Status Update

February 8, 2011

Web Performance Optimization: Analytics

Wim Leers

Web Performance Optimization

- Speed matters!
  - 0.1 s → direct manipulation
  - 1 s → good navigation
  - 10 s → attention kept
  - >10 s → *bye bye!*

Source: http://www.useit.com/alertbox/response-times.html, Jakob Nielsen, June 21, 2010

How to Measure? **Episodes**

- Measures “episodes” during page loading
- **Real measurements**: JS in browser, for *each* visitor
- Result: Episodes log file

Analytics

- **Automatically pinpoint causes of slow page loads**
- e.g.:
  - “http://uhasselt.be/ is slow in Belgium, for users of the ISP Telenet”
  - “http://uhasselt.be/studenten/dossier has slowly loading CSS”
  - “http://uhasselt.be/bib has slowly loading JS in Firefox 3”
  - …

Literature Study Subjects

- Data Stream Mining
- Anomaly Detection

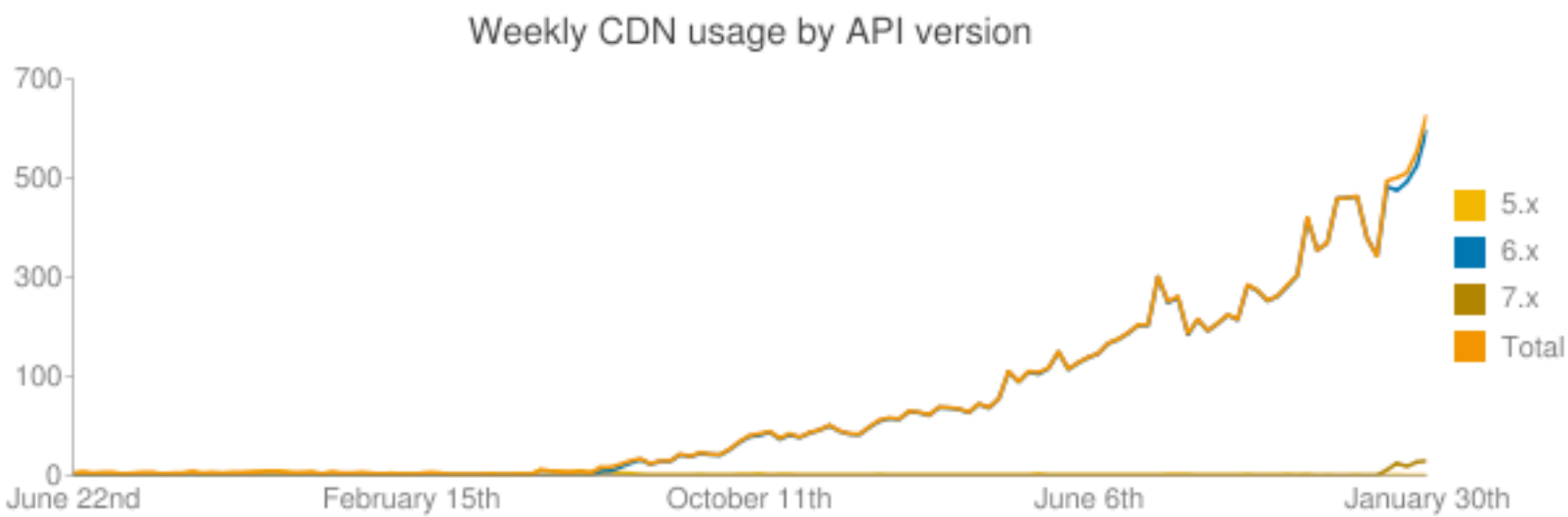
} **Data Mining:** finding patterns in data

- OLAP: Data Cube

} **OLAP:** querying multidimensional data

Work during: *the summer vacation*

- Keep track of evolutions in WPO
- Drupal CDN module updated
  - Goal: many users ⇒ easy-to-find master thesis testers
  - Currently: >600 websites use it, soon probably >1000



Work during: *October*

- Polishing of literature study based on feedback from prof. Wim Lamotte
- Finish literature study
  - Added section 7.9 (“Stream Cube: Data Cube for Data Streams”)
  - Re-added & updated the sections that weren’t sufficiently theoretical:
    - Section 4.2.1 (“All Fields Explained”)
    - Section 2 (“The Process”)

Work during: *November*

- `EpisodesParser`
  - Uses `QtConcurrent` to split up work in chunks, process them concurrently
- `QCachingLocale`
- *Performance Planet 2010 Advent Calendar* article:
  http://calendar.perfplanet.com/2010/wpo-analytics/
  (other authors include Google, Yahoo and Facebook employees, and most big names in the WPO industry)

Work during: *December*

- `QBrowsCap`
  - Uses the Browser Capabilities Project's dataset
- `QGeoIP`
  - Uses MaxMind.com's GeoIP databases + GeoIP C library
  - Not thread-safe, due to the GeoIP C library (spent a lot of time trying)

Work during: *January—February*

- `EpisodeDurationDiscretizer`
  - .csv file, like this:

```
pageready,fast,300,acceptable,2000,slow
domready,fast,150,acceptable,1000,slow
backend,fast,100,acceptable,500,slow
…
```

- `FPGrowth`
  - Optimization: set required item for transactions (`duration:slow`)
- `RuleMiner`
  - Optimization: set rule consequent (`duration:slow`)
- **5,000 lines of code** already! (Excluding the GeoIP C library.)

Sample flow

- **Step 1:** Episode log line:

```
"218.56.155.59 [Sunday, 14-Nov-2010 06:27:03 +0100] "?ets=css:203,headerjs:
94,footerjs:500,domready:843,tabs:110,ToThePointShowHideChangelog:
15,DrupalBehaviors:141,frontend:1547" 200 "http://driverpacks.net/driverpacks/
windows/xp/x86/chipset/10.09" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;
SV1)" "driverpacks.net""
```

- **Step 2:** Parsed + expanded (with concept hierarchy) into a transaction:

```
("episode:css", "duration:acceptable", "url:http://driverpacks.net/driverpacks/
windows/xp/x86/chipset/10.09", "status:200", "location:AS", "location:AS:China",
"location:AS:China:Shandong", "location:AS:China:Shandong:Zaozhuang",
"location:isp:China:AS4837 CNCGROUP China169 Backbone", "ua:WinXP", "ua:WinXP:IE",
"ua:WinXP:IE:6", "ua:WinXP:IE:6:0", "ua:IE", "ua:IE:6", "ua:IE:6:0",
"ua:isNotMobile")
```

Sample flow

- **Step 3:** Mined frequent itemsets per 4,000 page views (±40,000 transactions):

```
([size=1] {status:200}, [size=1] {ua:isNotMobile}, [size=2] {status:200,
ua:isNotMobile}, [size=1] {ua:WinXP}, [size=2] {status:200, ua:WinXP}, [size=2]
{ua:isNotMobile, ua:WinXP}, [size=3] {status:200, ua:isNotMobile, ua:WinXP}, [size=1]
{location:EU}, [size=2] {status:200, location:EU}, [size=2] {ua:isNotMobile,
location:EU}, [size=3] {status:200, ua:isNotMobile, location:EU}, [size=1]
{ua:Firefox}, [size=2] {status:200, ua:Firefox}, [size=2] {ua:isNotMobile,
ua:Firefox}, [size=3] {status:200, ua:isNotMobile, ua:Firefox}, [size=1] {ua:IE},
[size=2] {status:200, ua:IE}, [size=2] {ua:isNotMobile, ua:IE}, [size=3] {status:200,
ua:isNotMobile, ua:IE}, [size=2] {ua:WinXP, ua:IE}, [size=3] {status:200, ua:WinXP,
ua:IE}, [size=3] {ua:isNotMobile, ua:WinXP, ua:IE}, [size=4] {status:200,
ua:isNotMobile, ua:WinXP, ua:IE}, [size=1] {ua:Firefox:3}, [size=2] {status:200,
ua:Firefox:3}, [size=2] {ua:isNotMobile, ua:Firefox:3}, [size=3] {status:200,
ua:isNotMobile, ua:Firefox:3}, [size=2] {ua:Firefox, ua:Firefox:3}, [size=3] {status:
200, ua:Firefox, ua:Firefox:3}, [size=3] {ua:isNotMobile, ua:Firefox, ua:Firefox:3},
[size=4] {status:200, ua:isNotMobile, ua:Firefox, ua:Firefox:3}, [size=1]
{location:AS}, [size=2] {status:200, location:AS}, [size=2] {ua:isNotMobile,
location:AS}, [size=3] {status:200, ua:isNotMobile, location:AS} … )
```

Sample flow

- **Step 4:** Mined association rules from these frequent itemsets:

```
({status:200, ua:isNotMobile} => {duration:slow} (conf=1), {status:200, ua:WinXP} =>
{duration:slow} (conf=1), {ua:isNotMobile, ua:WinXP} => {duration:slow} (conf=1),
{status:200, ua:isNotMobile, ua:WinXP} => {duration:slow} (conf=1), {status:200,
location:EU} => {duration:slow} (conf=1), {ua:isNotMobile, location:EU} =>
{duration:slow} (conf=1), {status:200, ua:isNotMobile, location:EU} => {duration:slow}
(conf=1), {status:200, ua:Firefox} => {duration:slow} (conf=1), {ua:isNotMobile,
ua:Firefox} => {duration:slow} (conf=1), {status:200, ua:isNotMobile, ua:Firefox} =>
{duration:slow} (conf=1), {status:200, ua:IE} => {duration:slow} (conf=1),
{ua:isNotMobile, ua:IE} => {duration:slow} (conf=1), {status:200, ua:isNotMobile,
ua:IE} => {duration:slow} (conf=1), {ua:WinXP, ua:IE} => {duration:slow} (conf=1),
{status:200, ua:WinXP, ua:IE} => {duration:slow} (conf=1), {ua:isNotMobile, ua:WinXP,
ua:IE} => {duration:slow} (conf=1), {status:200, ua:isNotMobile, ua:WinXP, ua:IE} =>
{duration:slow} (conf=1), {status:200, ua:Firefox:3} => {duration:slow} (conf=1),
{ua:isNotMobile, ua:Firefox:3} => {duration:slow} (conf=1), {status:200,
ua:isNotMobile, ua:Firefox:3} => {duration:slow} (conf=1), {ua:Firefox, ua:Firefox:3}
=> {duration:slow} (conf=1), {status:200, ua:Firefox, ua:Firefox:3} => {duration:slow}
(conf=1), {ua:isNotMobile, ua:Firefox, ua:Firefox:3} => {duration:slow} (conf=1) …)
```

Current issues

- Mining **meaningful** rules
  - e.g. NOT: `{ua:IE, ua:WinXP} => {duration: slow}`
  - better: `{episode:pageReady, ua:IE, ua:WinXP} => {duration:slow}`
- **Optimizing** the current `FPGrowth` & `RuleMiner` logic to get **meaningful rules**
  ⇒ take these changes into account when implementing FP-stream
  - Optimization: set a required rule antecedent item (`episode:*`): rules should always be about *slow* (`duration:slow`) *episodes* (`episode:*`)
- **Concept hierarchy** filtering is not yet implemented:
  `{ua:IE, ua:IE8, ua:IE8:0} => {duration:slow}`

Performance characteristics

- Current performance characteristics for the sample flow:
  - Episodes log file of ±50,000 page views (±500,000 episodes): ±25 s
    - That's >2,000 page views analyzed per second
    - Or >20,000 episodes (transactions) analyzed per second
  - If we'd analyze a live site's data stream of up to 1,200 pageviews/s, that's **sufficient for websites with more than 100 million pageviews per day (or 3 billion pageviews per month)**

    ⇒ sufficient for >99% of all websites

  - *But*, it is possible that performance will get better or worse with FP-Stream.

Future

- Work should begin on FP-Stream in about 1 week
- After that, in order:
  - Basic UI (for conclusions (association rules) found)
  - OLAP + integrate this with UI
  - Advanced UI: visualizations (e.g. charts)
  - Anomaly detection (if there's still time) — not required for a useful application