INTEGRACIJA NESTRUKTURIRANIH PODATAKA U DATA WAREHOUSE

M. Srzentić, M. Arbanas

3.10.2010

SADRŽAJ

- Uvod
- Teoretska podloga
- Primjer iz prakse
- Zaključak
- Q&A

UVOD

- Velika količina podataka u nestrukturiranom obliku
- Različiti izvori podataka
  - e-mail, dokumenti, dopisi, ...
  - forumi, blogovi, Facebook, Twitter, Linkedin, ...
- Odgovori na posebnu vrstu pitanja
  - Kakvo je mišljenje o našim proizvodima i uslugama?
  - Koliko su zadovoljni naši klijenti?
  - ...
- Premala iskorištenost mogućnosti
  - nedostatak strukture
  - nedostatak kvalitetnih alata

TEORETSKA PODLOGA

TEORETSKA PODLOGA

EKSTRAKCIJA PODATAKA

- Različite metode ekstrakcije
  - export/import teksta
  - web crawleri
  - API web aplikacija
- Cilj je isti: dovesti sirove podatke u Data Staging Area

TEORETSKA PODLOGA

DOMENSKI POJMOVI

- Pojmovne kategorije
  - vremena/datuma
  - proizvoda i usluga
  - imena (osoba, org. dijelova)
  - geografskih lokacija
- Postoje specijalizirane taksonomije
  - preopćenite
  - preporučeno razvijati i koristiti svoje

TEORETSKA PODLOGA

DOMENSKI POJMOVI

•Primjer taksonomije proizvoda tvornice automobila:

1. Vozilo
    1.1. Osobno vozilo
        1.1.1.Automobil
            1.1.1.1. Model Mali
            1.1.1.2. Model Mali 2
            1.1.1.3. Model CC
            1.1.1.4. Model Sport
        1.1.2.Motocikl
            1.1.2.1. Model Ef 50
            1.1.2.2. Model Ef 125
            1.1.2.3. Model Ef S250
    1.2. Teretno vozilo
        1.2.1.Kamion
            1.2.1.1. Model 3T
            1.2.1.2. Model 7T
            1.2.1.3. Model 10T
        1.2.2.Kombi vozilo
            1.2.2.1. Model DG 2
            1.2.2.2. Model DGi

TEORETSKA PODLOGA

PARSIRANJE I TAGIRANJE

- Traže se pojmovi iz definiranih kategorija i tagiraju se
  - jednostavni (#) ili XML tag

"U **kolovozu** ove godine počinje prodaja novog modela **iPhonea** u našim trgovinama"

"U **<DATUM> kolovozu </DATUM>** ove godine počinje prodaja novog modela **<PROIZVOD> iPhonea </PROIZVOD>** u našim trgovinama"

- Temelj za narativni data mart
  - tagovi po kategorijama se pretvaraju u dimenzije

TEORETSKA PODLOGA

DOMENSKO ZNANJE

- Tezaurus – domensko znanje uređeno jednostavnim relacijama
- Ontologija – sličan koncept proširen atributima i većim brojem relacija
  - temelj semantičkog weba
  - OWL i RDF

<IVO IVIĆ – godina - 20>
<IVO IVIĆ – zanimanje - STUDENT>
<IVO IVIĆ – otac – PERO IVIĆ>
<IVO IVIĆ – majka – KATA IVIĆ>

- Preopširno definirane ontologije
  - potrebno naći optimalan broj tripleta

TEORETSKA PODLOGA

POVEZIVANJE POJMOVA

- Cilj: ujedinjavanje značenja i elimininiranje redundancije u pojmovima

<DOBAR – **sinonim** - ODLIČAN>
<DOBAR – **sinonim** - SUPER>
<DOBAR – **sinonim** - ZADOVOLJAN>
<DOBAR – **sinonim** - KORISTAN>
<LOŠ – **sinonim** - KATASTROFA>
<LOŠ – **sinonim** - NEPRIKLADAN>
<LOŠ – **sinonim** - NEUPOTREBLJIV>
<LOŠ – **sinonim** - PRESKUP>

- Moguće definirati i JAKO DOBAR, JAKO LOŠ i NEUTRALAN
  - povećava se mogućnost neispravnog tumačenja podataka

TEORETSKA PODLOGA

NARATIVNI I SINTETIZIRANI PODATKOVNI MODEL

- Narativni
  - dimenzije i fact tablice vrlo općenite, veze neobavezne
  - mora postojati referenca na originalne podatke
  - odgovara na veći broj pitanja
  - skloniji greškama
  - indikator povjerenja
- Sintetizirani
  - usko specijalizirana namjena
  - vrlo precizni podaci
  - odgovara na manji broj pitanja – vrlo fokusiran
- Obično jedan narativni i mnoštvo sintetiziranih data martova

PRIMJER IZ PRAKSE

- Mnoštvo primjera
  - obrada help-desk tiketa,
  - analiza odnosa s klijentima,
  - obrada prijavljenih šteta u osiguranju,
  - usporedba dijagnoza u medicini,
  - analiza interakcija lijekova u farmaciji,
  - rezultati marketinških kampanja...
- Cilj: Dobiti precizne, upotrebljive, općenite i/ili specifične informacije iz mnoštva sirovih podataka

PRIMJER IZ PRAKSE

- Marketinška kampanja izmišljene pivovare pri plasiranju novog proizvoda
  - Proizvođač: Pivovara Zlatica
  - Proizvod: Zlatno pivo
  - Izvor: Twitter
    - jednostavan izvor
    - olakšano pretraživanje
- Platforma
  - Oracle DB
  - Oracle Text
    - indeksiranje, pretraživanje i analiza teksta
    - ključne riječi, kontekst, teme, korijeni riječi...

PRIMJER IZ PRAKSE

DOHVAT KOMENTARA S TWITTERA

Pivoljubac	27.07.2010 12:46:17:	Izašlo novo pivo iz Zlatice, je li itko probao? #zlatnopivo
Majstor	27.07.2010 13:50:50:	Kod nas još nije stiglo u trgovine, ali jedva čekam... #zlatnopivo
Oskosk	27.07.2010 14:00:10:	Meni je pregorko, ne valja :(#zlatnopivo
Pro777	27.07.2010 14:07:34:	Jedno od najkvalitetnijih piva na tržištu, ali preskupo #zlatnopivo
Veseli	27.07.2010 14:31:10:	Meni je cijena OK, okus super, ja odsad pijem samo #zlatnopivo
Mrgud	27.07.2010 14:59:59:	Ambalaža totalni #fail, ostalo može proći

PRIMJER IZ PRAKSE

DEFINIRANJE POJMOVA O PROIZVODU I DOJMU

pivo, pivovara, zlatica, zlatno, hmelj, okus, cijena, ambalaža, skupo, jeftino, prihvatljivo, OK, dobro, loše, ukusno, najbolje, najgore, gorko, blago, fail, win, kvalitetno, super, katastrofa, ...

DOM_POJMOVI

ID	Number
POJAM	Varchar2(140)
KATEGORIJA	Varchar2(140)

- Dodatni pojmovi
  - sve izvedenice iz istog korijena
  - uobičajeni kolokvijalni nazivi (piva, bira, vopi, pivica, ...)
  - uobičajene pravopisne greške (cjena, supr, dorbo, ...)

PRIMJER IZ PRAKSE

PARSIRANJE I TAGIRANJE TWEETOVA

- Potrebno kreirati kontekstni indeks na koloni DSA_TWITTER_KOMENTARI.TWEET

```
CREATE INDEX dsa_twt_kom_twitter_idx
   ON DSA_TWITTER_KOMENTARI (tweet)
   INDEXTYPE IS CTXSYS.CONTEXT;
```

- Oracle Text automatski parsira i tokenizira tekst
  - tokeni spremljeni u jednu od indeksnih tablica (DR$<ime_indexa>$I)
- Tagiranje i povezivanje pomoću operatora CONTAINS

...

PRIMJER IZ PRAKSE

PARSIRANJE I TAGIRANJE TWEETOVA

```
SELECT t.id,
       t.tweet,
       p.pojam,
       p.kategorija
  FROM DSA_TWITTER_KOMENTARI t, DOM_POJMOVI p
 WHERE CONTAINS (t.tweet, p.pojam) > 0
```

ID	TWEET	POJAM	KATEGORIJA
2	Izašlo novo pivo iz Zlatice, je li itko probao? #zlatnopivo	pivo	PROIZVOD
4	Jedno od najkvalitetnijih piva na tržištu, ali preskupo #zlatnopivo	preskupo	DOJAM
4	Jedno od najkvalitetnijih piva na tržištu, ali preskupo #zlatnopivo	najkvalitetnijih	DOJAM
5	Meni je cijena OK, okus super, ja odsad pijem samo #zlatnopivo	okus	PROIZVOD
5	Meni je cijena OK, okus super, ja odsad pijem samo #zlatnopivo	cijena	PROIZVOD
5	Meni je cijena OK, okus super, ja odsad pijem samo #zlatnopivo	OK	DOJAM
5	Meni je cijena OK, okus super, ja odsad pijem samo #zlatnopivo	super	DOJAM
6	Ambalaža totalni #fail, ostalo može proći	ambalaža	PROIZVOD
6	Ambalaža totalni #fail, ostalo može proći	fail	DOJAM

PRIMJER IZ PRAKSE

DEFINIRANJE RELACIJA MEĐU POJMOVIMA

- Sinonimi pojmova iz kategorije DOJAM (relacija SYN)
  - svi koji znače pozitivno mišljenje – WIN
  - svi koji znače negativno mišljenje – FAIL
- Oracle Text nudi mogućnost korištenja postojećeg tezaurusa ili kreiranje novog
- Moguće definirati i hijerarhijske odnose
  - NT – narrower term (uži pojam, podpojam)
  - BT – broader term (širi pojam, nadpojam)
  - SYN – sinonim

PRIMJER IZ PRAKSE

DEFINIRANJE RELACIJA MEĐU POJMOVIMA

```
begin
CTXSYS.CTX_THES.CREATE_THESAURUS('thes_dojam');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN',
'najkvalitetnijih');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'super');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN',
'kvalitetno');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'blago');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'jeftino');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN',
'prihvatljivo');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'OK');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'dobro');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'ukusno');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'win', 'SYN', 'najbolje');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'fail', 'SYN',
'preskupo');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'fail', 'SYN',
'katastrofa');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'fail', 'SYN', 'gorko');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'fail', 'SYN', 'najgore');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'fail', 'SYN', 'skupo');
CTXSYS.CTX_THES.CREATE_RELATION('thes_dojam', 'fail', 'SYN', 'loše');
end;
```

PRIMJER IZ PRAKSE

STVARANJE DOJMA SINTEZOM POJMOVA

- Sinonimi se koriste operatorom CONTAINS

```
SELECT t.id, t.tweet, 'win' dojam
  FROM DSA_TWITTER_KOMENTARI t
 WHERE CONTAINS (tweet, 'syn(win, thes_dojam)') > 0
UNION ALL
SELECT t.id, t.tweet, 'fail' dojam
  FROM DSA_TWITTER_KOMENTARI t
 WHERE CONTAINS (tweet, 'syn(fail, thes_dojam)') > 0;
```

ID	TWEET	DOJAM
4	Jedno od najkvalitetnijih piva na tržištu, ali preskupo #zlatnopivo	win
5	Meni je cijena OK, okus super, ja odsad pijem samo #zlatnopivo	win
4	Jedno od najkvalitetnijih piva na tržištu, ali preskupo #zlatnopivo	fail
6	Ambalaža totalni #fail, ostalo može proći	fail

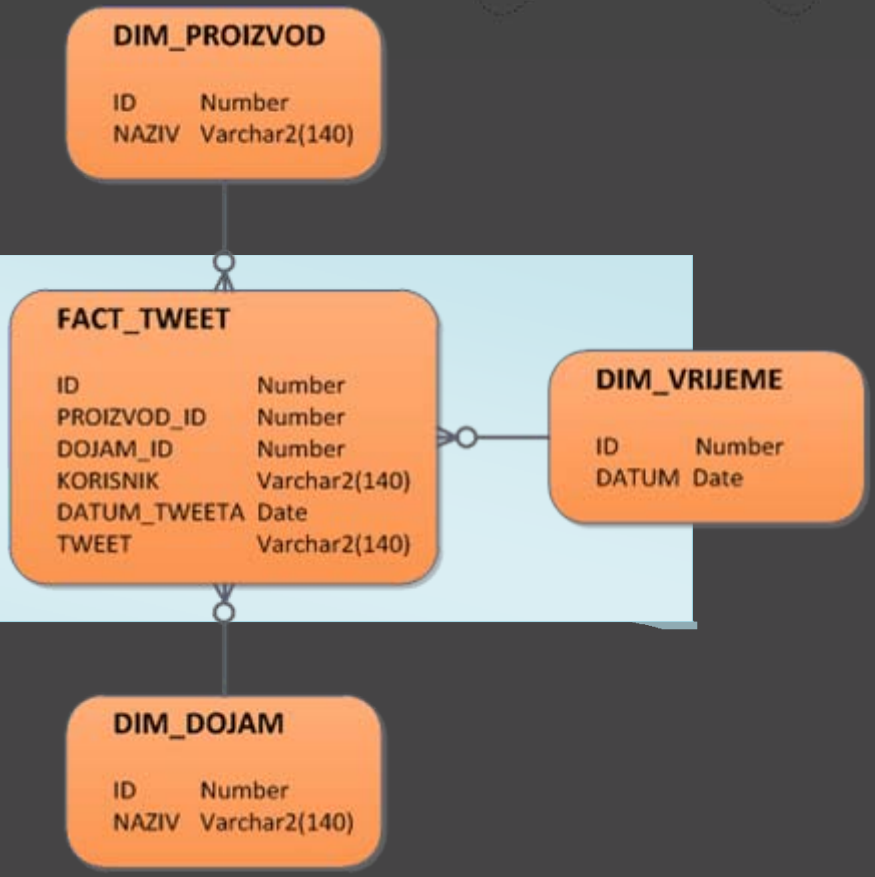
PRIMJER IZ PRAKSE

KREIRANJE DATA MARTOVA

- Narativni data mart
- Pojmovne kategorije pretvaraju se u dimenzije

```
INSERT INTO DIM_DOJAM (ID, NAZIV)
     SELECT id, pojam
       FROM dom_pojmovi
      WHERE kategorija = 'DOJAM';

INSERT INTO DIM_PROIZVOD (ID, NAZIV)
     SELECT id, pojam
       FROM dom_pojmovi
      WHERE kategorija = 'PROIZVOD';
```



PRIMJER IZ PRAKSE

KREIRANJE DATA MARTOVA

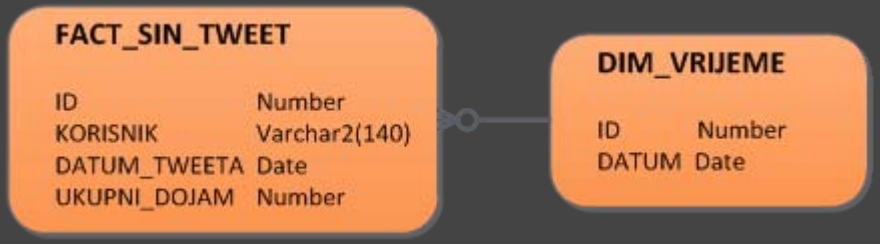
- Fact tablica puni se povezivanjem tweetova s kategorijama DOJAM i PROIZVOD
  - operator NEAR zbog umnožavanja redaka ("cijena OK, okus super")

```
INSERT INTO FACT_TWEET (ID, DOJAM_ID, PROIZVOD_ID, KORISNIK,
            DATUM_TWEETA, TWEET)
(SELECT id, dojam_id, proizvod_id, korisnik, datum_tweeta, tweet
  FROM (SELECT d.*, p.proizvod_id, proizvod_pojam
          FROM    (SELECT t.id,
                          p.id dojam_id,
                          p.pojam dojam_pojam,
                          korisnik,
                          datum_tweeta,
                          tweet
                     FROM DSA_TWITTER_KOMENTARI t, DOM_POJMOVI p
                    WHERE CONTAINS (t.tweet, p.pojam) > 0
                          AND kategorija = 'DOJAM') d
                JOIN
                   (SELECT t.id, p.id proizvod_id, p.pojam proizvod_pojam
                      FROM DSA_TWITTER_KOMENTARI t, DOM_POJMOVI p
                     WHERE CONTAINS (t.tweet, p.pojam) > 0
                           AND kategorija = 'PROIZVOD') p
                ON (d.id = p.id))
 WHERE CONTAINS (tweet,
       'near(('||proizvod_pojam||','||dojam_pojam||'),1,FALSE)') > 0)
```

PRIMJER IZ PRAKSE

KREIRANJE DATA MARTOVA

- FACT_SIN_TWEET sadrži sintetizirane podatke o dojmu
  - pozitivan tweet ima vrijednost 1, negativan tveet ima vrijednost -1, a neutralan tweet ima vrijednost 0




```
INSERT INTO FACT_SIN_TWITTER (ID, KORISNIK, DATUM_TWEETA, UKUPNI_DOJAM)
  (SELECT id, korisnik, datum_tweeta, SUM (dojam) ukupni_dojam
    FROM (SELECT t.*, 1 dojam
            FROM DSA_TWITTER_KOMENTARI t
           WHERE CONTAINS (tweet, 'syn(win, dojam_o_pivu)') > 0
          UNION ALL
          SELECT t.*, -1 dojam
            FROM DSA_TWITTER_KOMENTARI t
           WHERE CONTAINS (tweet, 'syn(fail, dojam_o_pivu)') > 0)
GROUP BY id, korisnik, datum_tweeta)
```

ZAKLJUČAK

- Gdje postoji potreba, postoji i način
- Najveći problem nedostatak alata
  - postojeći alati nisu specijalizirani
  - pokrivaju samo dio integracijskog procesa
  - mogući problemi u korištenju alata različitih proizvođača
- Oracle aduti
  - Oracle Text za transformacije teksta
  - moguća integracija u OWB i ODI
  - OBIEE, Reports, BI Publisher za prezentacijski sloj
- Prednost je što počivaju na zajedničkoj platformi - Oracle DB
- Treba riješiti još neke probleme, ali postojećim alatima se može zadovoljiti potreba


Q&A


WWW.NEOS.HR