pažljivije razmatranje. Ona je posebice nezaobilazna kada samostalno kreiramo vlastita algoritamska rješenja iz domene rudarenja podataka, a gdje se zahtijeva manipulacija sa kombinacijama atributa i njihovih vrijednosti.

© dr.sc. Goran Klepac
www.goranklepac.com
e-mail: goran**AT**goranklepac.com

Alati za rudarenje podataka koji su trenutno prisutni na tržištu relativno se uspješno nose sa ovom problematikom, posebice u domeni analize potrošačke košarice.