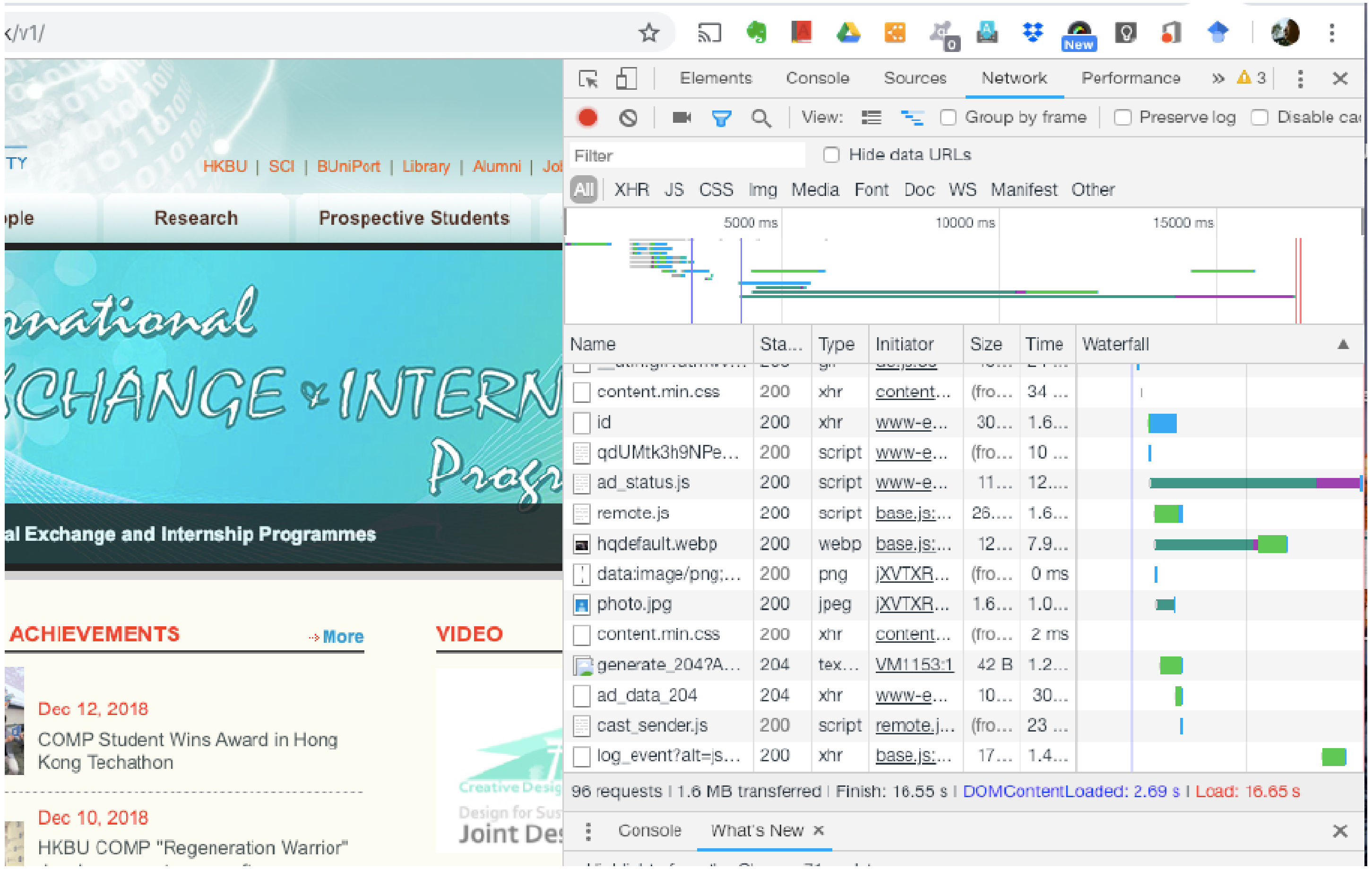
COMP7630 – Web Intelligence and its Applications

Web Usage Mining

Valentino Santucci

(valentino.santucci@unistrapg.it)

Downloading "one" web page …



Dec 12, 2018

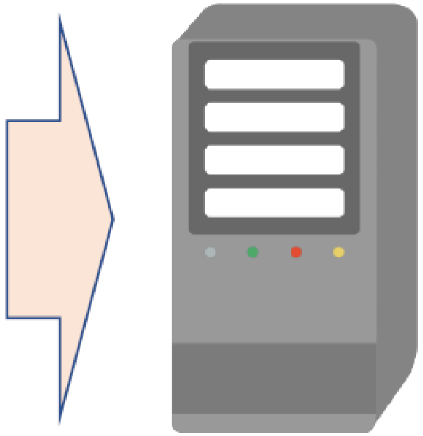
COMP Student Wins Award in Hong Kong Techathon

Dec 10, 2018

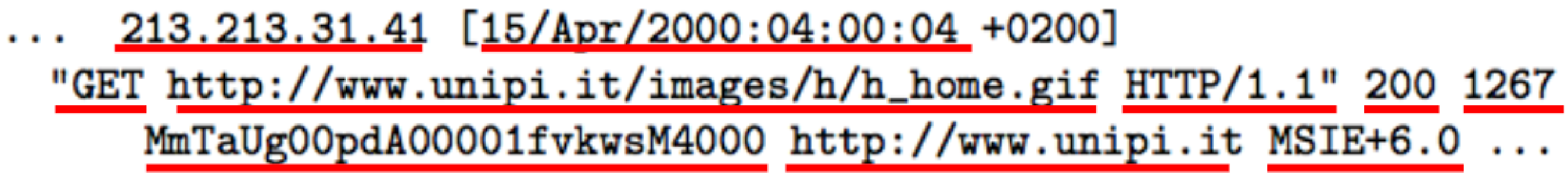
HKBU COMP "Regeneration Warrior"

Name	Sta...	Type	Initiator	Size	Time	Waterfall
content.min.css	200	xhr	content...	(fro...	34 ...	
id	200	xhr	www-e...	30...	1.6...	
qdUMtk3h9NPe...	200	script	www-e...	(fro...	10 ...	
ad_status.js	200	script	www-e...	11...	12....	
remote.js	200	script	base.js:...	26....	1.6...	
hqdefault.webp	200	webp	base.js:...	12...	7.9...	
data:image/png;...	200	png	jXVTXR...	(fro...	0 ms	
photo.jpg	200	jpeg	jXVTXR...	1.6...	1.0...	
content.min.css	200	xhr	content...	(fro...	2 ms	
generate_204?A...	204	tex...	VM1153:1	42 B	1.2...	
ad_data_204	204	xhr	www-e...	10...	30...	
cast_sender.js	200	script	remote.j...	(fro...	23 ...	
log_event?alt=js...	200	xhr	base.js:...	17...	1.4...	

96 requests | 1.6 MB transferred | Finish: 16.55 s | DOMContentLoaded: 2.69 s | Load: 16.65 s



Web Server Log



```
...  213.213.31.41 [15/Apr/2000:04:00:04 +0200]
 "GET http://www.unipi.it/images/h/h_home.gif HTTP/1.1" 200 1267
     MmTaUg00pdA00001fvkwsM4000 http://www.unipi.it MSIE+6.0 ...
```

- 213.213.31.41 is the IP address of the client
- 15/Apr/2000:04:00:04 is the date/time of this transaction (user activity)
- GET is the method of the transaction
- http://www.unipi.it/images/h/h home.gif is the URL requested
- HTTP/1.1 is the HTTP protocol
- 200 is the HTTP return code (200 means OK),
- 1267 is the size in bytes of the response sent to the client
- MmTaUg00pdA00001fvkwsM4000 is the *cookie* at client
- http://www.unipi.it is the URL from which the request was referred.
- MSIE+6.0 is the client environment provided by the client browser.

Web Usage Mining

- **Web usage mining**: automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web sites.
- **Goal**: analyze the behavioral patterns and profiles of users interacting with a Web site.
- The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common interests.

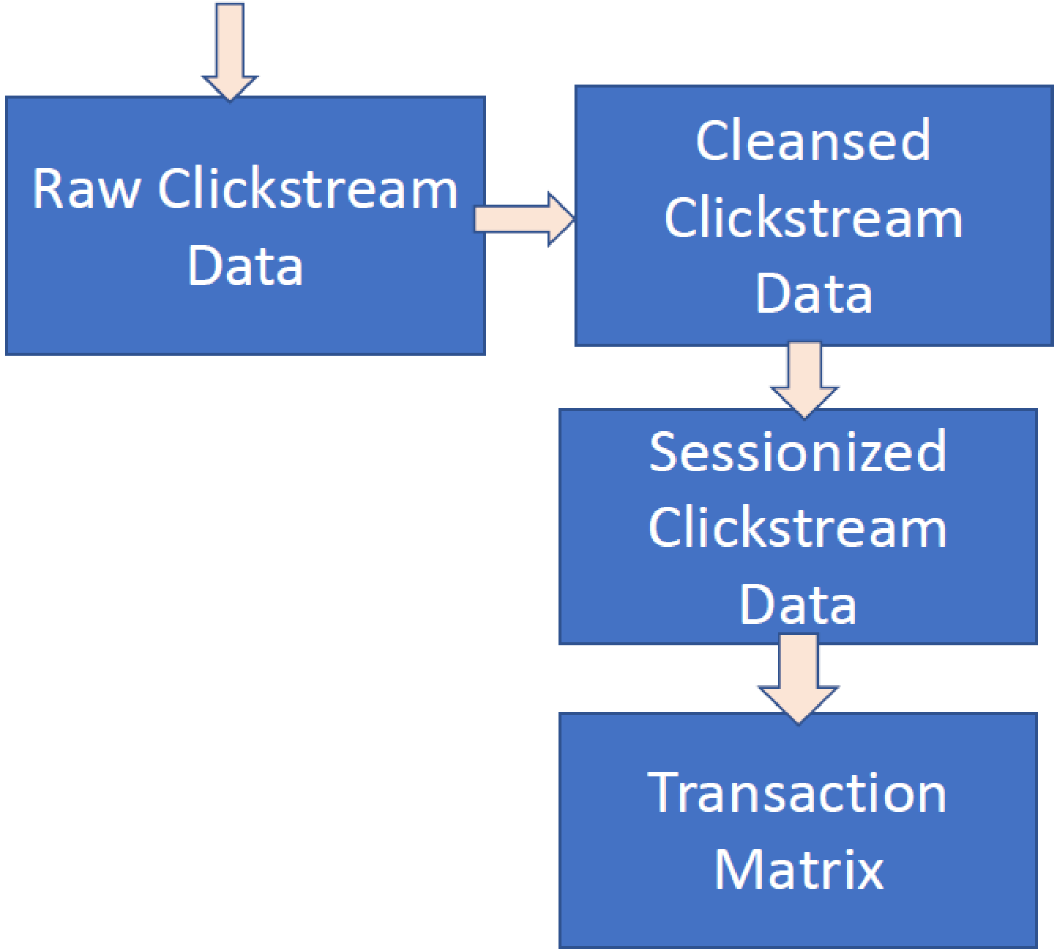
Preprocess & Analyze

- Collect and pre-process clickstreams

- Analyze clickstreams

Clickstream Data Preprocessing

```
...  213.213.31.41 [15/Apr/2000:04:00:04 +0200]
  "GET http://www.unipi.it/images/h/h_home.gif HTTP/1.1" 200 1267
      MmTaUg00pdA00001fvkwsM4000 http://www.unipi.it MSIE+6.0 ...
```

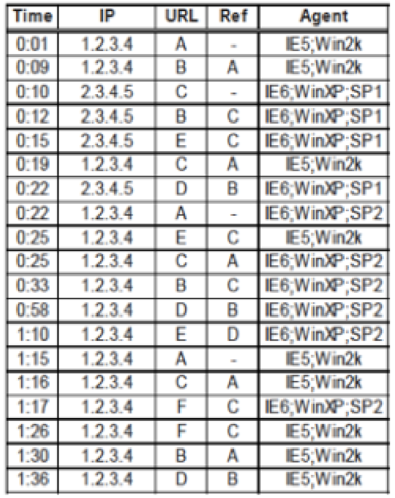


Raw Clickstream Data

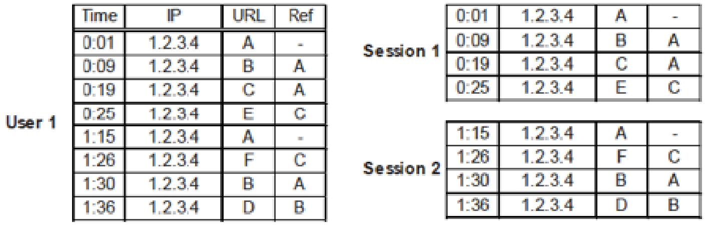
Cleansed Clickstream Data

Sessionized Clickstream Data

Transaction Matrix



Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k



User 1

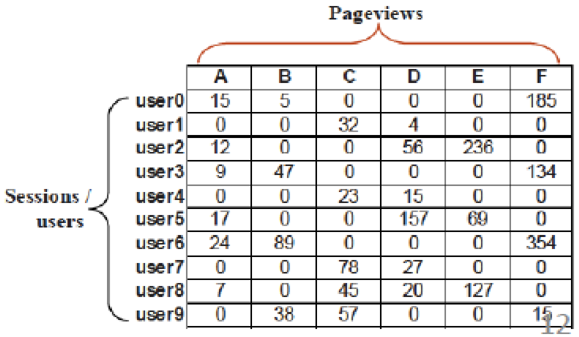
Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B



Pageviews

Sessions / users

	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

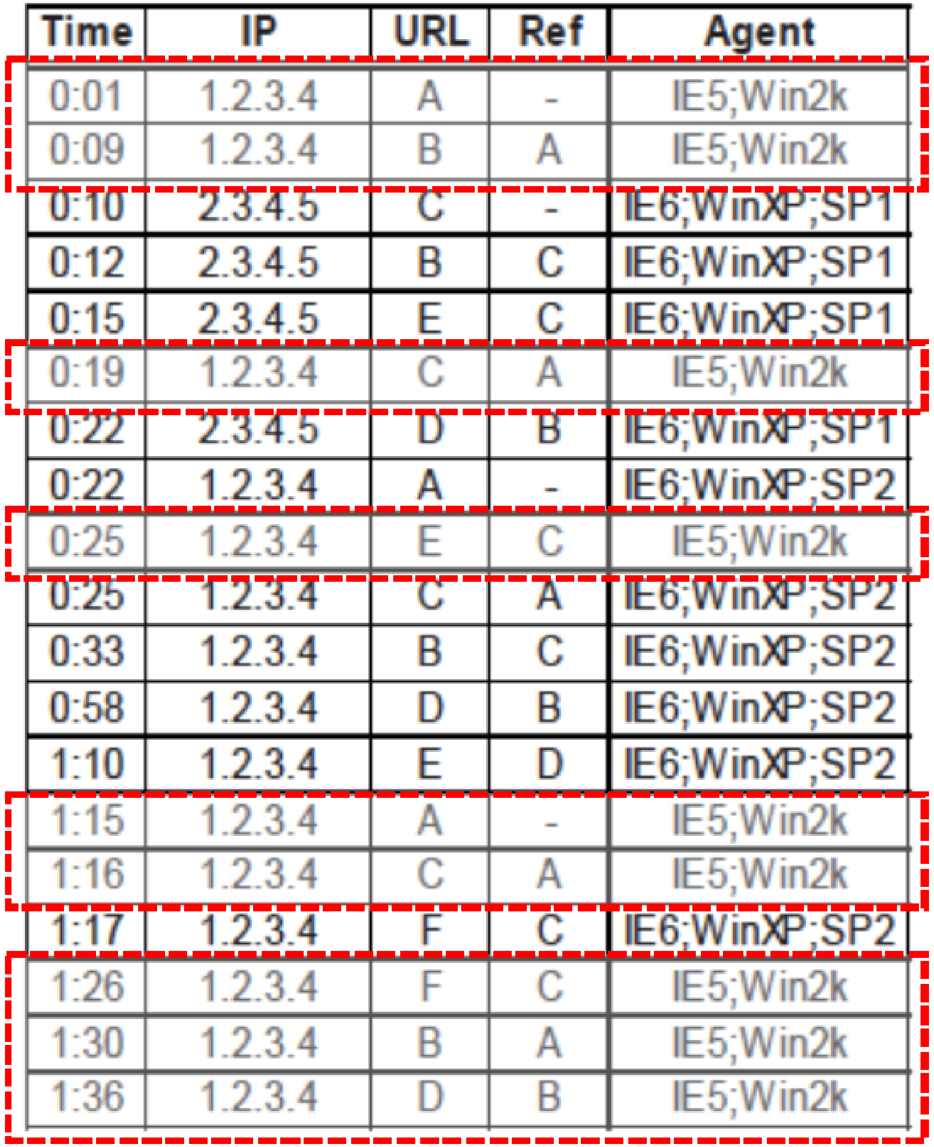
Cleansing

- Remove the following from the original log file
  - Entries with suffixes like .jpg, .jpeg, .css, etc.
  - Entries having status code failure (e.g. Forbidden, Method Not Allowed).

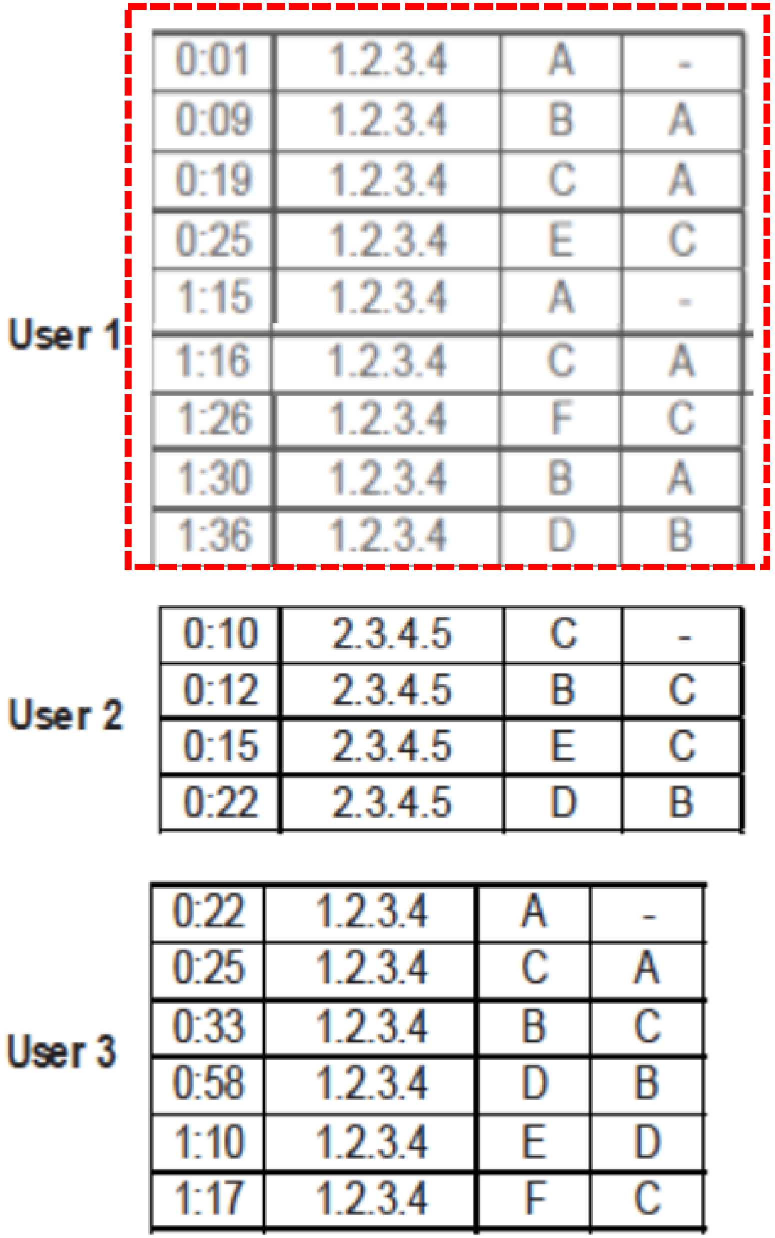
- Remove all records which do not contain method “GET” and “POST” (others like DELETE, TRACE, .. are not useful for understanding browsing behavior).

- Remove navigation sessions performed by crawlers/spiders.

User Identification (E.g., Same IP + Agent)



Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k



User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:16	1.2.3.4	C	A
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

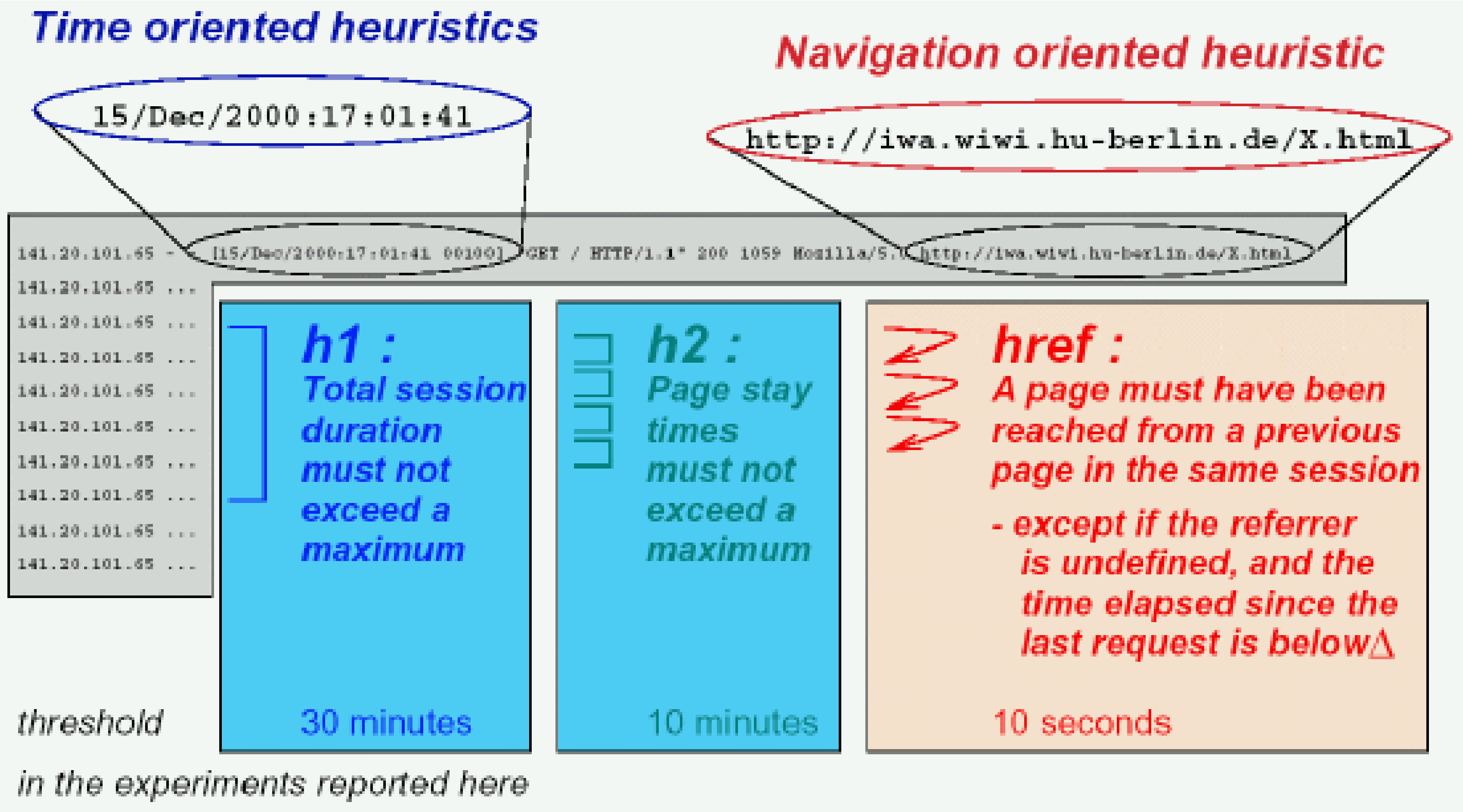
0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Identify sessions (sessionization)

- In Web usage analysis, these data are the sessions of the site visitors: the activities performed by a user from the moment she enters the site until the moment she leaves it.

- Difficult to obtain reliable usage data due to proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the inability of servers to distinguish among different visits.

Sessionization heuristics



Time oriented heuristics

Navigation oriented heuristic

```
141.20.101.65 - [15/Dec/2000:17:01:41 00100] GET / HTTP/1.1" 200 1059 Mozilla/5.0 http://iwa.wiwi.hu-berlin.de/X.html
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
141.20.101.65 ...
```

h1 : Total session duration must not exceed a maximum

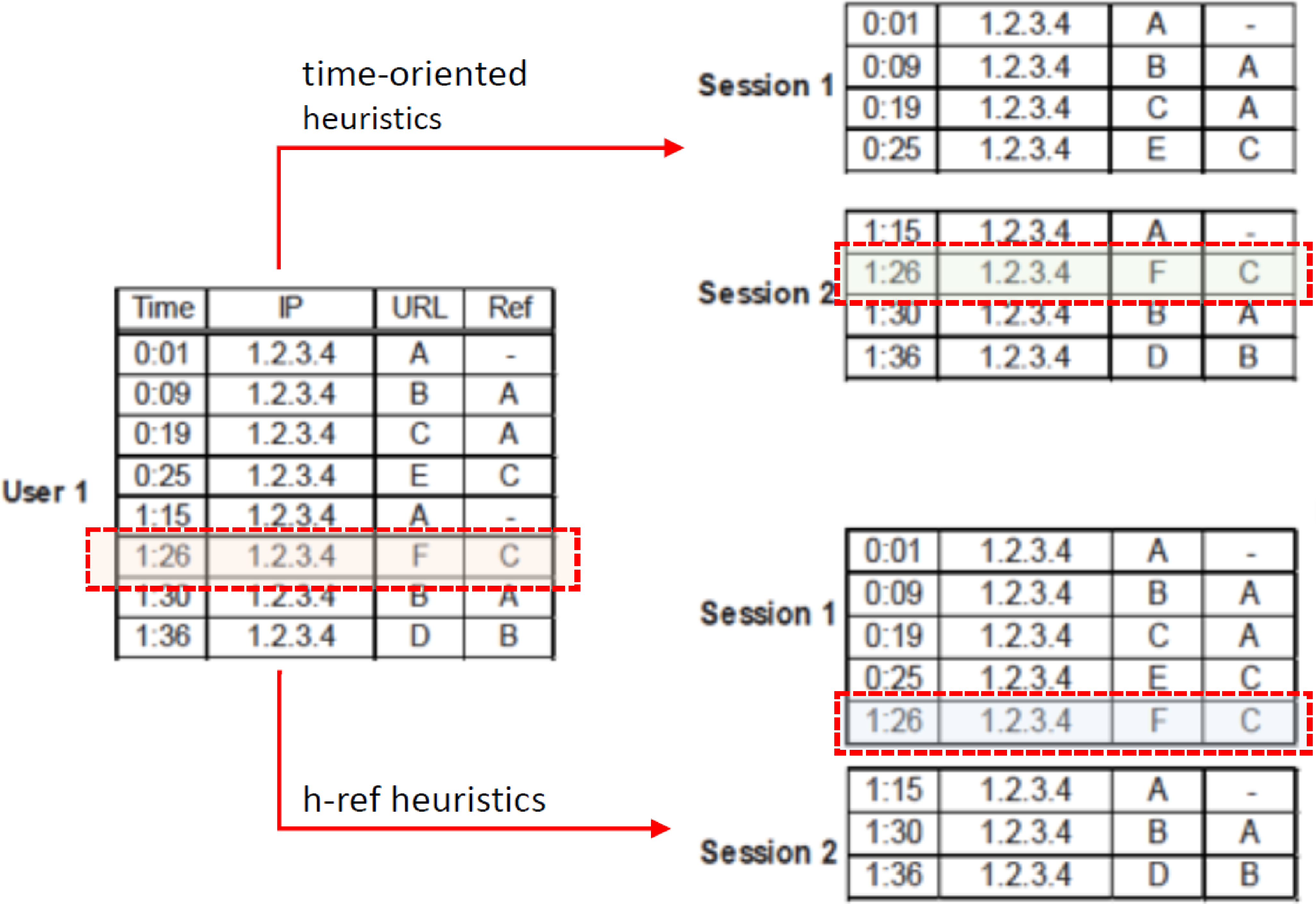
h2 : Page stay times must not exceed a maximum

href : A page must have been reached from a previous page in the same session

- except if the referrer is undefined, and the time elapsed since the last request is below Δ

threshold	30 minutes	10 minutes	10 seconds

in the experiments reported here



User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

time-oriented heuristics

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

h-ref heuristics

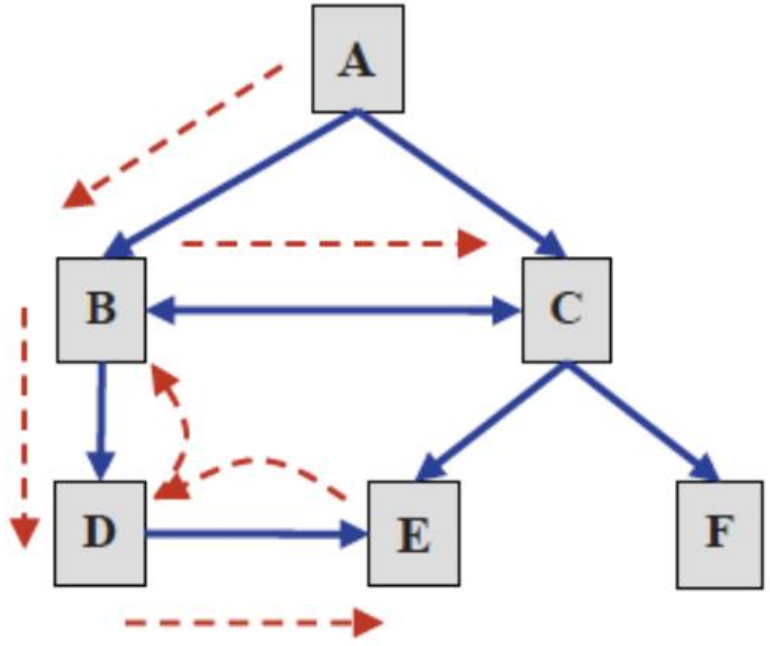
Session 1

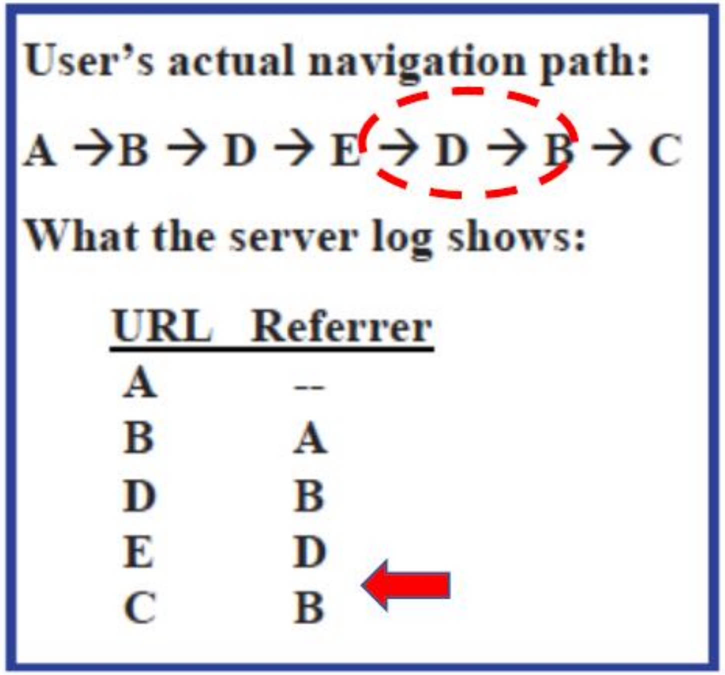
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:26	1.2.3.4	F	C

Session 2

1:15	1.2.3.4	A	-
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Caching and Path Completion





User's actual navigation path:

A →B → D → E → D → B → C

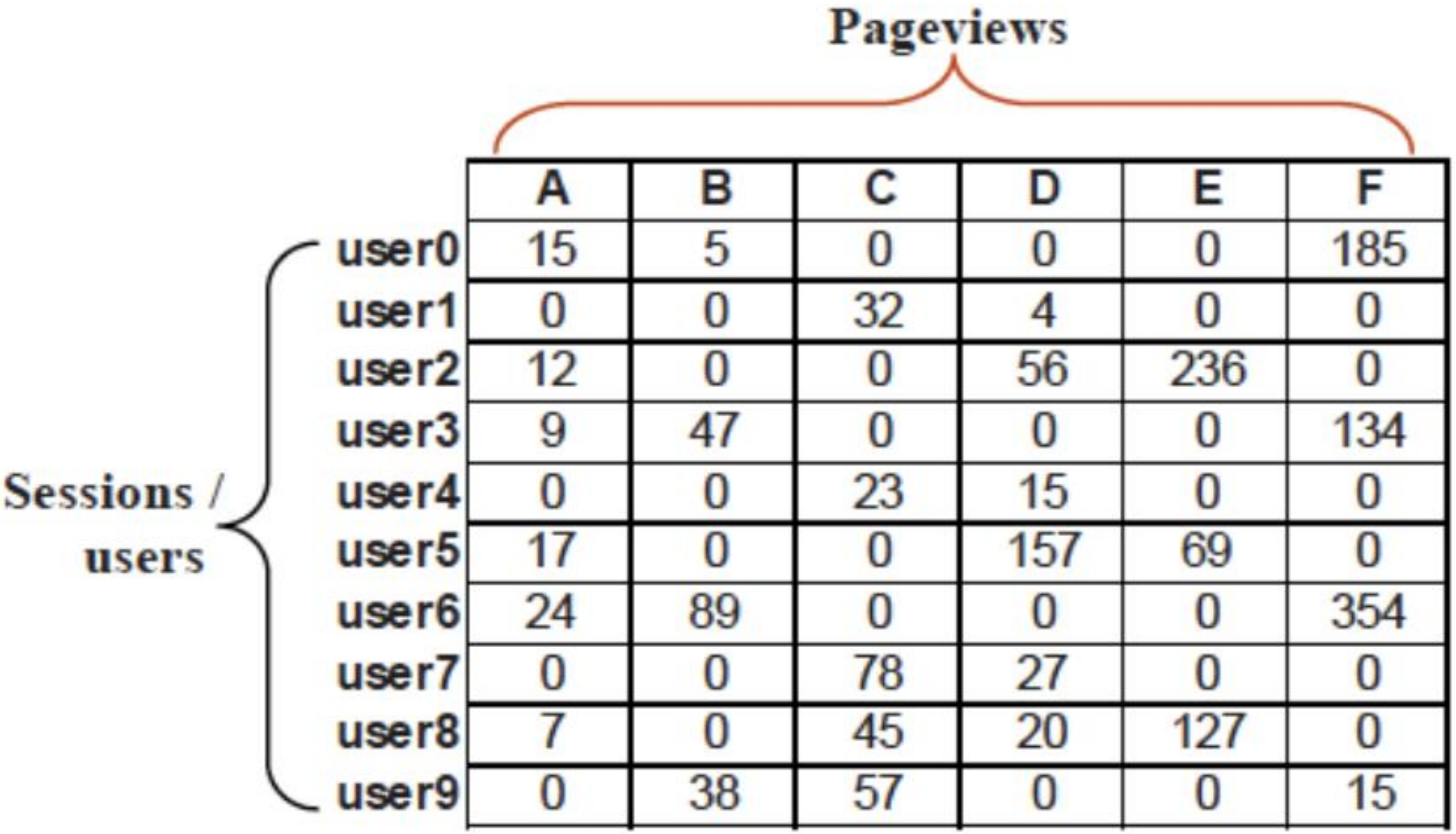
What the server log shows:

URL	Referrer
A	--
B	A
D	B
E	D
C	B

Fig. 12.7. Missing references due to caching.

Web-site structure is considered to infer the path

Transaction Matrix



Pageviews

Sessions / users

	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

Fig. 12.8. An example of a user-pageview matrix (or transaction matrix)

Analyze a Transaction Matrix

- Applicable algorithm?
  - Collaborative Filtering (#views as ratings)
  - Information Retrieval (a user as a doc)
  - Frequent Itemsets and Association Rules (a user as a "basket")

Sequential Pattern Mining

- If sequential patterns in user transactions are to be explored, sequential pattern mining techniques will be needed.
- User transactions modeled as Markov Chains
  - Markov Chains are models used to study systems that change over time.
  - They are characterized by a set of states and a transition matrix that describes the probability of moving from one state to another (states = pages)
  - The transition matrix is usually represented by a square matrix where each row corresponds to the current state/page, and each column corresponds to the next state.
  - Markov Chains have the property of memorylessness, meaning that the probability of moving to a future state only depends on the current state and not on any past states.
  - Markov Chains are a model which allows to identify the most common navigation paths followed by users on the website

References

- Liu, Bing. Web data mining: exploring hyperlinks, contents, and usage data. Berlin: springer, 2011. Chapter 12.