

VYUŽITIE VZOROV PARALELNÉHO PREHLIADANIA NA ADAPTÍVNOM WEBE

Autor práce: Bc. Martin Toma

Vedúci práce: Ing. Martin Labaj

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Bratislava, 10.06.2015

Paralelné prehliadanie webu

Využitie mechanizmov ako sú karty (angl. tabs) a okná (angl. windows), ktoré nám dovoľujú prehliadať viaceré webové stránky súčasne.

- ▷ veľmi aktuálne a populárne¹,
- ▷ stále však nedostatočne preskúmané²,
- ▷ skoro žiadne využitie získaných informácií³.

¹ Weinreich, H. a kol., 2006 ² Weinreich, H. a kol., 2008 ³ Huang, J. a kol., 2010

Vzory správania a WUM

Vzormi nazývame opakované, navzájom súvisiace udalosti. Vzory správania predstavujú spoločné črty a súvislosti.

- ▷ Web Usage Mining (WUM)
 - jedna z oblastí dolovania dát na webe,
 - proces aplikácie metód dolovania dát s cieľom objavenia vzorov¹.

¹ Weinreich, H. a kol., 2006

Aktuálny stav v oblasti

▷ Otvorené problémy

- Zisk kvalitných dát od dostatočného množstva používateľov.
- Analýza dát s inkrementálnou povahou.

▷ Potenciál v oblasti

- Možnosti využitia získaných dát a prípadných objavených vzorov.
- Prispôsobovanie, odporúčanie na základe charakteristík (vzorov) pri prehliadaní.

Ciele práce

1. Navrhnuť a implementovať metódu na získanie dostatočného množstva **korektných a vhodne štruktúrovaných** dát o **paralelnom prehliadaní**.
2. **Analyzovať** tieto dáta pomocou navrhnutej **metódy** za účelom objavenia **vzorov**.
3. Navrhnuť a implementovať ukážku **možnosti využitia** týchto vzorov.

Existujúce riešenia

- ▷ Získavanie dát
 - Brumo⁶,
 - DOBBS⁷,
 - klient vs. server.
- ▷ Objavovanie (dolovanie) sekvenčných vzorov*
- ▷ Využitie vzorov
 - rozšírenia pre webové prehliadače

⁶ Browser-based User Modelling and Personalization Framework (<http://brumo.fiit.stuba.sk>)

⁷ DERI Online Browsing Behavior Study (<http://dobbs.deri.ie/>)

Objavovanie sekvenčných vzorov

- ▷ Asociačné pravidlá a sekvenčné vzory
 - problém nákupného košíka⁵, $X \rightarrow Y$
 - v praxi sú používané skôr limitované sekvenčné vzory⁴.
- ▷ Markovove reťazce
 - nasledujúci stav na základe aktuálneho a K predchádzajúcich
- ▷ Algoritmus GSP⁸
 - dolovanie vzorov v sekvenčných dátach, výpočtovo efektívny,
 - atribúty: *customer, timestamp, min gap, max gap, window size, min support*.

⁴ Bonnin, G. a kol., 2010 ⁵ Agrawal, R. a kol., 1993 ⁸ Generalized Sequential Pattern

Využitie objavených vzorov

- ▷ Vybrali sme si odporúčanie akcií vo webovom prehliadači.
- ▷ Prečo?
 - možnosť riešiť problémy vyplývajúce z paralelného prehliadania,
 - potenciál zvýšenia používateľovej efektivity práce v prehliadači a s tým súvisiaca možnosť získania ďalších dát.

Paralelné prehliadanie ešte nebolo študované priamo z pohľadu prediktívneho modelovania webu (angl. Web Predictive Modelling)⁴.

⁴ Bonnin, G. a kol., 2010

Návrh riešenia I.

- ▷ Požiadavky na dáta
 - musia byť dostatočne presné,
 - prispôsobené na ďalšie spracovanie,
 - zabezpečené z pohľadu anonymity údajov.
- ▷ Navrhnuté riešenie
 - klient (Google Chrome extension⁹)
 - server (REST API a databáza).

⁹ https://developer.chrome.com/extensions/api_index

Návrh riešenia II.

▷ Klient

- zaznamenávanie operácií nad kartami,
- anonymizácia údajov ako ID používateľa či URL adresy,
- pravidelné odosielanie záznamov na server,
- detekcia vybraných vzorov a odporúčanie akcií v reálnom čase a zaznamenávanie ich výskytov.

▷ Server

- zabezpečenie perzistencie dát od klientov,
- objavovanie vzorov v dostupných dátach,
- poskytovanie štatistík.

Návrh riešenia III. (modifikácia GSP)

- ▷ Modifikácia v podpore duplicitných položiek.
- ▷ Špecifikácia atribútov s cieľom maximalizovať dĺžku netriviálnych sekvencií.
- ▷ Postup
 1. Prechod cez všetky záznamy vytvorí sekvencie ohraňované podľa zvolených premenných,
 2. Prechod vyhodnotí početnosť.
 3. Nepovinných krok spojí niektoré po sebe idúce akcie do logického celku.

Implementácia I. (klient)

- ▷ Zaznamenávanie operácií
 - Chrome API, UUID¹¹ ako ID používateľa a sedenia,
 - SHA1¹² na URL, domény, subdomény.

- ▷ Detekcia vzorov v reálnom čase
 - udržiavanie aktuálnej sekvencie,
 - výpočet pohyblivého priemeru v reálnom čase,
 - stavový automat pre každý vzor.

- ▷ Použité technológie
 - JavaScript, CoffeeScript, JQuery,
 - HTML5, CSS3 (SCSS), Bootstrap.

¹¹ A Universally Unique Identifier <https://www.ietf.org/rfc/rfc4122.txt>

¹² US Secure Hash Algorithm 1 <https://tools.ietf.org/html/rfc3174>

Implementácia II. (server)

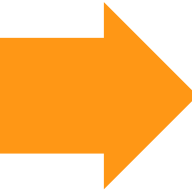
- ▷ Perzistencia dát
 - REST API pre ukladanie záznamov operácií nad kartami a detekovaných vzorov.
- ▷ Analýza dát a objavovanie vzorov
 - exploratory analýza, objavovanie vzorov, spájanie akcií do zmyslupnejších celkov,
 - implementované formou skriptov (Rake tasks).
- ▷ Štatistiky
 - API koncové body pre štatistiky prehliadania a odporúčania.
- ▷ Použité technológie
 - Ruby, Sinatra, PostgreSQL, Nginx a Passenger

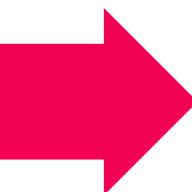
Príklad záznamu o používaní kariet

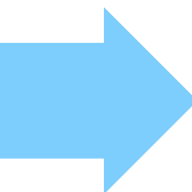
```
user_id: "2f54319b-9882-4ae0-932f-cb4d442e8526"  
session_id: "d36bf37d-5a27-4a72-b02f-9136fe3c1392"  
event_id: 2  
tab_id: 204  
window_id: 1  
url: "c8ca8846-0e12-11e5-a6c0-1697f925ec7b"  
path: "7dfc14e2-0d7b-4c17-9dd5-f1d72ca5e814"  
domain: "602e0188-09d8-46ff-b9ac-1523116949d7"  
subdomain: "db43cb77-0506-4f59-85cf-eb8d4475c04a"  
index_from: null  
index_to: null  
timestamp: 1425225019464
```

Výsledky práce

1. Získané dáta o používaní kariet v prehliadači.
2. Objavené a vybrané vzory a špecifikované akcie.

 **1,074,932** záznamov

 **1,814** sedení (angl. sessions)

 **97 / 47** aktívnych* používateľov

* minimálne 100 vyprodukovaných záznamov. Štatistiky ku dňu 01.06.2015.

Objavené vzory I.

- ▷ Zvolené parametre GSP
 - **min. rozdiel:** 5 sekúnd, **max. rozdiel:** 0 sekúnd
 - **veľkosť okna:** 100 sekúnd, **podpora:** 3 %
- ▷ Nutnosť filtrovania triviálnych sekvencií
 - manuálne s pochopením sekvencie.
- ▷ So spojením niektorých sekvencií vs. bez spojenia

Objavené vzory II.

▷ Najčastejšie netriviálne vzory

sekvencia udalostí	počet výskytov
aktualizácia → aktualizácia → aktualizácia	29 185
nová karta → aktivácia → aktualizácia	21 222
aktivácia → nová karta → aktivácia → aktualizácia	5 517
aktualizácia (8x)	5 263

Vybrané vzory a akcie I.

- ▷ Vzor hľadania medzi otvorenými kartami
 - implementované 4 verzie,
 - **sekvencia**: aktivácia (4x).
 - **podmienky**: aspoň 3 rôzne karty, ohraničené pohyblivým priemerom.
 - **akcia**: Zoradenie kariet podľa domény a subdomény.

- ▷ Vzor porovnávania obsahu dvoch kariet
 - **sekvencia**: aktivácia (4x)
 - **podmienky**: aspoň 4 aktivácie na 2 karty ohraničené pohyblivým priemerom.

Vybrané vzory a akcie II.

- ▷ Vzor sledovania zmien v konkrétnej karte
 - **sekvencia:** aktualizácia (3x)
 - **podmienky:** 3 aktualizácie konkrétnej karty na URL rovnakú ako je aktuálna, ohraničené pohyblivým priemerom.

- ▷ Vzor ukončenia aktuálnej úlohy v prehliadači
 - **sekvencia:** zatvorenie (4x)
 - **podmienky:** zatvorenie 4 kariet ohraničené pohyblivým priemerom.

Overenie

1. Korektnosť zozbieraných dát

- Sú nami zaznamenané akcie naozaj vykonané?
- Test so scenárom a očakávanými záznamami (ukážka).

2. Relevantnosť objavených vzorov

- Naozaj sa nami objavené vzory vyskytujú?
- Podpora používateľov a akceptovanie pri výskyte vzorov.

3. Presnosť odporúčania a podpora používateľov

- Je naša interpretácia týchto vzorov správna?
- Presnosť odporúčania (podiel akceptovaných odporúčení).

Výsledky odporúčania akcií

	Podpora používateľov	Prijatých	Zamietnutých	Podiel
Vzor hľadania (4 verzie)	49 %	83	530	15.66 %
Vzor porovnávania	51 %	209	608	25.58 %
Vzor sledovania zmien	34 %	39	94	29.32 %
Vzor ukončenia úlohy	34 %	74	52	58.73 %
Celkovo	<u>68 %</u>	<u>405</u>	<u>1284</u>	<u>31.54 %</u>

Zhodnotenie výsledkov

▷ Získané dáta

- dostatočné množstvo,
- korektnosť a vhodná štruktúra,
- dostatočný počet používateľov (avšak nedostatočná diverzita).

▷ Objavené vzory

- odpovedajú realite,
- vyjadrujú zaujímavé skutočnosti.

▷ Odporúčanie akcií

- na prvý pohľad negatívne výsledky,
- po zohľadnení možných dôvodov (nevedomosť o možnosti vrátenia stavu, atď.) poukazujú na potenciál v oblasti.

Zhrnutie

- ▷ Vytvorili sme platformu na získavanie **vhodných** dát o paralelnom prehliadaní.
- ▷ Objavili a analyzovali sme viaceré zaujímavé **vzory** vyskytujúce sa pri práci s kartami.
- ▷ Implementovali sme ukážku možného využitia týchto vzorov vo forme odporúčania akcií vo webovom prehliadači a tým poukázali na potenciál v oblasti.

Používateľská prezentácia

TabRec, ukážka, zhrnutie



TabRec

- ▷ Rozšírenie pre webový prehliadač Google Chrome.
- ▷ Funkcionalita
 - zaznamenávanie operácií nad kartami,
 - detekcia vybraných vzorov a odporúčanie akcií v reálnom čase,
 - poskytovanie základných štatistík o prehliadaní a odporúčaniach.



Demo ukážka - TabRec

Výhody a nevýhody riešenia

▷ Výhody

- jednoduchá inštalácia a aktualizácia,
- možnosť rozšírenia o ďalšie vzory a akcie,
- možnosť integrácie s inými službami (ALEF¹³).

▷ Nevýhody

- obmedzené možnosti v prispôsobení UI,
- potreba google účtu vyvoláva obavy zo súkromia.

¹³ Adaptive LEarning Framework, <http://alef.fiit.stuba.sk>

Zhrnutie

- ▷ Vytvorili sme platformu na získavanie **vhodných** dát o paralelnom prehliadaní.
- ▷ Objavili a analyzovali sme viaceré zaujímavé **vzory** vyskytujúce sa pri práci s kartami.
- ▷ Implementovali sme ukážku možného využitia týchto vzorov vo forme odporúčania akcií vo webovom prehliadači a tým poukázali na potenciál v oblasti.

Literatúra

- [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] Weinreich, H., Obendorf, H., Herder, E., Mayer, M.: Not quite the average: An empirical study of web use. ACM Trans. Web, 2(1): 5:1–5:31, March 2008.
- [3] Huang, J. and White, R.: Parallel browsing behavior on the web. Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10, page 13, 2010.
- [4] Bonnin, G., Brun, A., Boyer, A.: Towards tabbing aware recommendations. In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, IITM '10, pages 316–323, New York, NY, USA, 2010.
- [5] Agrawal, R., Imielinski, T. and Swami, A.: Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93, pages 207–216, New York, NY, USA, 1993.