

Analiza podataka Data Mining

Andela Milićević, Luka Dekić

Pregled

- 1 Upoznavanje sa pojmom Data Mining
- 2 Faze u procesu Data Mining-a
- 3 Data Mining tehnike
- 4 Primena Data Mining-a

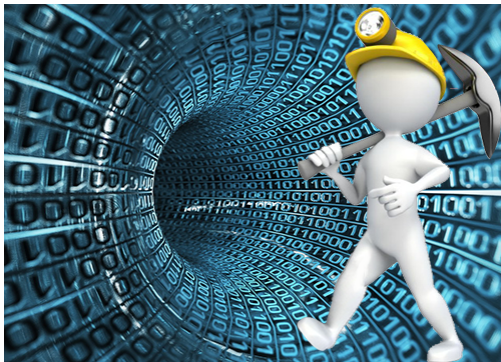
Šta je to Data Mining?

- Današnje uslove poslovanja karakterišu neizvesnost, rizik i konkurencija
- Preduzeća moraju svakodnevno da se bore za opstanak na tržištu i ostvarivanje boljih rezultata
- Skoro sve firme čuvaju velike količine podataka o poslovanju, klijentima i kretanjima u okruženju
- Dnevni unos podataka koje velike firme pohranjuju u svoje baze, meri se u terabajtima
- Kako znati baš *koji* podaci su značajni?

Šta je to Data Mining?

- Data Mining je proces pronalaženja skrivenih zakonitosti i veza među podacima
- Postupak izdvajanja interesantnih, novih i potencijalno korisnih informacija, sadržanih u velikim bazama podataka
- Multidisciplinarno područje koje obuhvata baze podataka, ekspertne sisteme, teoriju informacija, statistiku, matematiku, logiku i čitav niz drugih oblasti
- Data Mining se zove i Knowledge Discovery in Databases (KDD) tj. otkrivanje znanja u bazama podataka

Šta je to Data Mining?



Faze u procesu Data Mining-a

1. Sakupljanje podataka

- Poslovni podaci su uskladišteni u brojnim sistemima, na Internetu, u bazama podataka kompanija
- Prvi korak je prenos relevantnih podataka u bazu podataka gde se podaci analiziraju
- Poželjno je da se podaci sakupe iz različitih izvora, a nakon toga sempluju

Faze u procesu Data Mining-a

2. Filtiranje podataka i transformacija

- Cilj filtriranja podataka je odstranjivanje irelevantnih i suvišnih informacija iz skupa podataka
- To podrazumeva uklanjanje duplih i nepotpunih informacija, njihovu transformaciju, biranje podgrupa podataka itd.
- Neke od tehnika koje se primenjuju su transformacija tipova podataka, neprekidna transformacija kolona, rad sa vrednošću koja nedostaje, brisanje abnormalnih slučajeva

Faze u procesu Data Mining-a

3. Kreiranje i izbor modela

- Pre kreiranja modela treba da razumemo cilj datog Data Mining projekta
- Za svaki Data Mining problem postoji nekoliko odgovarajućih algoritama
- Preciznost algoritma zavisi od prirode podataka kao što su: broj stanja atributa koji se koriste za predviđanje, prenos vrednosti svakog atributa, veza između atributa itd.
- U ovom delu projekta potrebno je sastaviti tim poslovnih analitičara koji su eksperti u određenoj oblasti

Faze u procesu Data Mining-a

4. Procena kvaliteta modela

- U ovoj fazi kreira se skup modela korišćenjem algoritama i DM tehnika
- Nakon kreiranja mora se izvršiti i evaluacija tog modela
- Postoji nekoliko popularnih alata za evaluaciju kvaliteta modela, a najpoznatiji je lift dijagram

Faze u procesu Data Mining-a

5. Kreiranje izveštaja

- Nakon kreiranja modela i evaluacije kvaliteta tog modela, vrši se kreiranje izveštaja

6. Ocenjivanje modela

- U mnogim projektima, pronalaženje obrazaca i modela je samo pola posla
- Konačni cilj je upotreba tog modela za predviđanje

Faze u procesu Data Mining-a

7. Integracija Data Mining modela u aplikaciju

- Sve više poslovnih aplikacija uključuje Data Mining komponentu
- Na primer, on-line knjižara može dati potencijalnim kupcima preporuke knjiga
- Ovo je ključni korak za uvođenje Data Mining-a u masovnu upotrebu

Faze u procesu Data Mining-a

8. Upravljanje modelom

- Trajanje jednog Data Mining modela je ograničeno
- Nova verzija modela se mora praviti često
- Određivanje preciznosti i kreiranje novih verzija ovog modela bi trebalo biti postignuto korišćenjem automatizovanih procesa

Data Mining tehnike

- Iako je Data Mining mlada tehnologija, koriste se odavno poznate matematičke tehnike i algoritmi
- Ono što je povezalo ranija saznanja i velike baze podataka jeste pojeftinjenje prostora za skladištenje podataka i procesorske snage
- Data Mining tehnike se mogu podeliti u dve grupe
 1. Discovery Data Mining – tehnike za otkrivanje novih znanja
 2. Predictive Data Mining – tehnike za predviđanja

Data Mining tehnike

- Metoda najbližeg suseda
Nearest neighbor classification
 - Zbog načina rada, koji je sličan ljudskom načinu razmišljanja, ova metoda je jedna od najjednostavnijih
 - Zasniva se na traženju podataka koji imaju najbližija svojstva i poznato ponašanje
 - Podatak koji ima najbližija svojstva je najbliži sused, pa se pretpostavlja da će se slično i ponašati

Data Mining tehnike

- Stablo odlučivanja

Decision Tree

- To je tehnika odlučivanja utemeljena na odnosima između strategija i stanja
- Koristi se za rešavanje problema u finansijama, bankarstvu, osiguranju
- Korišćenjem serije pitanja i pravila za kategorizaciju podataka, predviđaju se ishodi. Stablo odlučivanja nastaje grananjem kao posledica ispunjenja uslova određenih pitanja. Svako pitanje će podeliti podatke u podskupove koji su homogeniji nego viši skup

Data Mining tehnike

- Neuronske mreže
Neural networks

- Ova tehnika Data Mining-a zamišljena je da deluje slično ljudskom mozgu
- Kao što ljudski mozak nakon učenja izvlači određene pretpostavke na osnovu ranijih zapažanja, tako i ove mreže predviđaju promene i dešavanja u sistemu nakon procesa učenja
- DM na osnovu ove tehnike počinje „učenjem” mreže pomoću podataka koji su već poznati, a koji se odnose na vrednost koju želimo prognozirati. Nakon toga znanje se proverava, sve dok rezultati provere ne budu zadovoljavajući
- Ovo je najkomplikovanija metoda, ali daje najtačnije modele

Data Mining tehnike

- Pored ovih tehnika postoji niz drugih algoritama na kojima se temelje modeli za Data Mining
- Neki od njih su fuzzy logika, memorijski zasnovano rasuđivanje, klastering, analiza potrošačke korpe

Primena Data Mining-a

- Data Mining najveću primenu ima u trgovini. Pored toga, primenjuje se i u bankarstvu. U medicini može da se odredi koju terapiju treba prepisati pacijentu. Elektrane ili telefonske kompanije mogu da predvide kada će i koliki biti vrhunac opterećenja, kako bi ga izbegle. . .
- Potencijali koji leže u Data Mining tehnologiji su ogromni i već se uveliko primenjuju, a dalji razvoj računarske tehnologije će ga sve više i više uvoditi u svakodnevni život

Literatura

- Max Bramer, British Library Cataloguing in Publication data, Springer-Verlag London limited, 2007
(<http://lib.mdp.ac.id/ebook/Karya%20Umum/Data-Mining-Undergraduate-Topics.pdf>)
- <http://www.sk.rs/2005/05/skpr01.html>

HVALA NA PAŽNJI!