Utilization of Parallel Web Browsing Patterns for Browser Action Recommendation

Autor: Bc. Martin Toma
Vedúci práce: Ing. Martin Labaj

FIIT STU, 14.05.2014

Obsah prezentácie

- Krátka motivácia a ciele
- Získanie dát
- Predspracovanie dát
- Dolovanie sekvencií

- Využitie v odporúčaní
  - výber vzoru a akcie
  - jednotlivé verzie vzoru

- Zhrnutie a rady
- TabRec

Motivácia a ciele

- Paralelné prehliadanie
  - veľmi aktuálne a populárne [1],
  - stále však nedostatočne preskúmané [2],
  - skoro žiadne využitie už získaných informácií [3].
- Naše ciele
  - získať presné a ľahko spracovateľné dáta (Brumo[1]),
  - identifikovať vzory v paralelnom prehliadaní,
  - využiť tieto zistenia na odporúčanie akcií v prehliadači.

[1] http://brumo.fiit.stuba.sk/

Získanie dát

- Existujúce dáta (projekt Brumo)
  - nie ideálna štruktúra,
  - niektoré udalosti detekované nespoľahlivo.
- Vlastná implementácia
  - definícia požadovanej štruktúry,
  - spoľahlivejšie Chrome API [2],
  - pokrytie väčšieho množstva používateľov.

[2] https://developer.chrome.com/extensions/api_index

Štruktúra získaných dát

id	540777
user_id	"79b9aec6-dba9-457c-9997-87ba96cdf515"
event_id	5
created_at	"2015-02-18T16:39:52.716Z"
tab_id	204
timestamp	1424275741642
window_id	1
index_from	null
index_to	null
url	"f651fc16595664af9b86c974abdd6a89e4a946d4"
session_id	"27e96a3f-728b-4d09-9767-090fe362e912"
domain	"82b49183978af605fb03e940fea84ff94b7de626"
path	"6f04060ab8d9c13468b842e6a19056d79d000eef"
subdomain	"f35755423a19541c5fa5a36f8951e1e834cc57a5"

Predspracovanie dát

- Minimalizácia návrhom štruktúry dát
- Zlepšenie reprezentácie
  - vytvorenie abstraktnejších udalostí
- Pre použitie v algoritmoch
  - GSP[3] v RM vyžaduje tzv. *"wide column"* formát,
  - GSP ignoruje duplikátne položky v transakcii.

[3] Generalized Sequential Patterns

Dolovanie sekvencií I

- Existujúce prístupy
  - asociačné pravidlá a sekvenčné vzory, GSP, FP-Growth a modifikácie
    - implementované v Rapidminer
    - hlavný problém ignorancia duplikátnych položiek
  - implementácia iného sofistikovaného riešenia príliš náročná na čas

Dolovanie sekvencií II

- Naše riešenie
  - Problém je nájsť najčastejšie a čo najdlhšie sekvencie
  - GSP pracuje s niekoľkými premennými
    - customer_id, timestamp, min/max gap, window_size, min_support
  - 1. Prechod cez všetky záznamy vytvorí sekvencie podľa premenných
  - 2. Prechod vyhodnotí početnosť

Využitie v odporúčaní

- Prehliadanie - riešenie úloh
  - vysoko-úrovňové
    - nájdenie určitej informácie,
    - porovnávanie,
    - učenie sa.
  - nízko-úrovňové
    - nájdenie konkrétne karty,
    - upratanie priestoru (zoradenie atď.),
    - uloženie, zatvorenie, etc.
- Identifikácia zámeru umožňuje automatizáciu
  - hlavne nízko-úrovňových úloh ako zoraďovanie, hľadanie, uloženie...

Výber vzoru a akcie

- Manuálna analýza najčastejších sekvencií
- Podrobnejšia štúdia záznamov v ktorých sa vzory vyskytli
- Snaha pochopiť čo používateľ chce dosiahnuť

Verzie “multi activate” vzoru I

- Verzia 1
  - 4 prepnutia medzi kartami detekované v konštantnom časovom okne (max_gap v GSP)
- Verzia 2
  - Pridaný timeout po akceptovaní alebo zrušení odporúčanej akcie.

Verzie “multi activate” vzoru II

- Verzia 3
  - Pridanie prvej verzie running average a pridanie podmienky nesusednosti kariet.
- Verzia 4
  - Vylepšený running average (threshold) a pridaná podmienka minimálneho počtu 3 rôznych kariet pri prepínaní.

Aktuálny stav

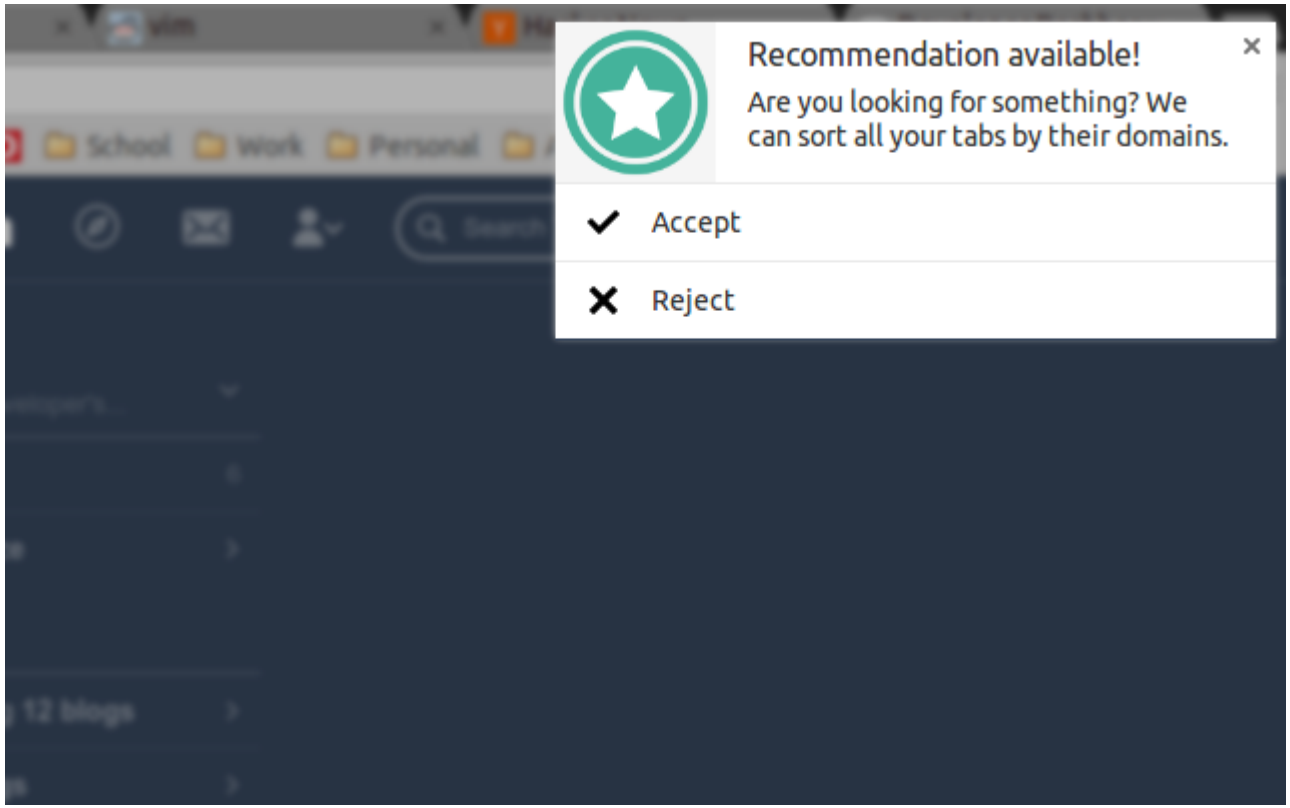
- Vyladené a funkčné zaznamenávanie paralelného prehliadania (data source).
- Implementovaný vlastný algoritmus hľadania najčastejších sekvencií založený na GSP.
- Vybraný vzor implementovaný, detekcia v reálnom čase a odporúčaná akcia.
- Personalizácia v podobe výpočtu *"running average gap"* a odporúčacieho módu.

Aktuálny pohľad na dáta

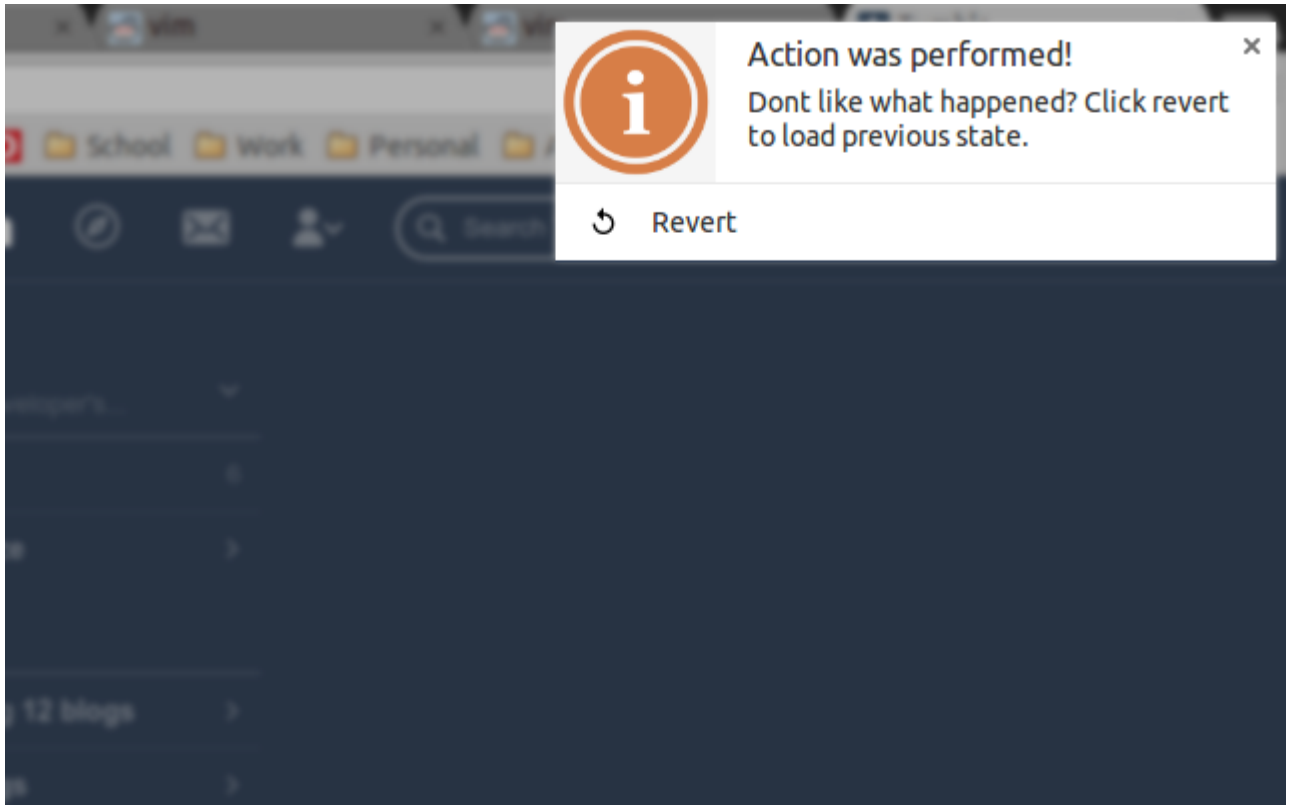
Počet používateľov	57 / 23
Počet sedení	1 011
Počet zaznamenaných akcií	602 877
Počet vykonaných odporúčaní	325 [4]
Akceptovaných	48
Neakceptovaných / vrátených	273 / 4

[4] prvý záznam 1.4.2015

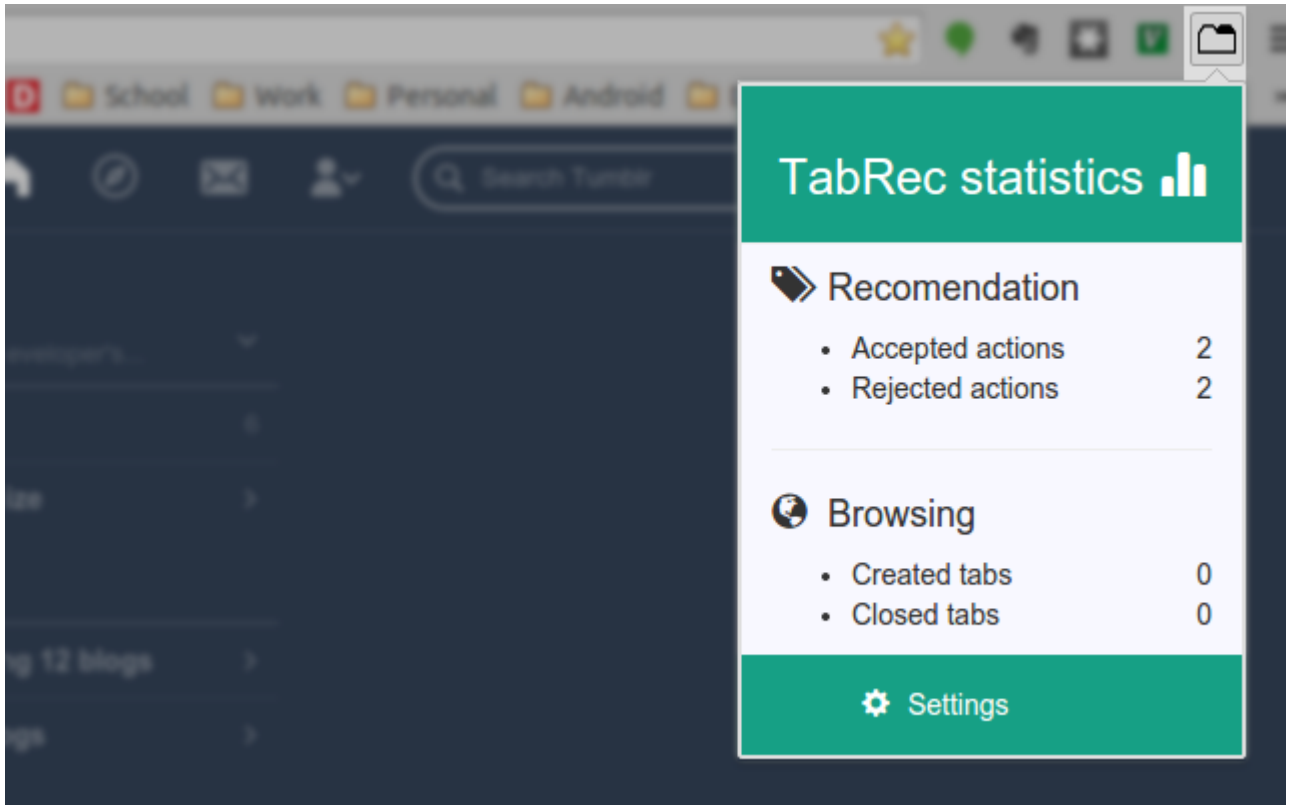
TabRec - detekovaný vzor



TabRec - obnovenie stavu



TabRec - štatistiky



Čo ďalej

- Demo na IIT.SRC 2015
- Doladenie *"Multi Activate"* vzoru
- Vyhodnotenie podľa verzie vzoru
- Doladenie klienta aj API na možné pokračovanie - jednoduché pridávanie ďalších vzorov.

Rady na záver

- Získanie dát
  - kvalita dát vs. trvanie implementácie vlastného zdroja vs. možnosť pred-spracovania dostupných dát,
  - nepodceniť časovú náročnosť vlastnej implementácie.
- Predspracovanie dát
  - vhodný zdroj dát veľmi pomôže,
  - problém inkrementálnosti,
  - často veľká časť analýzy dát.

Rady na záver II

- Dolovanie sekvencií
  - adaptácia existujúcich vs. vlastné metódy,
  - opäť často inkrementálny proces,
  - overenie výsledkov.
- Odporúčanie akcií
  - ani najlepšie logy nezachytia všetky používateľove myšlienky,
  - predpokladať čo používateľ urobí je zložité :)

Zdroje

[1] Weinreich, H., Obendorf, H., Herder, E., Mayer, M.: Off the Beaten Tracks: Exploring Three Aspects of Web Navigation. In: Proceedings of the 15th International Conference on World Wide Web. WWW ’06, ACM, 2006, pp. 133–142.

[2] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. Not quite the average: An empirical study of web use. ACM Trans. Web, 2(1): 5:1–5:31, March 2008.

[3] Jeff Huang and Ryen W. White. Parallel browsing behavior on the web. Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT ’ 10, page 13, 2010.

TabRec extension

1. http://tabber.fiit.stuba.sk
2. Klik na *"Add to chrome"*
3. Otvorí karta s nastaveniami, poprosíme vyplniť a uložiť.

★ https://github.com/martin-svk/tabrec

@ martin.toma.svk@gmail.com