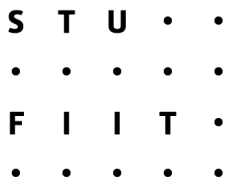
SluiceBox - a pattern mining tool

Tomáš Kramár

Faculty of Informatics and Information Technology

9. apríla 2008



Cieľ

Dolovanie vzorov používania (pravidiel) fazetového prehliadača *Factic* v doméne pracovných ponúk.

Asociačné pravidlá

Vzory, pravidelnosti v dátach, v správaní návštevníkov

/products.html $\rightarrow$ /order.html [0.8, 0.9]

90% ľudí, ktorí klikli na odkaz products.html, klikli aj na odkaz order.html a toto pravidlo sa vyskytlo v 80% všetkých transakcií.

Druhy pravidiel

- Asociačné pravidlá
- Sekvenčné pravidlá - zachytávajú aj poradie akcií

restriction:offer-job#bCar → restriction:region#GB [0.19, 1]

restriction:region#GB → restriction:offer-job#bCar [0.19, 0.8]

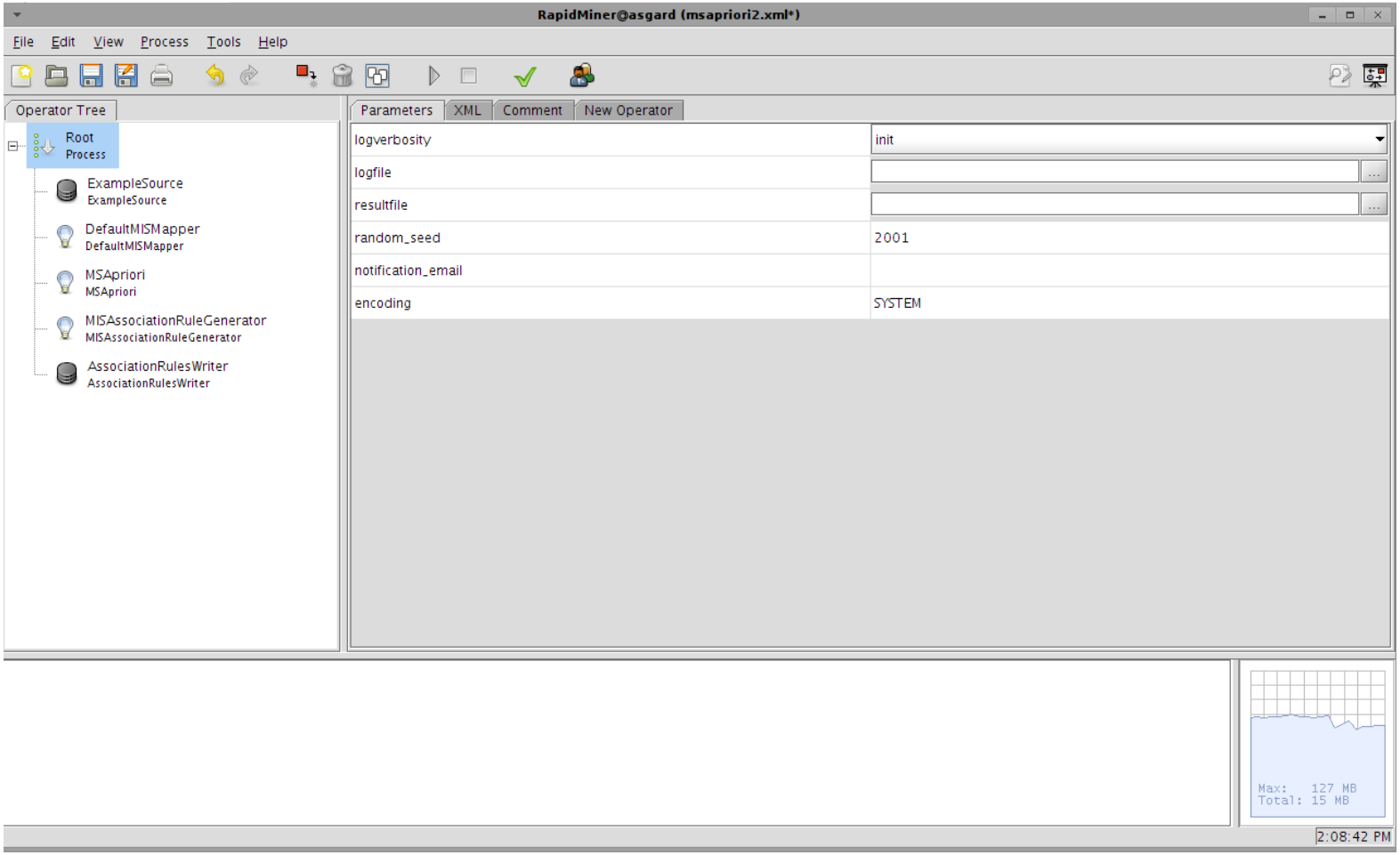
Problémy

- generovanie systémom “každý s každým”
- obrovské množstvá pravidiel
- potreba inteligentného filtrovania

RapidMiner

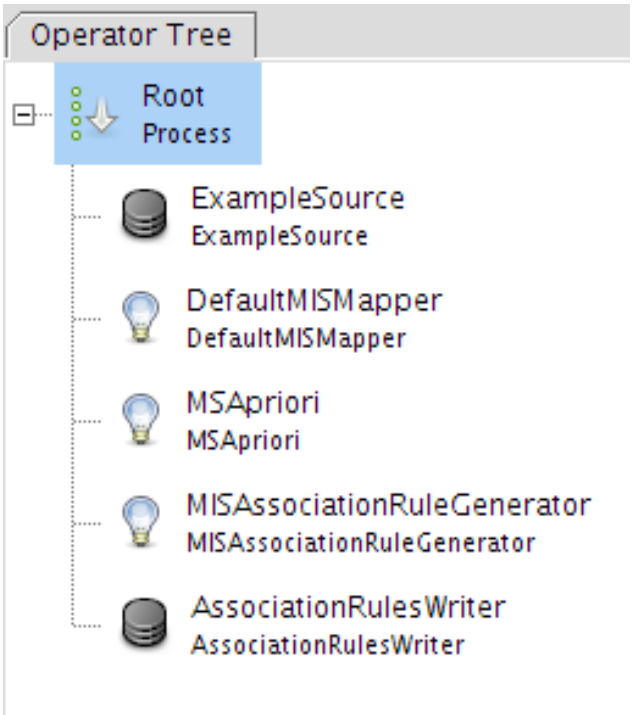
- operátory
- reťazenie operátorov
- dobrá modularita
- bugy
- slabá podpora dolovania vzorov

RapidMiner GUI



Obrázok: RapidMiner GUI

RapidMiner operator tree



Obrázok: Operator tree

SluiceBox

- webová aplikácia
- dolovanie logov z *Factic*-u
- logy zachytené *SemanticLog*-om
- *RapidMiner* ako process engine
- zameranie na filtrovanie
  - dolovanie s podporou MIS
  - klasifikátory zaujímavosti
  - šablóny

Moduly

Moduly:

- import dát – interne sa doluje nad textovými reťazcami – možnosť dolovania z iných systémov, textových dokumentov
- dolovanie – končí zápisom do db
- filtrovanie – začína načítaním z db

Uplatnenie *a posteriori* princípu filtrovania

Úprava vstupných dát

- Náhrada URI

http://nazou.fiit.stuba.sk/nazou/ontologies/v0.6.17/region#World
=> restriction: region#World

- Nahradenie položiek za kategórie

offer-job-inst#Offer_jakubik_f650b3e70bd54d92df0a6edf5958756b
=> showOfferDetails

- Odfiltrovanie položiek (login)

Dolovanie s podporou MIS

každá položka má inú frekventovanosť – *rare item problem*

id	event
1	ShowOverview
2	ShowDetails
4	UserLogin
5	UserLogout
6	PageNext
7	PagePrevious
8	PageSelect
9	SelectRestriction
10	FacetEnable
11	FacetDisable
12	SelectItemsPerPage
13	SelectSortingOrder

Klasifikátory zaujímavosti

Vyjadrujú silu pravidla. Neexistuje najlepší klasifikátor, ale existujú dobré a zlé klasifikátory. Každý z nich má istú sémentiku.

- počítajú sa pri vygenerovaní pravidla
- pravidlá sa zapíšu do databázy
- *a posteriori* filtrovanie
- support, confidence, laplace, gain, conviction . . .

Šablóny

Typy

- inkluzívne
- reštriktívne

Každá akcia musí mať priradenú svoju kategóriu (typ udalosti).

Príklady:

Inclusive

Any* $\rightarrow$ **FacetDisable**
vyberie akcie, ktoré viedli k zakázaniu fazety

Restrictive

SelectRestriction, Any* $\rightarrow$ **Any** vyberie akcie, ktoré nesúvisia s manipuláciou faziet

Class association rules

Vhodné nahradenie konkrétnych položiek za ich kategórie.
Namiesto:

restriction:region#GB, restriction:classification#pcProgrammer → showOfferDetails
restriction:region#GB, restriction:classification#pcProfessionals → showOfferDetails
restriction:region#GB, restriction:classification#pcTechnicians → showOfferDetails

len:

restriction:region#GB, restriction:classification → showOfferDetails

Záver

Výstup:

- implementácia algoritmov
- plugin do RapidMiner-u
- SluiceBox