

Dolovanie vzorov používania webového sídla

Tomáš Kramár

Fakulta Informatiky a Informačných Technológií

13. júna 2008

S T U . .
.
F I I T .
.

Vzory a pravidlá



Vyjadrujú pravidelnosti v dátach, v správaní návštevníkov

`/produkt.html → /objednaj.html [0.8, 0.9]`

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

Čo je dolovanie pravidiel?

Je to hľadanie takých pravidiel, ktorých metriky majú väčšiu hodnotu, ako používateľom zadané minimálne hodnoty.



Objavené pravidlá môžeme využiť na:

- Získanie prehľadu o správaní návštevníkov
- Presunutie položiek (tovarov, odkazov) pre uľahčenie navigácie (zväčšenia zisku)
- Odporúčanie odkazov (e-learning) alebo produktov
- Predpovedanie správania

Problémy



Problémy dolovania pravidiel:

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

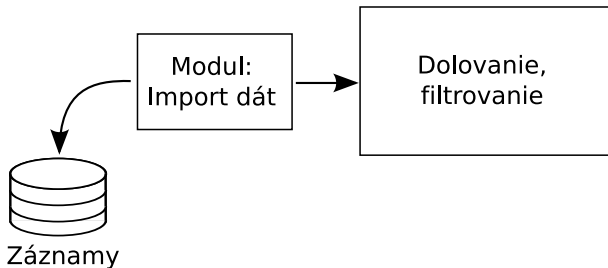
Žiadny z existujúcich najpoužívanějších open-source nástrojov tieto problémy nerieši.



Prehľad:

- webová aplikácia
- dolovanie logov zachytených nástrojom *SemanticLog* z fazetového prehliadača *Factic* z projektu *NAZOU*
- rozširuje *RapidMiner*
- zameranie na filtrovanie
 - dolovanie s podporou viacerých minimálnych podpôr
 - klasifikátory zaujímavosti
 - šablóny

Architektúra



Obrázok: Hrubá architektúra nástroja SluiceBox

Vstupný modul je zodpovedný za načítanie dát a ich transformáciu.

Kategorizácia vstupných dát



- Viacero použitých algoritmov (MS-Apriori, šablóny) vyžaduje pre svoju činnosť kategorizáciu položiek.
- Kategorizácia prebieha vo vstupnom module, kde je jednotlivým položkám priradený prefix ich kategórie.

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

- Položky sa tak stávajú čitateľnejšie a informatívnejšie

Dolovanie s podporou viacerých minimálnych podpôr



- Jedna hodnota minimálnej podpory nie je vhodná pre všetky položky – niektoré položky sa v dátach vyskytujú prirodzene častejšie, tie ktoré sa vyskytujú menej často sú však vzácnejšie.
- Napr.: chlieb vs. kosačka
- SluiceBox implementuje algoritmus *MS-Apriori*.
- Minimálna podpora sa nenastavuje pre každú položku zvlášť (zdĺhavé, nepraktické), ale len pre kategóriu.

Dolovanie s podporou viacerých minimálnych podpôr



Association rules with multiple minimum support

Parameters

Default Support	0.2
Confidence	0.5
showOfferDetails	0.2
selectSortingOrder	0.2
restriction	0.2
pageSelect	0.2
facetDisable	0.2
facetEnable	0.2

Mine

Obrázok: Zadávanie minimálnych podpôr pre kategórie

Klasifikátory zaujímavosti



- Každému nájdenému pravidlu priradíme viacero *váh*, ktoré číselne vyjadrujú silu pravidla. Klasifikátor zároveň vyjadruje určitú sémantiku a prispieva k interpretácii pravidla.

```
/products.html → /order.html [podpora=0.8, spoľahlivosť=0.9]  
/products.html → /order.html [podpora=0.1, spoľahlivosť=0.9]  
/products.html → /order.html [podpora=0.8, spoľahlivosť=0.1]
```

- počítajú sa pri vygenerovaní pravidla
- pravidlá sa zapíšu do databázy
- *a posteriori* filtrovanie

Filtrovanie pravidiel podľa zadanej minimálnej hodnoty klasifikátora.

Klasifikátory zaujímavosti



16 items found, displaying 11 to 16.

[\[First/Prev\]](#) **1**, **2** [\[Next/Last\]](#)

Rule	Support	Confidence	Conviction	Lift	Laplace	Gain	Ps
[restriction:region#World] => [restriction:region#Europe_Western]	0.33	Support	1.87	1.53	0.91	-0.6	0.12
[restriction:region#Europe_Western] => [restriction:region#World]	0.33	Support of a rule represents the percentage of transaction, where the rule was found. The higher the support, the better the rule.					
[restriction:region#Europe_Western] => [restriction:region#Europe]	0.33	0.12	1.87	1.53	0.91	-0.6	0.15
[restriction:region#Europe] => [restriction:region#Europe_Western]	0.33	0.83	3.2	1.79	0.95	-0.47	0.15
[restriction:region#World] => [restriction:region#Europe]	0.4	Support might vary from <0, 1>, 0 being the worst and 1 the best.					
[restriction:region#Europe] => [restriction:region#World]	0.4	-	-	2.14	1	-0.4	0.21

Export options: [CSV](#) | [Excel](#) | [XML](#) | [PDF](#)

Obrázok: Zoznam pravidiel

Šablóny



Typy

- inkluzívne
- reštriktívne

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

Príklady:

Inclusive

Any* → **FacetDisable**

vyberie akcie, ktoré viedli k zakázaniu fazety

Restrictive

SelectRestriction, Any* → **Any** vyberie akcie, ktoré nesúvisia s manipuláciou faziet

Šablóny

**Templates**

Define a template here.

`facetDisable* => sortingOrder`**sortingOrder is not a valid category.**Inclusive ☒Restrictive ☐

Filter

List of all available categories. Click on the category to copy its name into template definition.

- [Any](#)
- [showOfferDetails](#)
- [selectSortingOrder](#)
- [restriction](#)
- [pageselect](#)
- [facetDisable](#)
- [facetEnable](#)

Stored templates

Choose a template from the list.

```
selectSortingOrder, showOfferDetails, facetEnable => facetEnable*  
restriction, selectPage => restriction  
restriction => restriction
```

[Or create a new one](#)

Obrázok: Editor šablón

Experimenty



Vzorka záznamov z nástroja Factic: 30 transakcií, v každej priemerne 7 klikov.

Pravidlá z Facticu

```
[restriction:region#World] => [restriction:region#Europe] [0,4; 0,86]  
[restriction:region#Europe_Western] => [showOfferDetails] [0,3; 0,64]  
[restriction:region#World] => [showOfferDetails] [0,33; 0,71]  
[restriction:#pcProfessionals] => [showOfferDetails] [0,27; 0,73]
```

Zhrnutie



- Webová aplikácia
- Zameranie na filtrovanie (dolovanie s podporou viacerých minimálnych podpôr, šablóny, klasifikátory zaujímavosti)
- Možnosť použiť na viacero zdrojov dát – vymeniteľný vstupný modul
- Implementované s použitím nástroja RapidMiner
- Implementované algoritmy sú obalené do RapidMiner operátorov
- RapidMiner plugin
- Článok na IIT.SRC 2008