Web Mining - CSE3024
Slot – L33-L34
Faculty – Prof. Sridhar R
Dhrubanka Dutta
17BCE1019

Web usage mining lab on 18th March 2019

Following is a set of web usage logs from an organization. Assume the following rules.

a. The time heuristic h1 for a session duration = 30 minutes.

b. The time heuristic h2 for the average duration for the user visit of a page = 1 minute.

c. Include a page in a session if the page that refers is in that session.

d. During a session, the IP address, browser and OS is the same.

You need to printout a table containing session number, IP address, session start time and session end time.

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	-	IE4;Win98
0:12	2.3.4.5	B	C	IE4;Win98
0:15	2.3.4.5	E	C	IE4;Win98
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE4;Win98
0:22	1.2.3.4	A	-	IE4;Win98
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE4;Win98
0:33	1.2.3.4	B	C	IE4;Win98
0:58	1.2.3.4	D	B	IE4;Win98
1:10	1.2.3.4	E	D	IE4;Win98
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	IE4;Win98
1:25	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

CODE:

```
time = ["0:01", "0:09","0:10","0:12", "0:15","0:19",
        "0:22", "0:22","0:25","0:25", "0:33","0:58",
        "1:10", "1:15","1:16","1:17", "1:25","1:30",
        "1:36"]

ip = ["1.2.3.4","1.2.3.4","2.3.4.5","2.3.4.5","2.3.4.5",
      "1.2.3.4","2.3.4.5","1.2.3.4","1.2.3.4","1.2.3.4",
      "1.2.3.4","1.2.3.4","1.2.3.4","1.2.3.4","1.2.3.4",
      "1.2.3.4","1.2.3.4","1.2.3.4","1.2.3.4"]

url = ['A','B','C','B','E','C','D','A','E','C','B','D','E','A','C','F','F','B','D']

ref = ['0','A','0','C','C','A','B','0','C','A','C','B','D','0','A','C','C','A','B']

agent = ["5:2K","5:2K","4:98","4:98","4:98","5:2K","4:98","4:98","5:2K","4:98","4:98","4:98",
         "4:98","5:2K","5:2K","4:98","5:2K","5:2K","5:2K"]

mlis, check = [],[]

mlis.append(time)
mlis.append(ip)
mlis.append(url)
mlis.append(ref)
mlis.append(agent)

for i in range(len(mlis[0])):
    check.append(0)
    sl1 = int(mlis[0][i][0])
    sl2 = int(mlis[0][i][2:4])
    mlis[0][i] = sl1*60+sl2

ipset = list(set(ip))
agentset = list(set(agent))

mainlis = []
newvar = mlis[0][0] + mlis[0][len(mlis[0])-1]
interval = newvar-mlis[0][0]
toint = (interval//30)+1

for j in range(2):
    for k in range(2):
        lis = []
        for i in range(len(mlis[1])):
            if mlis[1][i] == ipset[j] and mlis[4][i]==agentset[k]:
                lis.append(i)
        mainlis.append(lis)

mainlis2 = []
lis = []
```

```
for x in range(2):
    for i in mainlis:
        for j in range(len(i)):
            val = mlis[0][i[j]]
            if mlis[3][i[j]]=='0':
                mins = mlis[0][i[j]]
                maxs = mins+30
                if lis:
                    mainlis2.append(lis)
                    lis = []
            if val>=mins and val<maxs:
                lis.append(i[j])

mainlis2 = mainlis2[0:4]
finlis = []
lis = []
for i in mainlis2:
    val = mlis[2][i[0]]
    lis.append(val)
    for j in range(1,len(i)):
        lis.append(mlis[2][i[j]])
    finlis.append(lis)
    lis = []

count = 0
print("h1 heuristic along with h2 with the reference url's order is in below")
for i in mainlis2:
    if i:
        for j in i:
            print(mlis[0][j],mlis[1][j],mlis[2][j],mlis[3][j],mlis[4][j])
        print(finlis[count])
        print("\n")
        count=count+1
```

OUTPUT: