INTELLIGENTNI SISTEMI
OSNOVE GROZDENJA

Osnove Grozdenja

Prof. Jurij F. Tasič
Emil Plesnik

Grozdenje: definicija

"The process of organizing objects into *groups* whose members are *similar in some way*"

J.A. Hartigan, 1975

"An algorithm by which objects are grouped in *classes*, so that intra-class similarity is maximized and inter-class similarity is minimized"

J. Han and M. Kamber, 2000

Grozdenje: definicija

- Grozdenje je algoritem nenadzorovanega učenja
  - *Spomnimo: “Izkoristiti zakonitosti vhodov za predstavitev, ki omogoča sklepanje in predikcijo"*
- Posebno pozornost namenimo
  - *skupinam/razredom (proti vrednostim zunaj področja)*
  - *oddaljenost/podobnost*
- Kaj potrebujemo za dobro grozdenje?
  - *Ni (neodvisnega) najboljšega kriterija*
    - **redukcija podatkov** (iskanje predstavnikov homogenih skupin)
    - **naravni podatkovni tipi** (opis neznanih lastnosti naravnih grozdov)
    - **uporabni podatkovni razredi** (iskanje uporabnih in ustreznih skupin)
    - **zaznavanje vrednosti zunaj področij** (iskanje nenavadnih podatkovnih objektov)

Primeri (nekaj) uporabe grozdenja

- Tržne raziskave
  - *Iskanje skupin strank, ki se po značilnostih ujemajo s ciljnim oglaševanjem*
- Biologija
  - *Klasifikacija rastlin in živali glede na njihove značilnosti*
- Zavarovalništvo, telekomunikacije
  - *Grozdenje strank s podobnimi značilnostmi/vedenjem*
  - *Zaznavanje prevar/poneverb*
- Splet:
  - *Klasifikacija dokumentov*
  - *Grozdenje spletnih dnevnikov za odkrivanje skupin s podobnim vzorcem dostopanj*
  - *Priporočilni sistemi (“Če vam je bilo všeč tole, vam bo najbrž tudi tole.")*

Primer: Grozdenje CDjev

- Intuitivno: glasba je razvrščena v kategorije, od katerih jih je vsaki osebi všeč zgolj nekaj

  Ampak kaj dejansko so kategorije?

- Predstavitev CDja z njihovimi uporabniki/kupci
- Podobni CDji imajo podobne uporabnike in obratno
- Zamislimo si enodimenzionalni prostor za vsakega uporabnika

  Vrednosti v posamezni dimenziji so lahko samo 0 ali 1

- Točka CDja v tem prostoru je $(x_1, x_2, \ldots, x_k)$, kjer je $x_i = 1$ če je i_{ti} uporabnik kupil CD

  Primerjava z matriko korelacij

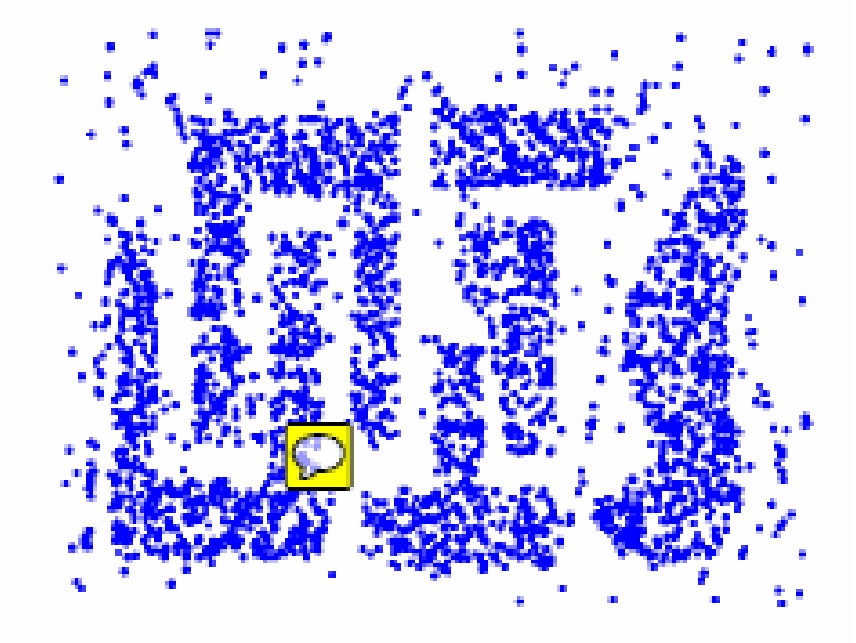
  (vrstice = uporabniki, stolpci = CDji)

Zahteve

- Skalabilnost
- Upravljanje z različnimi tipi atributov
- Odkrivanje grozdov poljubnih oblik
- Minimalne zahteve za določitev vhodnih parametrov
- Delovanje v prisotnosti šuma in vrednosti zunaj področij
- Neobčutljivost na red vhodnih spremenljivk
- Večdimenzionalnost
- Interoperabilnost in uporabnost

Vprašanje

Kaj storiti z naborom podatkov naslednje oblike?



Problemi

V zvezi z grozdenjem obstaja več problemov :

- trenutne tehnike grozdenja vseh zahtev ne obravnavajo zadostno (in sočasno);
- obravnavanje velikih večdimenzionalnih naborov podatkov je lahko težavno s stališča časovne kompleksnosti;
- učinkovitost metod je odvisna od definicije razdalje (za grozdenje na osnovi razdalje);
- če neko naravno merilo razdalje ne obstaja, ga moramo sami določiti, kar pa ni zmeraj enostavno, še posebej v večdimenzionalnih prostorih;
- rezultat algoritma grozdenja (ki je lahko v več primerih poljuben) lahko interpretiramo na več načinov.

Klasifikacija algoritmov grozdenja

- Izključojoči vs Prekrivajoči
- Hirearhični vs Enoplastni
- Od zgoraj navzdol vs Od spodaj navzgor
- Deterministični vs Verjetnostni
- Podatki: simboli ali številke

Podobnost skozi razdalje

- Razdalje so običajno uporabljene kot merilo podobnosti ali razlike med dvema podatkovnima objektoma
- Najpreprostejši primer: imamo številski atribut A
  - *Razdalja (X, Y) = A(X) − A(Y)*
- Več številčnih atributov
  - *Razdalja (X, Y)* = Evklidska razdalja med X in Y
- Nominalni atributi
  - Razdalja je 1, če sta vrednosti različni in 0, če sta enaki
- Ali so vsi atributi enako pomembni?
  - Uteževanje atributov je lahko potrebno

Merilo razdalje

Dva večja razreda merila razdalje:

- Evklidska
  - Evklidski prostor ima določeno število realnih dimenzij in “gostih" točk
  - Ideja povprečja dveh točk
  - Evklidska razdalja je zasnovana na lokaciji točk v takšnem prostoru
- Ne Evklidska
  - Ne Evklidska razdalja je zasnovana na lastnostih točk in ne na njihovi lokaciji v prostoru

Merilo razdalje

Aksiomi merila razdalje:

- *d* je *merilo razdalje*, če je realna funkcija parov točk, tako da velja:
    1. $d(x, y) \geq 0$
    2. $d(x, y) = 0$ če $x = y$
    3. $d(x, y) = d(y, x)$
    4. $d(x, y) \leq d(x, z) + d(z, y)$ (trikotniška enakost)

Razdalje za numerične atribute

- Razdalja Minkowskega:

$$d_{ij} = \sqrt[q]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^q}$$

- Kjer sta $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ in $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ dva p-dimenzionalna podatkovna objekta in je q pozitivno celo število
- Če je $q = 1$, potem je d Manhattan razdalja:

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

Razdalje za numerične atribute

- Razdalja Minkowskega

$$d_{ij} = \sqrt[q]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^q}$$

- Kjer sta $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ in $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ dva p-dimenzionalna podatkovna objekta in je q pozitivno celo število
- Če je $q = 2$, potem je d Evklidska razdalja:

$$d_{ij} = \sqrt[2]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^2}$$

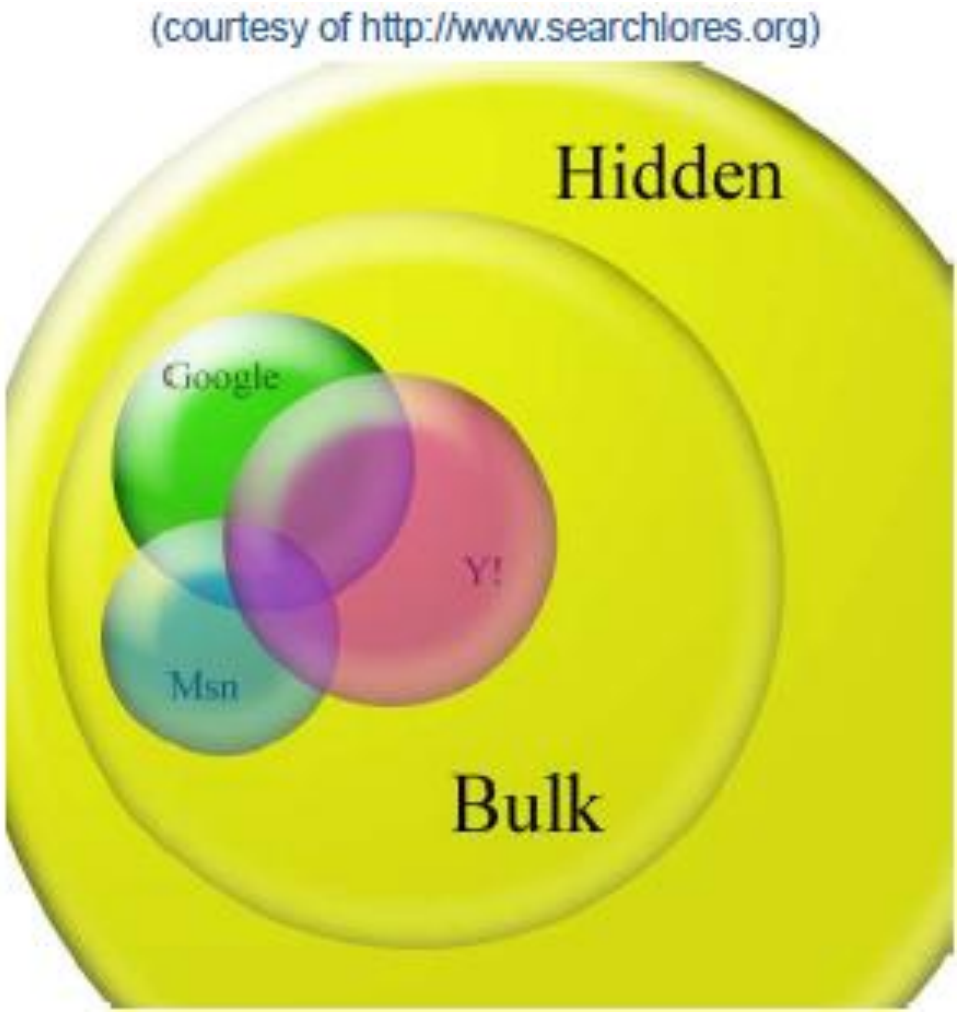
Iskanje/zbiranje podatkov

- večina algoritmov, ki jih boste srečali je ustreznih za grozdenje podatkov – vendar katerih podatkov?!
- Grozdenje je zgolj tehnika, ki jo lahko uporabimo, če najprej zberemo podatke
  - Zato pa moramo razumeti delovanje izvora podatkov, npr. spleta
  - Če poznamo osnove delovanja spleta, lahko:
    - Učinkovito zberemo podatke za grozdenje
    - Uporabimo splet v našo korist, tako da hitro najdemo podatke, ki nas zanimajo

Grozdenje besedil – zakaj?

- Količina podatkov na spletu se povečuje vsak dan:
  - več kot 25 milijard Google indeksov
  - več kot 25 milijard jih ni indeksiranih z nobenim od iskalnikov
  - na tisoče novih strani vsak dan
  - Zahvaljujoč blogom in forumom se število povečuje še hitreje

Struktura spleta



In kaj sedaj?

- Splet:
  - Ne pomeni samo irc, ftp, usenet itd.
  - Ni povsem pokrit z iskalniki
  - Raste zelo hitro
- Kako lahko najdemo informacije na spletu?
  - S pristopom “normalnega uporabnika”
  - S pristopom “iskalnika”

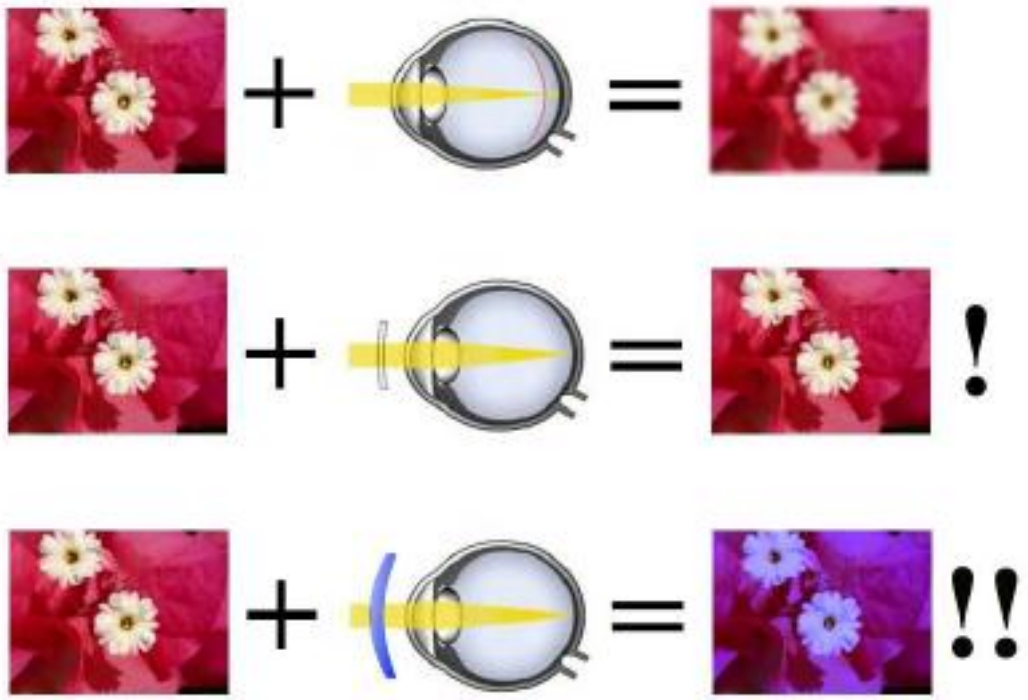
Kako brskati?

- Z uporabo privzetih nastavitev brskalnika vidimo zgolj tisto, kar drugi želijo, da vidimo na način, kot oni želijo:
  - Reklamna okna in banner-ji
  - Množica neuporabne kode, ki je dejansko ne vidimo, vendar jo kljub temu prenesemo
  - Z “aktivno” nedostopno vsebino
  - Prikaz v oknih fiksno določenih velikosti
  - Pomikanje po vnaprej določenih poteh
- Ampak to nam je pač dano, kajne?

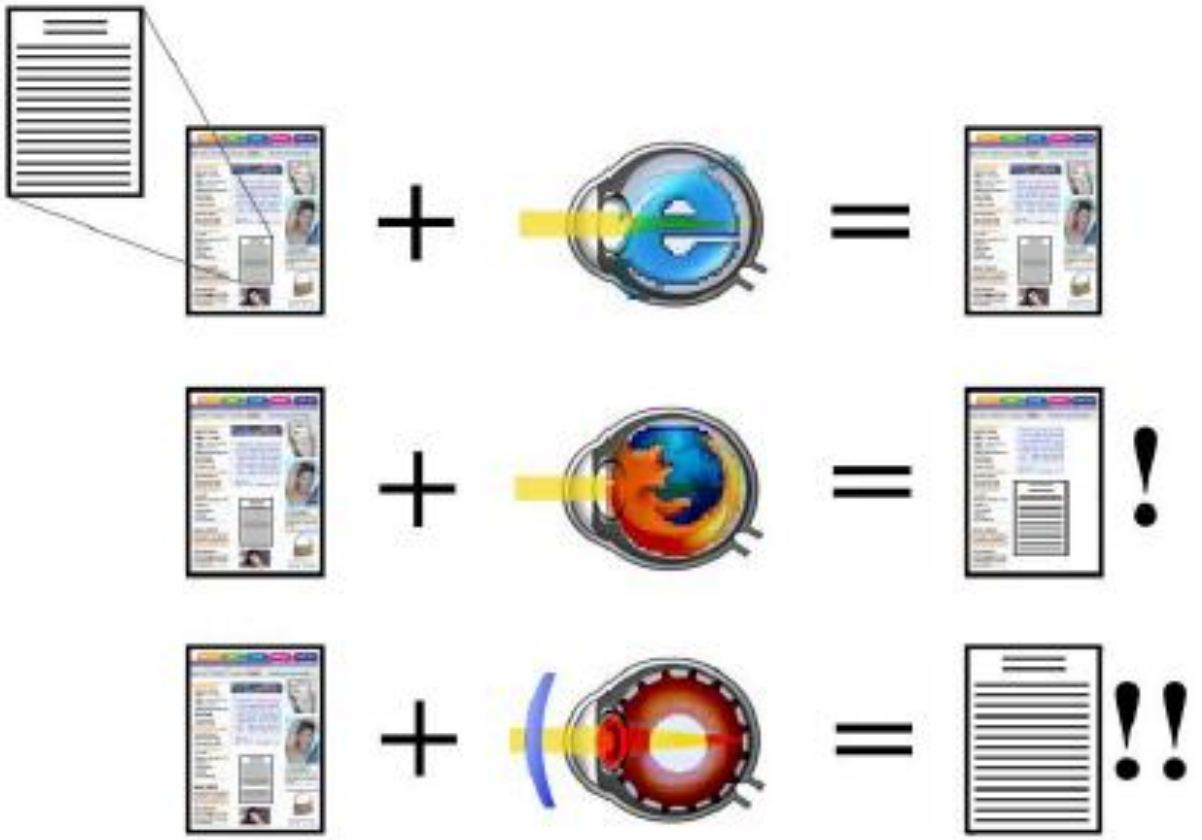
Kako deluje drugače?

- Napačno! PC ni kot televizija. Dosežemo lahko, da počne, kar mi želimo:
  - Prenos zgolj tistih vsebin, ki jih mi želimo
  - Prikazovanje spletnih strani, kot mi želimo
  - Zbiranje in obdelava podatkov

Prava očala



Moč očal



Tehnike in tehnologije

- S katerimi tehnikami lahko uporabnik pridobi največ iz spleta?
  - Osnovne
    - Alternativni brskalniki
    - “leecher” in “teleporter” orodja
    - “spider” in “scraper” orodja
    - Proxy orodja
  - Napredne
    - Učenje enovrstičnih ukazov z uporabo curl, wget, lynx
    - Učenje iskanja
    - Učenje pridobivanja podatkov s spletnih strani
    - Učenje iskanja med pridobljenimi podatki

Učenje iskanja

- Med učenjem iskanja se spoznamo tudi z delovanjem obstoječih iskalnikov in dobimo kakšno idejo za lastni iskalnik:
  - Besedno iskanje (primer “zvezde”)
  - Uporaba izraza “index off”
  - Uporaba več iskalnikov
  - Grozdenje iskalnikov
    - Kaj pomeni grozdenje v tem primeru?
  - Uporaba “folksonomije”
  - Uporaba blogov in forumov

Osnove bot-ov

(ali: spoznajmo kako pridobiti želene podatke s spletnih strani)

- Kaj je bot?
- Kaj naj bi bot delal za nas?
  1. Obiskoval spletne strani in sledil povezavam
  2. Pridobival uporabne podatke
  3. Obdeloval podatke (ali jih zgolj shranil za druge aplikacije)
- Kako ustvariti bot-a?
  - Kompromis med zmogljivostjo in kompleksnostjo
  - Primeren je katerikoli programski jezik (boljša izbira so hitrejši)
  - Uporaba potrebnih knjižnic (http, razčlenjevanje besedil itd.)

Prepoznavanje spletnih vzorcev

- Predstavitveni/brskalni vzorci
- Vzorci v okviru spletnih strani/razredov spletnih strani
- Orodja: vaši možgani
- V obeh primerih je v pomoč avtomatsko generirana koda

Brskanje z boti

- Naloge bot-ov:
  - Prenos spletnih strani
  - Zbiranje ali sledenje povezavam, ki zadoščajo določenim pogojem (vsebovanim v označenem besedilu ali v sami povezavi) do določene globine ali za vedno
  - Izpolnjevanje obrazcev

Spletne tehnologije

- Izdelava dobrega bota zahteva poznavanje naslednjih področij:
  - HTTP
    - GET in POST
    - Referer
    - UserAgent
    - Cookie
    - Proxy
  - HTML
    - Obrazci
    - Dinamično generirana koda

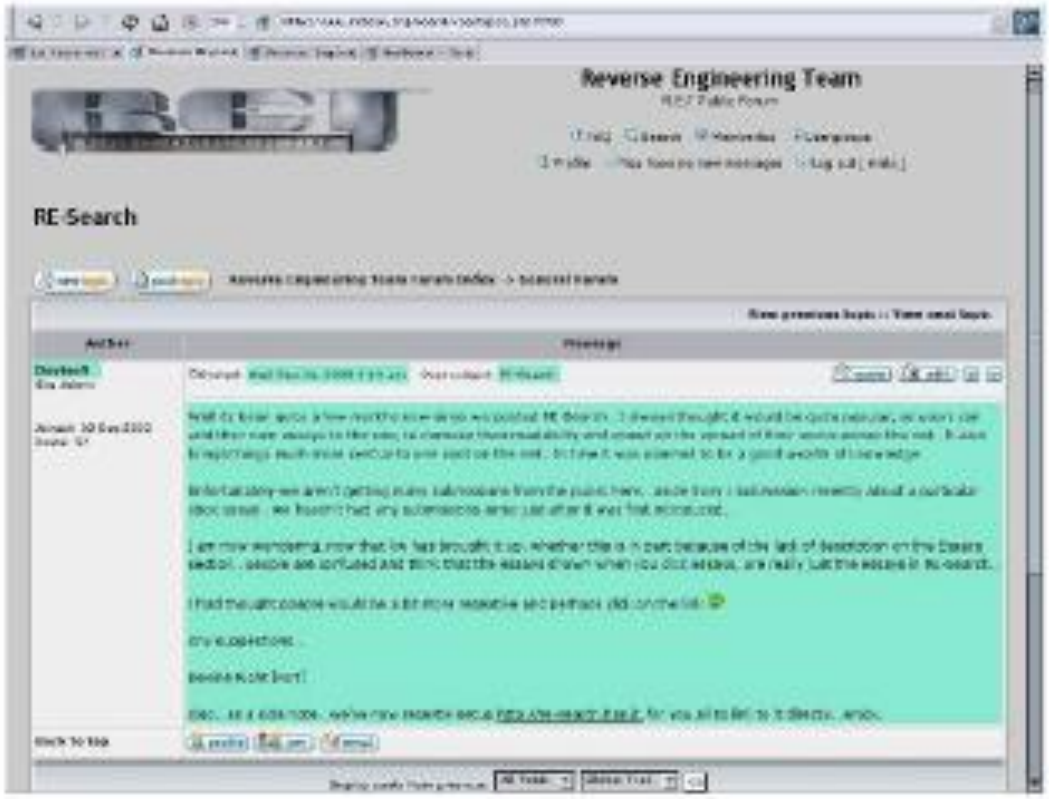
Primer

- Projekt TWO (The Working Offline) bralnik
  - Prenos spletnih strani s spletnih forumov
  - Pridobivanje informacij s prenesenih strani
  - Normalizacija podatkov in shranjevanje v podatkovno bazo

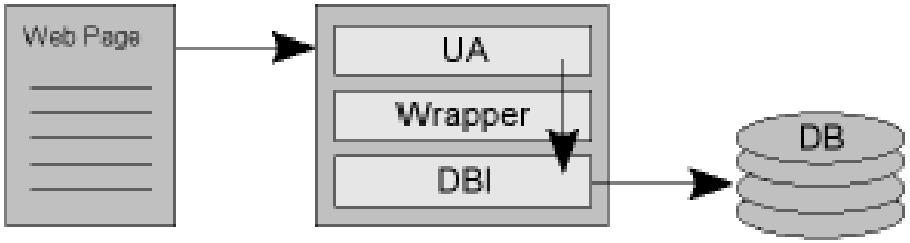
  - Dostopen na: http://two.sourceforge.net/

Ideja

Slika: za ogled strani na sliki je potrebno prenesti 140 kB podatkov. Zanimivi (ozačeni) del pa obsega manj kot 1,4 kB, kae je manj kot 1% vsebine!



Struktura TWO



Zmogljivosti TWO

```
------------------------------
Test started at       22:49:00
Test finished at      00:28:00
------------------------------
Total test time       01:39:00
------------------------------


------------------------------
Downloaded pages          2245
Saved messages           13693
Bytes count           94967139
------------------------------


------------------------------
DB size before test (KB) 3288
DB size after test (KB) 11180
------------------------------
Total data size.(KB).... 7892
------------------------------


------------------------------
Forum data size (KB)     92741
TWO's data size (KB)      7892
------------------------------
Saved space.(KB).........84849
Saved space (perc)........91%
------------------------------
```

```
Pages count: 2244
Bytes count: 94943907
Messages count: 13692
Saving message #17986
-----
GET http://board.anticrack.de/viewtopic.php?t=2491
User-Agent: Two/0.01
Cookie: phpbb2mysql_sid=d68ccd982732c219e55cff36ec5fa44f;
Cookie2; $Version="1"



Pages count: 2245
Bytes count: 94967139
Messages count: 13693
Saving message #17989
nala@kami:~/prj/last$
```

```
kami:/var/lib/mysql# du two
3288    two
kami:/var/lib/mysql# du two
11180   two
kami:/var/lib/mysql#
```

Bibliografija

- "Metodologie per Sistemi Intelligenti" course - Clustering Tutorial Slides by P.L. Lanzi
- "Data mining" course - Clustering, Part I Tutorial slides by J.D. Ullman
- Satnam Alag: "Collective Intelligence in Action" (Manning, 2009)
- http://www.searchlores.org
- http://davide.eynard.it/malawiki/PowerBrowsing
- As usual, more info on del.icio.us

- Literatura:
  - Sargur N. Srihari, Lecture notes, University at Buffalo, State University of New York

HVALA ZA POZORNOST