Web Performance Optimization: Analytics

Wim Leers

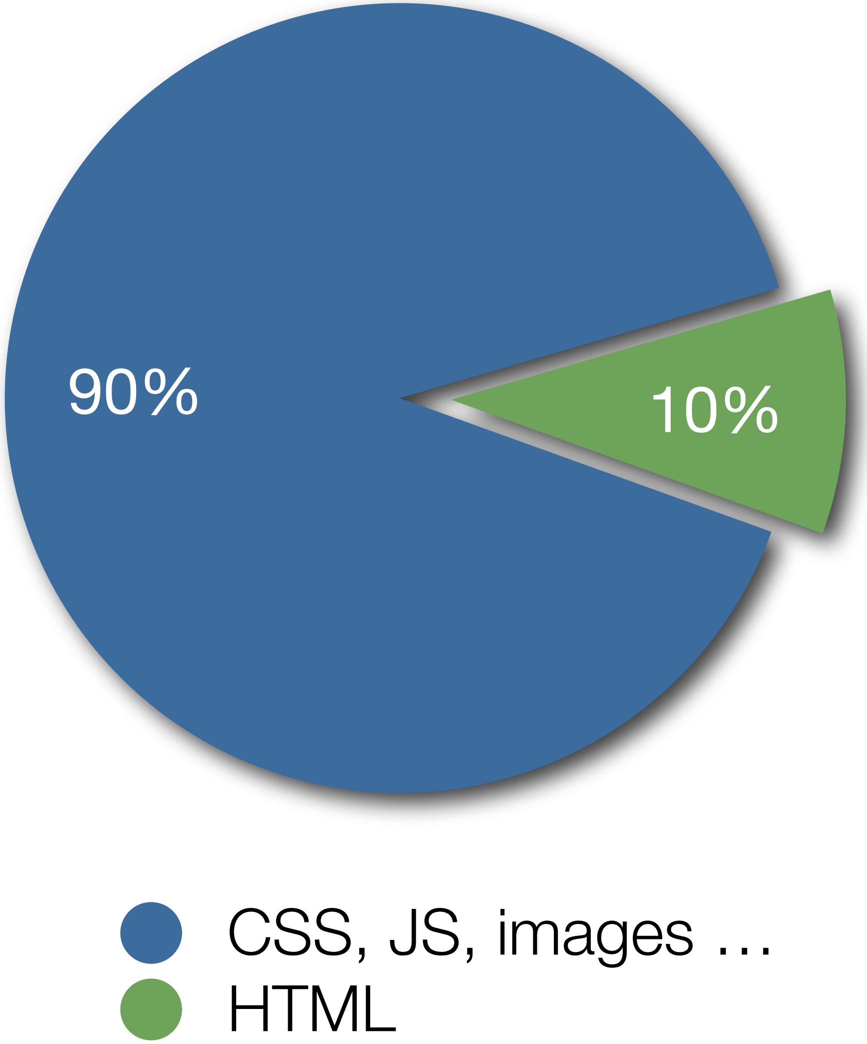
Promotor: Prof. dr. Jan Van den Bussche

Why Optimize? **Speed matters**

- Speed → satisfaction → more & happier visitors
- Search engines reward speed → more visitors
- Examples
  - Google: +0.5s → -20% searches
  - Amazon: +0.1s → -1% sales

Source: http://www.slideshare.net/stubbornella/designing-fast-websites-presentation, Nicole Sullivan, Yahoo!

What to Optimize? **Front-end**



How to Measure? **Episodes**

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

What to Optimize *Exactly?* **WPO 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”
  - …

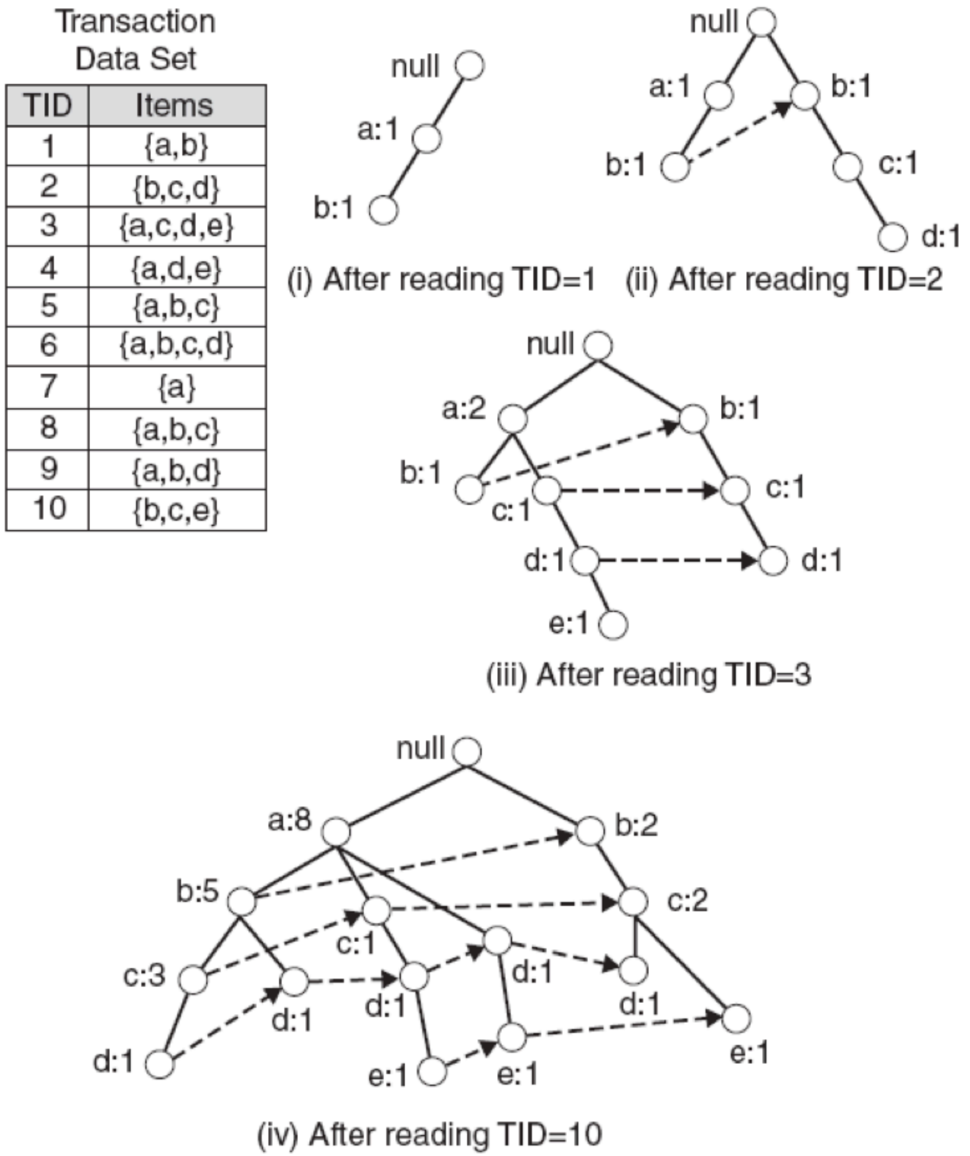
The Theory: **Data Stream Mining**

- Data mining: **finding patterns in data**
- Implemented well-known algorithms:
    - **FP-Growth**: mining frequent patterns from **static data sets**
    - **FP-Stream:** mining frequent patterns from **data streams**
        - Possibly infinite data streams ⇒ approximation necessary
    - **Apriori:** mining **association rules** from frequent itemsets

FP-Growth: **FP-Tree**

Prefix tree or *Trie*

- Efficiently store transactions
- Maximize compression by ordering items in the transaction by descending frequency



Transaction Data Set

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

(i) After reading TID=1

(ii) After reading TID=2

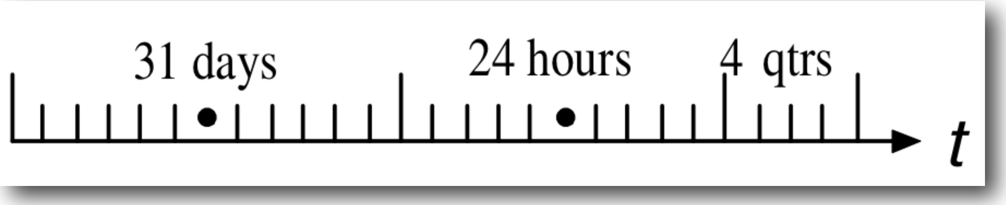
(iii) After reading TID=3

(iv) After reading TID=10

Source: Introduction to Data Mining, Nan; Steinbach; Kumar, 2005

FP-Stream: **Tilted-Time Window Model**

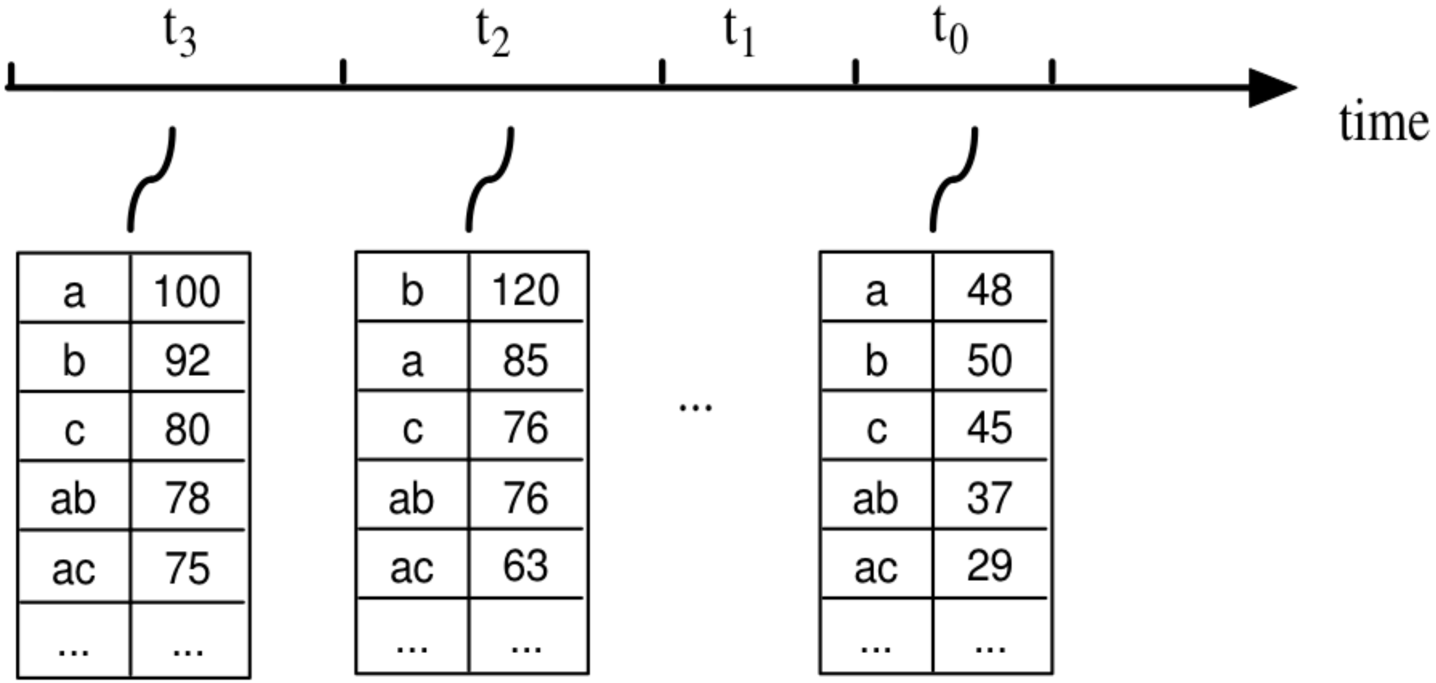
The more recent, the more detail.



Source: Mining Frequent Patterns in Data Streams at Multiple Time Granularities, Giannella; Han et al., 2003

FP-Stream: **Frequent Patterns in** `TiltedTimeWindow`

- Suppose: $\{t_0, t_1, t_2, t_3\}$ are all full; next window w_n arrives
- Result: reset $\{t_3\}$; $t_3 = t_2$; $t_2 = t_1 + t_0$; reset $\{t_1, t_0\}$; $t_0 = w_n$



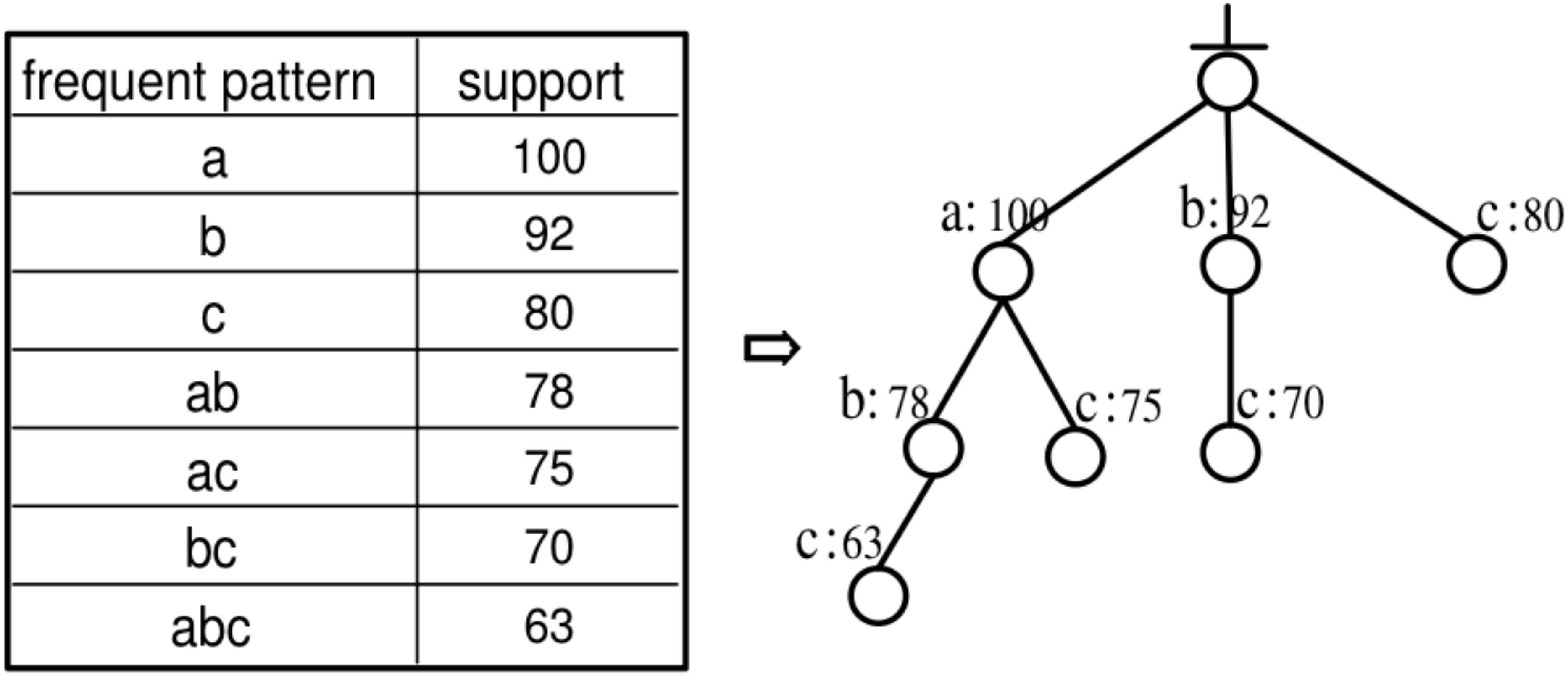
a	100
b	92
c	80
ab	78
ac	75
...	...

b	120
a	85
c	76
ab	76
ac	63
...	...

a	48
b	50
c	45
ab	37
ac	29
...	...

Source: Mining Frequent Patterns in Data Streams at Multiple Time Granularities, Giannella; Han et al., 2003

FP-Stream: **PatternTree**



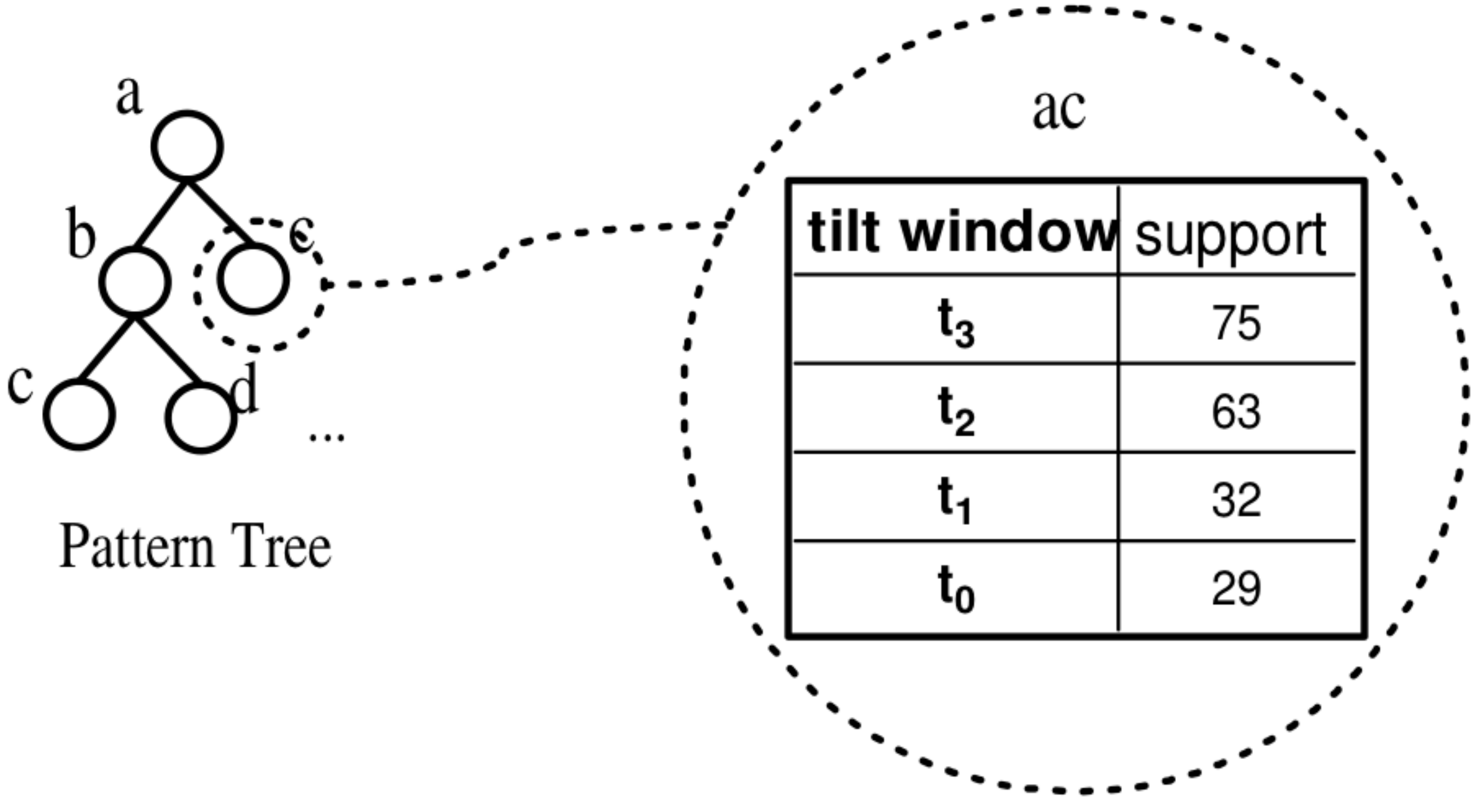
frequent pattern	support
a	100
b	92
c	80
ab	78
ac	75
bc	70
abc	63

Frequent Patterns

Pattern Tree

Source: Mining Frequent Patterns in Data Streams at Multiple Time Granularities, Giannella; Han et al., 2003

FP-Stream: **PatternTree**



tilt window	support
t_3	75
t_2	63
t_1	32
t_0	29

Pattern Tree

Tilted-time Window Table

Source: Mining Frequent Patterns in Data Streams at Multiple Time Granularities, Giannella; Han et al., 2003

Architecture

- 3 modules (connected through Qt’s signal/slot mechanism: low coupling)
  - `EpisodesParser`: log file → transactions (episodes)
  - `Analytics`
    - <u>Processing:</u> episodes → `PatternTree`
    - <u>Upon request:</u> `PatternTree` → frequent patterns → association rules
  - `UI`
- ±9,000 lines of C++/Qt

Implementing `EpisodesParser`

- New libraries
  - `QCachingLocale`: speed up locale queries
  - `QBrowsCap`: user agent → operating system + browser
  - `QGeoIP`: IP → location + ISP

Implementing `Analytics`

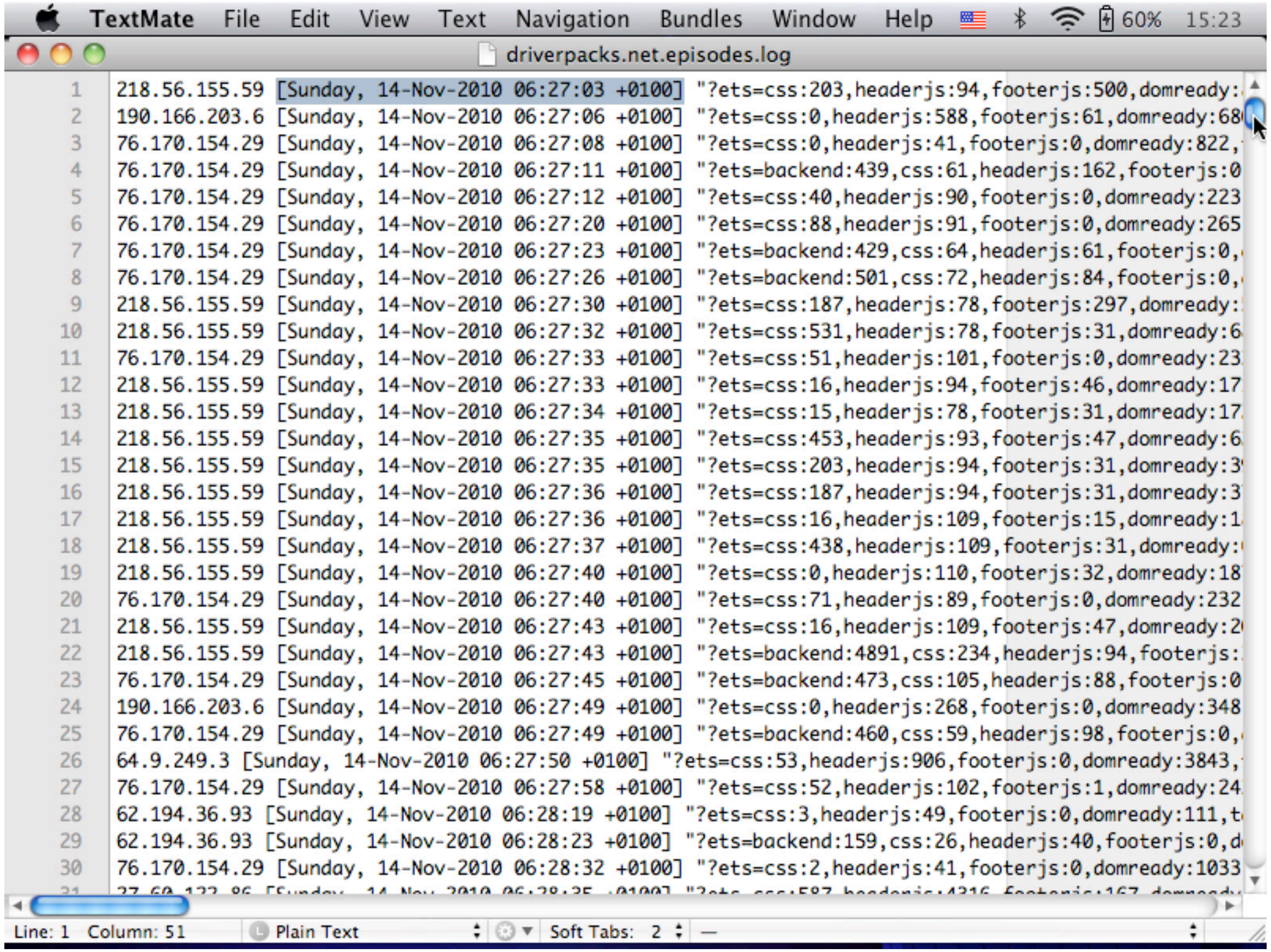
- Phase 1: frequent itemset mining on **static data sets → FP-Growth**
  - Phase 1b: **optimize** FP-Growth
  - Phase 1c: **Apriori** to mine association rules
- Phase 2: **FP-Growth + item constraints** (not covered by literature)
- Phase 3: frequent itemset mining on **data streams → FP-Stream**
- Phase 4: **FP-Stream + item constraints** (not covered by literature)

Note: FP-Stream uses FP-Growth!

Implementing `UI`

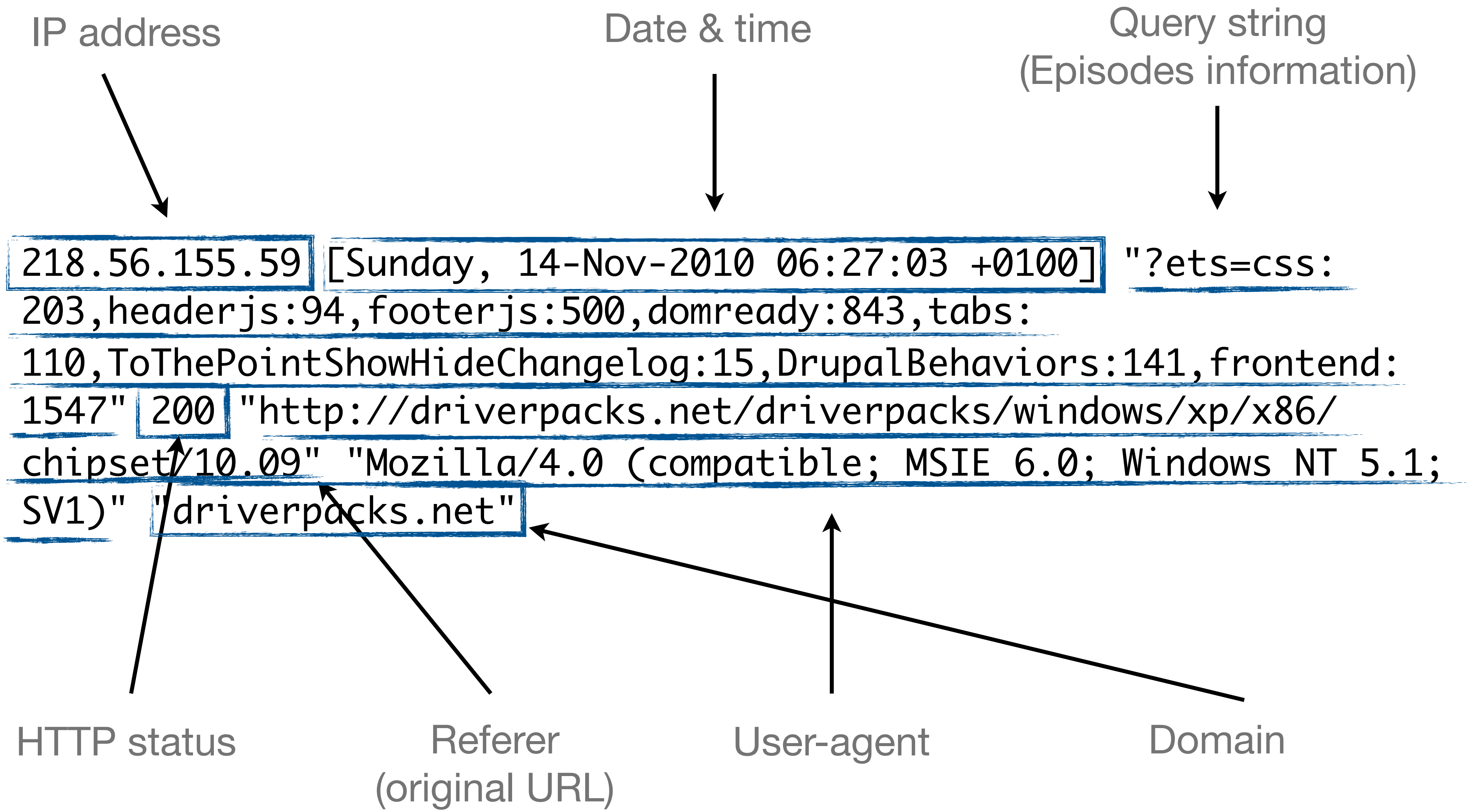
Not interesting.

Sample Flow: **Episodes Log File**



```
218.56.155.59 [Sunday, 14-Nov-2010 06:27:03 +0100] "?ets=css:203,headerjs:94,footerjs:500,domready:
190.166.203.6 [Sunday, 14-Nov-2010 06:27:06 +0100] "?ets=css:0,headerjs:588,footerjs:61,domready:68
76.170.154.29 [Sunday, 14-Nov-2010 06:27:08 +0100] "?ets=css:0,headerjs:41,footerjs:0,domready:822,
76.170.154.29 [Sunday, 14-Nov-2010 06:27:11 +0100] "?ets=backend:439,css:61,headerjs:162,footerjs:0
76.170.154.29 [Sunday, 14-Nov-2010 06:27:12 +0100] "?ets=css:40,headerjs:90,footerjs:0,domready:223
76.170.154.29 [Sunday, 14-Nov-2010 06:27:20 +0100] "?ets=css:88,headerjs:91,footerjs:0,domready:265
76.170.154.29 [Sunday, 14-Nov-2010 06:27:23 +0100] "?ets=backend:429,css:64,headerjs:61,footerjs:0,
76.170.154.29 [Sunday, 14-Nov-2010 06:27:26 +0100] "?ets=backend:501,css:72,headerjs:84,footerjs:0,
218.56.155.59 [Sunday, 14-Nov-2010 06:27:30 +0100] "?ets=css:187,headerjs:78,footerjs:297,domready:
218.56.155.59 [Sunday, 14-Nov-2010 06:27:32 +0100] "?ets=css:531,headerjs:78,footerjs:31,domready:6
76.170.154.29 [Sunday, 14-Nov-2010 06:27:33 +0100] "?ets=css:51,headerjs:101,footerjs:0,domready:23
218.56.155.59 [Sunday, 14-Nov-2010 06:27:33 +0100] "?ets=css:16,headerjs:94,footerjs:46,domready:17
218.56.155.59 [Sunday, 14-Nov-2010 06:27:34 +0100] "?ets=css:15,headerjs:78,footerjs:31,domready:17
218.56.155.59 [Sunday, 14-Nov-2010 06:27:35 +0100] "?ets=css:453,headerjs:93,footerjs:47,domready:6
218.56.155.59 [Sunday, 14-Nov-2010 06:27:35 +0100] "?ets=css:203,headerjs:94,footerjs:31,domready:3
218.56.155.59 [Sunday, 14-Nov-2010 06:27:36 +0100] "?ets=css:187,headerjs:94,footerjs:31,domready:3
218.56.155.59 [Sunday, 14-Nov-2010 06:27:36 +0100] "?ets=css:16,headerjs:109,footerjs:15,domready:1
218.56.155.59 [Sunday, 14-Nov-2010 06:27:37 +0100] "?ets=css:438,headerjs:109,footerjs:31,domready:
218.56.155.59 [Sunday, 14-Nov-2010 06:27:40 +0100] "?ets=css:0,headerjs:110,footerjs:32,domready:18
76.170.154.29 [Sunday, 14-Nov-2010 06:27:40 +0100] "?ets=css:71,headerjs:89,footerjs:0,domready:232
218.56.155.59 [Sunday, 14-Nov-2010 06:27:43 +0100] "?ets=css:16,headerjs:109,footerjs:47,domready:2
218.56.155.59 [Sunday, 14-Nov-2010 06:27:43 +0100] "?ets=backend:4891,css:234,headerjs:94,footerjs:
76.170.154.29 [Sunday, 14-Nov-2010 06:27:45 +0100] "?ets=backend:473,css:105,headerjs:88,footerjs:0
190.166.203.6 [Sunday, 14-Nov-2010 06:27:49 +0100] "?ets=css:0,headerjs:268,footerjs:0,domready:348
76.170.154.29 [Sunday, 14-Nov-2010 06:27:49 +0100] "?ets=backend:460,css:59,headerjs:98,footerjs:0,
64.9.249.3 [Sunday, 14-Nov-2010 06:27:50 +0100] "?ets=css:53,headerjs:906,footerjs:0,domready:3843,
76.170.154.29 [Sunday, 14-Nov-2010 06:27:58 +0100] "?ets=css:52,headerjs:102,footerjs:1,domready:24
62.194.36.93 [Sunday, 14-Nov-2010 06:28:19 +0100] "?ets=css:3,headerjs:49,footerjs:0,domready:111,t
62.194.36.93 [Sunday, 14-Nov-2010 06:28:23 +0100] "?ets=backend:159,css:26,headerjs:40,footerjs:0,d
76.170.154.29 [Sunday, 14-Nov-2010 06:28:32 +0100] "?ets=css:2,headerjs:41,footerjs:0,domready:1033
37.60.133.86 [Sunday, 14-Nov-2010 06:28:35 +0100] "?ets=css:587,headerjs:4316,footerjs:167,domready:
```

Sample Flow: **Episodes Log Line**



IP address

Date & time

Query string
(Episodes information)

```
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"
```

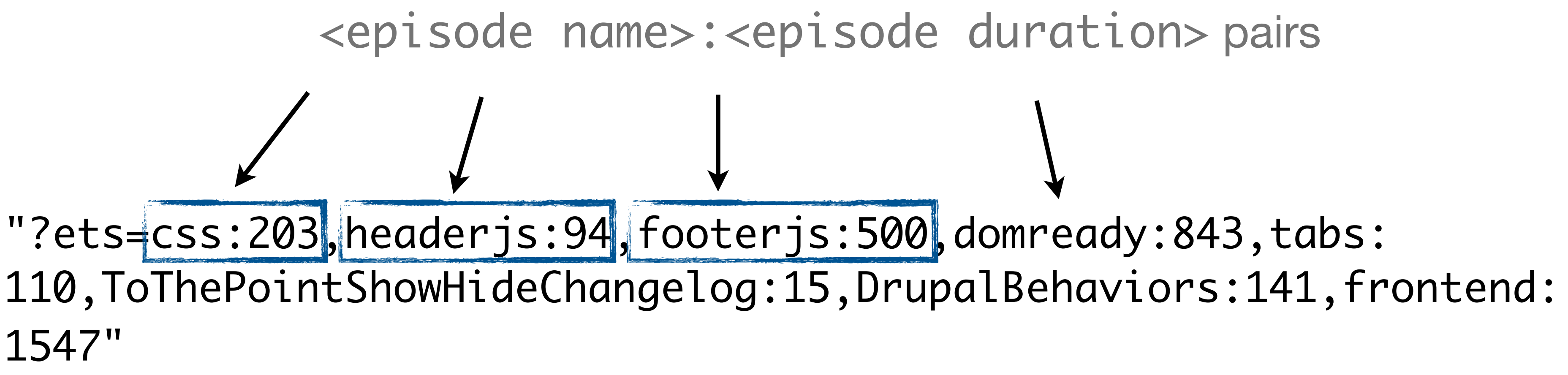
HTTP status

Referer
(original URL)

User-agent

Domain

Sample Flow: **Episodes Information**



`<episode name>:<episode duration>` pairs

```
"?ets=css:203,headerjs:94,footerjs:500,domready:843,tabs:
110,ToThePointShowHideChangelog:15,DrupalBehaviors:141,frontend:
1547"
```

(one for each episode in the page load)

Sample Flow: **Episodes Log Line → Transactions**

```
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"
```

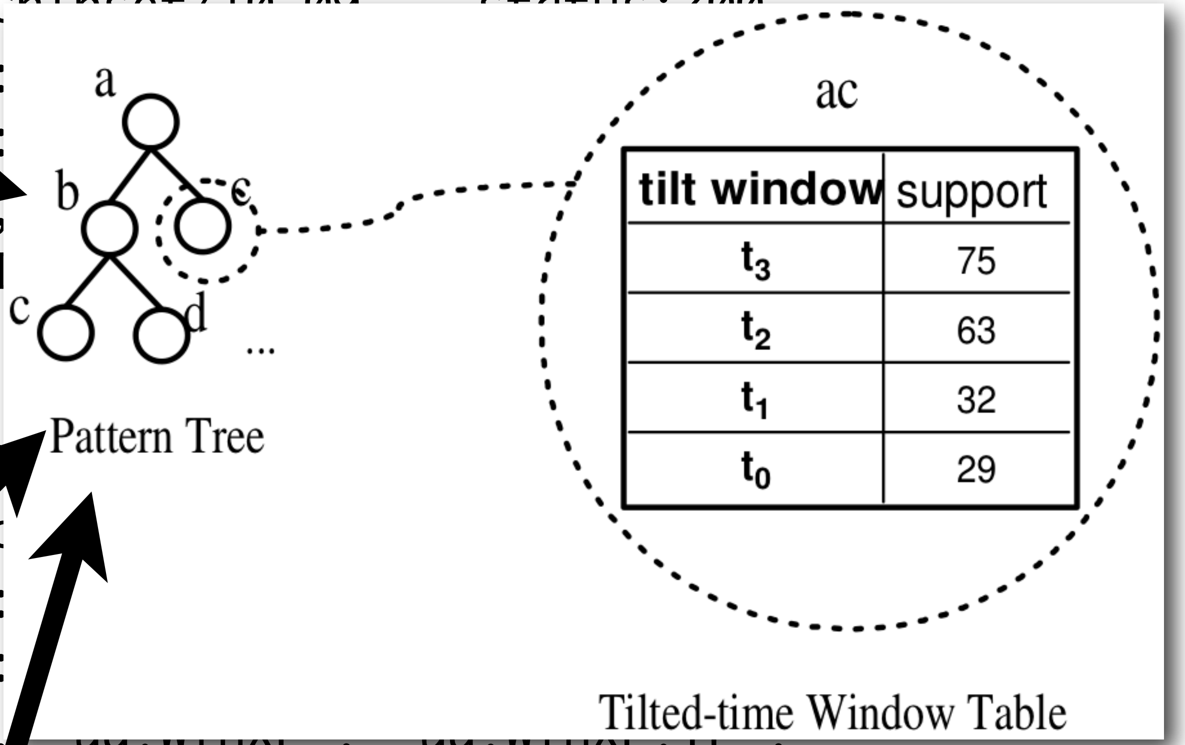
1 transaction per episode

```
("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")
```

```
("episode:headerjs", "duration:fast", "url:http://driverpacks.net/
driverpacks/windows/xp/x86/chipset/10.09", "status:200",
"location:AS", "location:AS:China", "location:AS:China:Shandong",
```

Sample Flow: **Transactions → PatternTree**

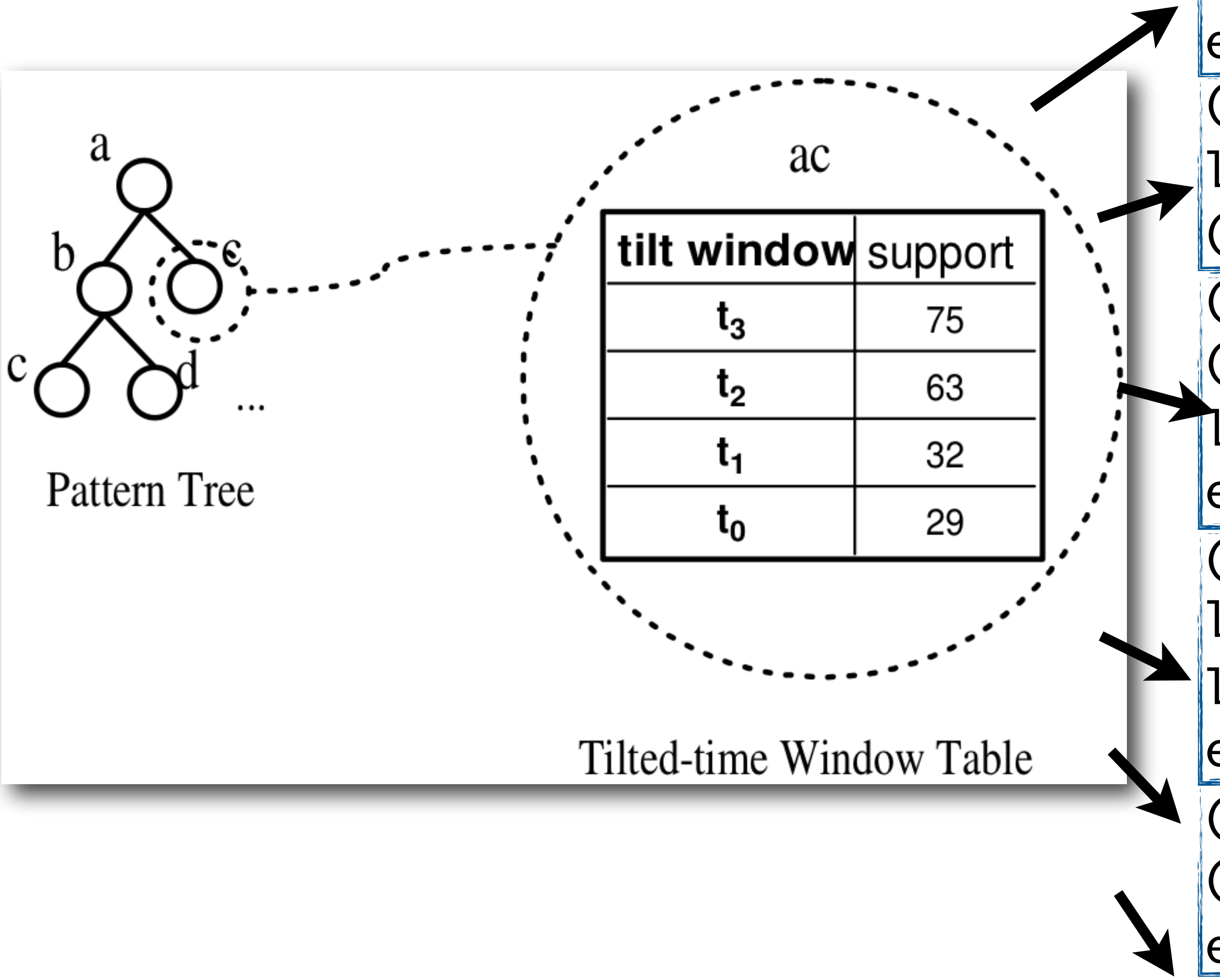
```
("episode:css", "duration:acceptable", "url:http://driverpacks.net/
driverpacks/windows/xp/x86/chipset/10.09", "status:200",
"location:AS", "location:AS:
"location:AS:China:Shandong:
CNCGROUP China169 Backbone",
"ua:WinXP:IE:6", "ua:WinXP:I
6:0", "ua:isNotMobile")

("episode:headerjs", "durat
driverpacks/windows/xp/x86/c
"location:AS", "location:AS:
"location:AS:China:Shandong:
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")

("episode:footerjs", "duration:acceptable", "url:http://
driverpacks.net/driverpacks/windows/xp/x86/chipset/10.09", "status:
```



tilt window	support
t_3	75
t_2	63
t_1	32
t_0	29

Tilted-time Window Table

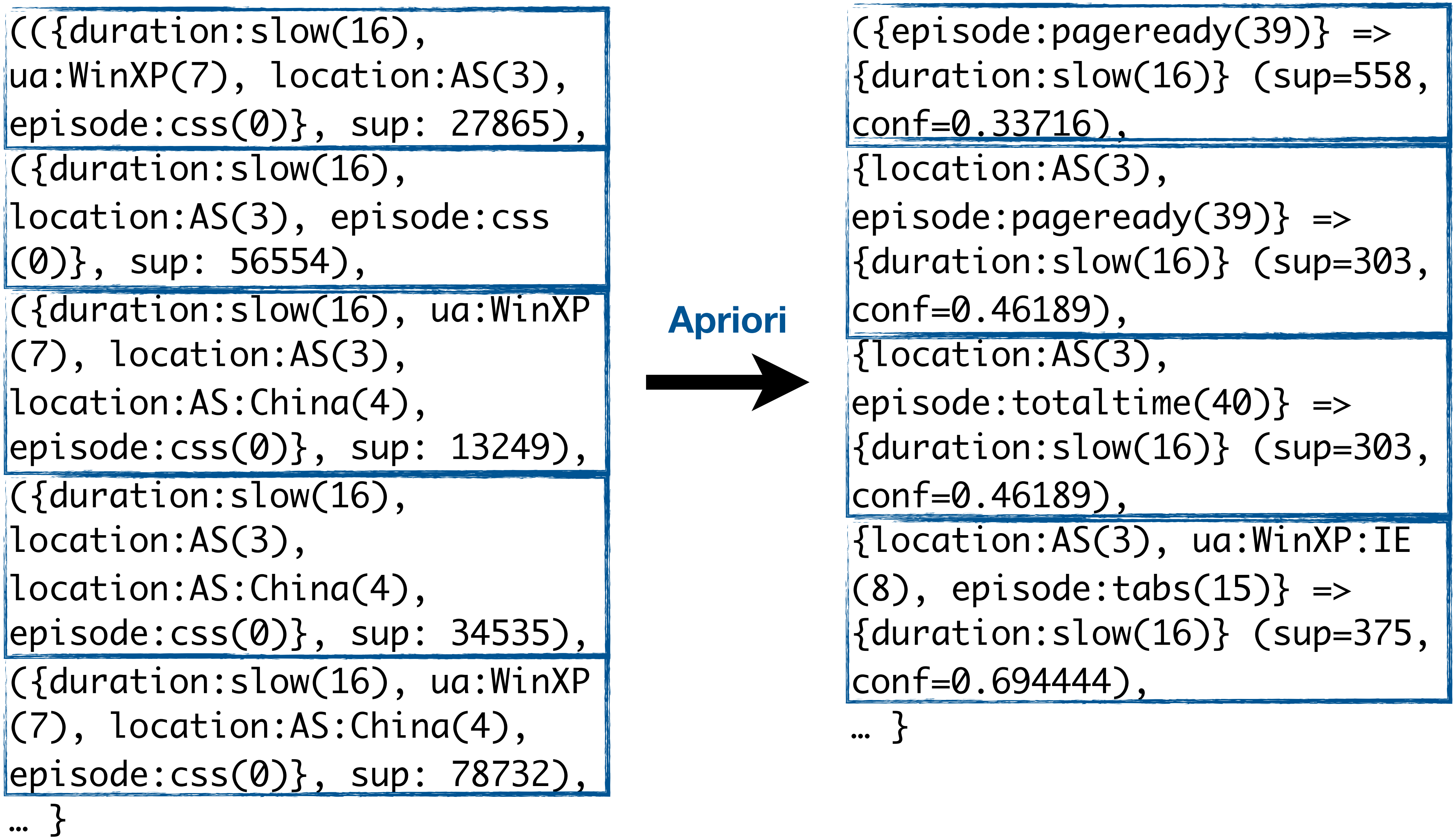
Sample flow: **PatternTree → Frequent Patterns**



tilt window	support
t_3	75
t_2	63
t_1	32
t_0	29

```
(({duration:slow(16),
ua:WinXP(7), location:AS(3),
episode:css(0)}, sup: 27865),
({duration:slow(16),
location:AS(3), episode:css
(0)}, sup: 56554),
({duration:slow(16), ua:WinXP
(7), location:AS(3),
location:AS:China(4),
episode:css(0)}, sup: 13249),
({duration:slow(16),
location:AS(3),
location:AS:China(4),
episode:css(0)}, sup: 34535),
({duration:slow(16), ua:WinXP
(7), location:AS:China(4),
episode:css(0)}, sup: 78732),
… }
```

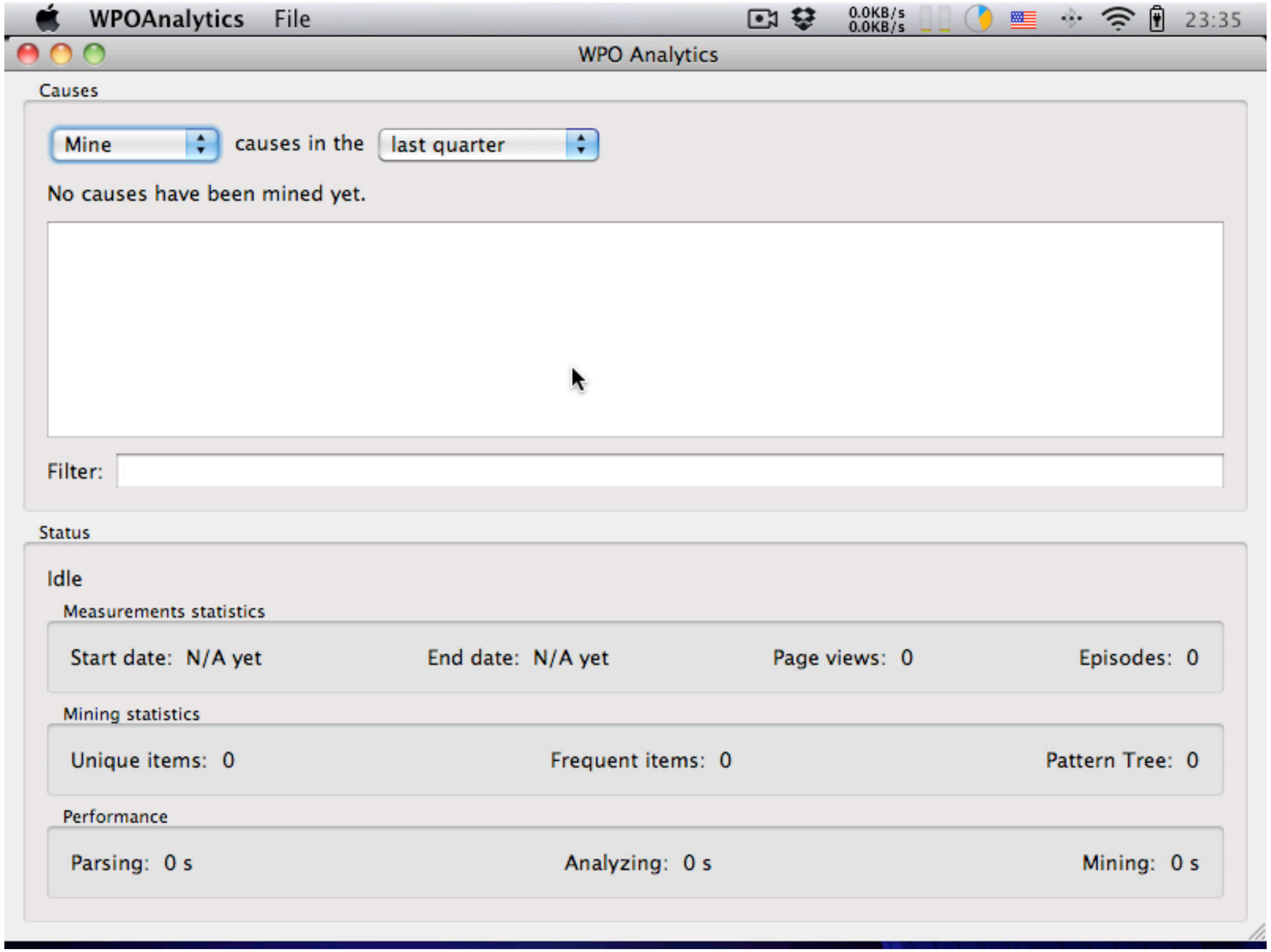
Sample Flow: **Frequent Patterns → Association Rules**



```
(({duration:slow(16),
ua:WinXP(7), location:AS(3),
episode:css(0)}, sup: 27865),
({duration:slow(16),
location:AS(3), episode:css
(0)}, sup: 56554),
({duration:slow(16), ua:WinXP
(7), location:AS(3),
location:AS:China(4),
episode:css(0)}, sup: 13249),
({duration:slow(16),
location:AS(3),
location:AS:China(4),
episode:css(0)}, sup: 34535),
({duration:slow(16), ua:WinXP
(7), location:AS:China(4),
episode:css(0)}, sup: 78732),
… }
```

```
({episode:pageready(39)} =>
{duration:slow(16)} (sup=558,
conf=0.33716),
{location:AS(3),
episode:pageready(39)} =>
{duration:slow(16)} (sup=303,
conf=0.46189),
{location:AS(3),
episode:totaltime(40)} =>
{duration:slow(16)} (sup=303,
conf=0.46189),
{location:AS(3), ua:WinXP:IE
(8), episode:tabs(15)} =>
{duration:slow(16)} (sup=375,
conf=0.694444),
… }
```

WPO Analytics: **Demo**



Causes

causes in the

No causes have been mined yet.

Filter:

Status

Idle

Measurements statistics

Start date: N/A yet
End date: N/A yet
Page views: 0
Episodes: 0

Mining statistics

Unique items: 0
Frequent items: 0
Pattern Tree: 0

Performance

Parsing: 0 s
Analyzing: 0 s
Mining: 0 s

Performance & Applicability

- On a 2.66 GHzCore 2 Duo:
  - Parser: >4,000 lines (page views)/s
  - FP-Stream: >12,000 episodes/s

    (FP-Growth: >16,500 episodes/s, but FP-Stream has some overhead)
- Assume:
  - 10 episodes per tracked page load
  - 1,200 lines (page views)/s

  } ⇒ 12,000 Episodes/s can be achieved
- Analyzing a live site's data stream of up to 1,200 pageviews/s makes this tool usable for **websites with more than 100 million pageviews per day (or 3 billion pageviews per month)**

  ⇒ sufficient for >99% of all websites!

Questions?

Thanks for your time!