Poniedziałek, 10:00

Eksploracja danych internetowych

Zadanie 1

Eksploracja użycia na podstawie pliku logów

Barbara Morawska 234096

Andrzej Sasinowski 234118

Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej

Politechnika Łódzka

2020/2021

1. Cel

Celem ćwiczenia było wykonanie analizy wybranego pliku logów w formacie Common Log Format. Przed wykonaniem czynności związanych z eksploracją danych należało odpowiednio przetworzyć plik. Do wykonania analizy klastrowej oraz znalezienia reguł asocjacyjnych wykorzystano program Weka.

2. Opis oraz przygotowanie danych

2.1. Opis oryginalnego pliku

Zadanie zostało wykonane przy użyciu pliku `access_log_Jul95`, w którym znajdują się żądania HTTP wysyłane do strony internetowej Centrum Kosmicznego Johna F. Kennedy'ego w lipcu 1995 roku. Każdy wiersz pliku zawiera kolejno adres IP hosta, pole identyfikacji oraz nazwę użytkownika, znacznik czasu, żądanie (w którym można wyróżnić metodę, adres oraz protokół), kod odpowiedzi HTTP oraz liczbę bajtów w odpowiedzi. Do przejrzeniu pliku okazało się jednak, że pola identyfikacji oraz nazwy użytkownika najczęściej są polami pustymi (co jest zaznaczone za pomocą znaku „-").

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085
burger.letters.com - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0" 200 4179
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 304 0
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/video/livevideo.gif HTTP/1.0" 200 0
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985
d104.aa.net - - [01/Jul/1995:00:00:13 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
129.94.144.152 - - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:18 -0400] "GET /facts/about_ksc.html HTTP/1.0" 200 3977
net-1-141.eden.com - - [01/Jul/1995:00:00:19 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0916.jpg HTTP/1.0" 200 34029
```

Rysunek 1: Przykładowe 20 wierszy omawianego pliku.

2.2. Przygotowanie pliku (utworzenie plików w formacie arff)

Do przetworzenia pliku, a także do wyodrębnienia sesji oraz użytkowników i konwersji pliku na format arff wykorzystano przygotowane skrypty w języku Python. Przed przetworzeniem danych za pomocą skryptów przygotowano dodatkowy plik, w którym znajdowało się pierwsze 50000 rekordów z oryginalnego pliku. Tak otrzymany plik przetworzono za pomocą skryptu `main.py`, w którym wyodrębniono dodatkowo metodę, adres strony oraz protokół

z kolumny, w której znajduje się żądanie użytkownika. Następnie wybrano tylko te rekordy, które korzystały z metody GET, otrzymanym kodem odpowiedzi był kod 200, a dodatkowo usunięto te rekordy, w których żądano plików graficznych (w tym celu odfiltrowano wszystkie adresy o rozszerzeniach jpg, gif, bmp, xbm, png, jpeg, mpg oraz mpeg). Przed wyodrębnieniem sesji przygotowano plik, w którym znajdowały strony posortowane pod względem liczby wizyt. Obliczono procentową wartość odwiedzin, a następnie wybrano tylko te strony, dla których wartość ta była większa niż 0.5% (skrypt `trending.py`). Za pomocą napisanego skryptu mining.py wyodrębniono sesje, przyjmując 30 minut jako maksymalny próg czas, który może upłynąć pomiędzy dwoma rekordami tego samego użytkownika, aby można było mówić o sesji. Do nadania atrybutów oraz przeprowadzeniu transformacji koszykowej dla wyodrębnionych sesji wykorzystano skrypt `adding_attributes.py`, dzięki czemu obliczono dodatkowo czas sesji (w sekundach), liczbę odwiedzonych stron w ramach sesji oraz przeciętny czas na jedną stronę (również w sekundach). Po przeprowadzeniu transformacji koszykowej uzyskano dla każdej sesji flagi dla najpopularniejszych stron (wartość 0 oznacza, że strona nie została odwiedzona, natomiast 1 – została odwiedzona). Przy pomocy tego samego skryptu dokonano dyskretyzacji, przyporządkowując każdy z powyższych trzech wartości numerycznych do kategorii odpowiadającym określonym przedziałom liczbowym.

```
@RELATION session_attributes.arff

@ATTRIBUTE sessionID STRING
@ATTRIBUTE /shuttle/countdown/ {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/images/images.html {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/mission-sts-71.html {0, 1}
@ATTRIBUTE / {0, 1}
@ATTRIBUTE /ksc.html {0, 1}
@ATTRIBUTE /shuttle/countdown/liftoff.html {0, 1}
@ATTRIBUTE /shuttle/missions/missions.html {0, 1}
@ATTRIBUTE /htbin/cdt_main.pl {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/movies/movies.html {0, 1}
@ATTRIBUTE /history/apollo/apollo.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/apollo-13.html {0, 1}
@ATTRIBUTE /history/history.html {0, 1}
@ATTRIBUTE /shuttle/countdown/countdown.html {0, 1}
@ATTRIBUTE /shuttle/technology/sts-newsref/stsref-toc.html {0, 1}
@ATTRIBUTE /shuttle/resources/orbiters/atlantis.html {0, 1}
@ATTRIBUTE /htbin/cdt_clock.pl {0, 1}
@ATTRIBUTE /software/winvn/winvn.html {0, 1}
@ATTRIBUTE /shuttle/countdown/lps/fr.html {0, 1}
@ATTRIBUTE /shuttle/missions/sts-67/mission-sts-67.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/apollo-13-info.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-11/apollo-11.html {0, 1}
@ATTRIBUTE /facilities/lc39a.html {0, 1}
@ATTRIBUTE /shuttle/missions/sts-70/mission-sts-70.html {0, 1}
@ATTRIBUTE /shuttle/countdown/tour.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/ {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/images/ {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/news/ {0, 1}
@ATTRIBUTE /shuttle/missions/sts-63/mission-sts-63.html {0, 1}
@ATTRIBUTE session_time NUMERIC
@ATTRIBUTE visited_sites NUMERIC
@ATTRIBUTE avg_time_per_site NUMERIC
@ATTRIBUTE session_time_categories {t<1min, 1min<=t<2min, 3min<=t<5min, 5min<=t<10min, 10min<=t<15min, 15min<=t<30min}
@ATTRIBUTE session_average_time_per_site_categories {t<1min, 1min<=t<2min, 2min<=t<5min, 5min<=t<30min}
@ATTRIBUTE done_things_categories {x<=3, 3<=x<6, 6<=x<10, 10<=x<20, x>=20}

@DATA
c5b67c26-1246-11eb-91d9-e935f39b3778,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,80,4,20.0,t<1min,t<1min,3<=x<6
c4984ea8-1248-11eb-91d9-e935f39b3778,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,52,2,26.0,t<1min,t<1min,x<=3
15227d8e-124f-11eb-91d9-e935f39b3778,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,23,2,11.5,t<1min,t<1min,x<=3
ed709d92-1247-11eb-91d9-e935f39b3778,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,804,2,402.0,5min<=t<10min,5min<=t<30min,x<=3
ed709d93-1247-11eb-91d9-e935f39b3778,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,211,2,105.5,1min<=t<2min,1min<=t<2min,x<=3
ed709d91-1247-11eb-91d9-e935f39b3778,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1589,3,529.6666666666666,5min<=t<10min,5min<=t<30min,x<=3
7bd12576-124f-11eb-91d9-e935f39b3778,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,146,3,48.66666666666664,t<1min,t<1min,x<=3
60822b0a-1248-11eb-91d9-e935f39b3778,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,904,5,180.8,3min<=t<5min,2min<=t<5min,3<=x<6
d10b951c-1240-11eb-91d9-e935f39b3778,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,2,1.5,t<1min,t<1min,x<=3
98312ac2-124a-11eb-91d9-e935f39b3778,1,1,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,343,6,57.166666666666664,t<1min,t<1min,3<=x<6
```

Rysunek 2: Plik arff zawierający wyodrębnione sesje wraz z atrybutami oraz kategoriami dla czasu sesji, liczby odwiedzonych stron oraz średniego czasu na jedną stronę.

Skrypt `mining_users.py` wykorzystano do wyodrębnienia użytkowników oraz przeprowadzenia transformacji koszykowej, dzięki czemu każdemu użytkownikowi nadano odpowiednie wartości atrybutów odpowiadające stronom, które odwiedzali użytkownicy.

```
@RELATION user_attributes.arff

@ATTRIBUTE userID STRING
@ATTRIBUTE /shuttle/countdown/ {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/images/images.html {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/mission-sts-71.html {0, 1}
@ATTRIBUTE / {0, 1}
@ATTRIBUTE /ksc.html {0, 1}
@ATTRIBUTE /shuttle/countdown/liftoff.html {0, 1}
@ATTRIBUTE /shuttle/missions/missions.html {0, 1}
@ATTRIBUTE /htbin/cdt_main.pl {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/movies/movies.html {0, 1}
@ATTRIBUTE /history/apollo/apollo.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/apollo-13.html {0, 1}
@ATTRIBUTE /history/history.html {0, 1}
@ATTRIBUTE /shuttle/countdown/countdown.html {0, 1}
@ATTRIBUTE /shuttle/technology/sts-newsref/stsref-toc.html {0, 1}
@ATTRIBUTE /shuttle/resources/orbiters/atlantis.html {0, 1}
@ATTRIBUTE /htbin/cdt_clock.pl {0, 1}
@ATTRIBUTE /software/winvn/winvn.html {0, 1}
@ATTRIBUTE /shuttle/countdown/lps/fr.html {0, 1}
@ATTRIBUTE /shuttle/missions/sts-67/mission-sts-67.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/apollo-13-info.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-11/apollo-11.html {0, 1}
@ATTRIBUTE /facilities/lc39a.html {0, 1}
@ATTRIBUTE /shuttle/missions/sts-70/mission-sts-70.html {0, 1}
@ATTRIBUTE /shuttle/countdown/tour.html {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/ {0, 1}
@ATTRIBUTE /history/apollo/apollo-13/images/ {0, 1}
@ATTRIBUTE /shuttle/missions/sts-71/news/ {0, 1}
@ATTRIBUTE /shuttle/missions/sts-63/mission-sts-63.html {0, 1}


@DATA
199.72.81.55,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
unicomp6.unicomp.net,1,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
199.120.110.21,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
205.212.115.106,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
d104.aa.net,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
129.94.144.152,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
ppptky391.asahi-net.or.jp,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
205.189.154.54,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
waters-gw.starway.net.au,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
ppp-mia-30.shadow.net,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Rysunek 3: Plik arff zawierający wyodrębnionych użytkowników wraz z atrybutami.

3. Eksploracja danych

Przeprowadzono proces analizy klastrowej dla sesji oraz dla użytkowników, a także znaleziono reguły asocjacyjne dla pliku sesji (korzystając z wyznaczonych wcześniej wartości typu kategorycznego). Proces klastrowania dla pliku sesji został podzielony na analizę odwiedzonych stron w trakcie sesji (atrybuty dla stron) oraz analizę wartości numerycznych (czas trwania sesji, liczba odwiedzonych stron podczas sesji oraz średni czas na jedną stronę). Do analizy skupień użyto algorytmu K-średnich, który jest algorytmem iteracyjnym. Początkowo wyznaczane są środki skupienia każdego z klastrów, następnie przy użyciu jednej

z miar (w przypadku zadania użyto miary Euklidesowej) liczony jest dystans każdej próbki od środków skupienia – próbki te dołączane są do klastra, który jest najbliżej. Następnie dla każdego klastra wartości wszystkich próbek tego klastra są uśredniane i w ten sposób wyznaczany jest nowy środek skupienia. Eksperyment powtórzono trzy razy dla każdego z trzech opisanych wcześniej przypadków – wyznaczając kolejno 2, 5 oraz 8 klastrów. Do wyznaczenia początkowych środków każdego klastra skorzystano z metody losowej, ustawiając wartość ziarna na 854.

3.1. Analiza skupień dla sesji (odwiedzone strony)

Algorytm dla 2 klastrów zatrzymał się po 3 iteracjach, a suma błędów kwadratowych jest równa 6904.

		Cluster#	
Attribute	Full Data	0	1
	(2987.0)	(1682.0)	(1305.0)
/shuttle/countdown/	0	1	0
/shuttle/missions/sts-71/images/images.html	0	0	0
/shuttle/missions/sts-71/mission-sts-71.html	0	1	0
/	0	0	0
/ksc.html	0	0	0
/shuttle/countdown/liftoff.html	0	0	0
/shuttle/missions/missions.html	0	0	0
/htbin/cdt_main.pl	0	0	0
/shuttle/missions/sts-71/movies/movies.html	0	0	0
/history/apollo/apollo.html	0	0	0
/history/apollo/apollo-13/apollo-13.html	0	0	0
/history/history.html	0	0	0
/shuttle/countdown/countdown.html	0	0	0
/shuttle/technology/sts-newsref/stsref-toc.html	0	0	0
/shuttle/resources/orbiters/atlantis.html	0	0	0
/htbin/cdt_clock.pl	0	0	0
/software/winvn/winvn.html	0	0	0
/shuttle/countdown/lps/fr.html	0	0	0
/shuttle/missions/sts-67/mission-sts-67.html	0	0	0
/history/apollo/apollo-13/apollo-13-info.html	0	0	0
/history/apollo/apollo-11/apollo-11.html	0	0	0
/facilities/lc39a.html	0	0	0
/shuttle/missions/sts-70/mission-sts-70.html	0	0	0
/shuttle/countdown/tour.html	0	0	0
/history/apollo/apollo-13/	0	0	0
/history/apollo/apollo-13/images/	0	0	0
/shuttle/missions/sts-71/news/	0	0	0
/shuttle/missions/sts-63/mission-sts-63.html	0	0	0

Rysunek 4: Wartości środków skupienia klastrów dla flag stron podczas analizy sesji (2 klastry).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	1682	56%
1	1305	44%

Rysunek 5: Rozkład wszystkich danych na poszczególne klastry (2 klastry).

Algorytm dla 5 klastrów zatrzymał się po 4 iteracjach, a suma błędów kwadratowych jest równa 5937.

Attribute	Full Data (2987.0)	Cluster# 0 (1335.0)	1 (464.0)	2 (153.0)	3 (920.0)	4 (115.0)
/shuttle/countdown/	0	1	0	0	0	1
/shuttle/missions/sts-71/images/images.html	0	0	0	0	0	0
/shuttle/missions/sts-71/mission-sts-71.html	0	1	0	0	0	0
/	0	0	1	0	0	0
/ksc.html	0	0	0	0	0	1
/shuttle/countdown/liftoff.html	0	0	0	0	0	0
/shuttle/missions/missions.html	0	0	0	0	0	0
/htbin/cdt_main.pl	0	0	0	0	0	0
/shuttle/missions/sts-71/movies/movies.html	0	0	0	0	0	0
/history/apollo/apollo.html	0	0	0	1	0	0
/history/apollo/apollo-13/apollo-13.html	0	0	0	1	0	0
/history/history.html	0	0	0	0	0	0
/shuttle/countdown/countdown.html	0	0	0	0	0	0
/shuttle/technology/sts-newsref/stsref-toc.html	0	0	0	0	0	0
/shuttle/resources/orbiters/atlantis.html	0	0	0	0	0	0
/htbin/cdt_clock.pl	0	0	0	0	0	0
/software/winvn/winvn.html	0	0	0	0	0	0
/shuttle/countdown/lps/fr.html	0	0	0	0	0	0
/shuttle/missions/sts-67/mission-sts-67.html	0	0	0	0	0	0
/history/apollo/apollo-13/apollo-13-info.html	0	0	0	1	0	0
/history/apollo/apollo-11/apollo-11.html	0	0	0	0	0	0
/facilities/lc39a.html	0	0	0	0	0	0
/shuttle/missions/sts-70/mission-sts-70.html	0	0	0	0	0	0
/shuttle/countdown/tour.html	0	0	0	0	0	0
/history/apollo/apollo-13/	0	0	0	0	0	0
/history/apollo/apollo-13/images/	0	0	0	0	0	0
/shuttle/missions/sts-71/news/	0	0	0	0	0	0
/shuttle/missions/sts-63/mission-sts-63.html	0	0	0	0	0	0

Rysunek 6: Wartości środków skupienia klastrów dla flag stron podczas analizy sesji (5 klastrów).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	1335	45%
1	464	16%
2	153	5%
3	920	31%
4	115	4%

Rysunek 7: Rozkład wszystkich danych na poszczególne klastry (5 klastrów).

Algorytm dla 8 klastrów zatrzymał się po 4 iteracjach, a suma błędów kwadratowych jest równa 5228.

		Cluster#							
Attribute	Full Data	0	1	2	3	4	5	6	7
	(2987.0)	(622.0)	(294.0)	(148.0)	(1120.0)	(207.0)	(324.0)	(264.0)	(8.0)
/shuttle/countdown/	0	1	0	0	0	1	1	1	0
/shuttle/missions/sts-71/images/images.html	0	0	0	0	0	0	1	0	0
/shuttle/missions/sts-71/mission-sts-71.html	0	0	0	0	0	0	1	1	1
/	0	0	1	0	0	0	1	0	0
/ksc.html	0	0	0	0	0	1	0	0	0
/shuttle/countdown/liftoff.html	0	0	0	0	0	0	0	0	0
/shuttle/missions/missions.html	0	0	0	0	0	0	0	0	0
/htbin/cdt_main.pl	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-71/movies/movies.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo.html	0	0	0	1	0	0	0	0	0
/history/apollo/apollo-13/apollo-13.html	0	0	0	1	0	0	0	0	0
/history/history.html	0	0	0	0	0	0	0	0	0
/shuttle/countdown/countdown.html	0	0	0	0	0	0	0	0	0
/shuttle/technology/sts-newsref/stsref-toc.html	0	0	0	0	0	0	0	0	0
/shuttle/resources/orbiters/atlantis.html	0	0	0	0	0	0	0	0	0
/htbin/cdt_clock.pl	0	0	0	0	0	0	0	0	0
/software/winvn/winvn.html	0	0	0	0	0	0	0	0	0
/shuttle/countdown/lps/fr.html	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-67/mission-sts-67.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/apollo-13-info.html	0	0	0	1	0	0	0	0	0
/history/apollo/apollo-11/apollo-11.html	0	0	0	0	0	0	0	0	0
/facilities/lc39a.html	0	0	0	0	0	0	0	0	1
/shuttle/missions/sts-70/mission-sts-70.html	0	0	0	0	0	0	0	0	0
/shuttle/countdown/tour.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/images/	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-71/news/	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-63/mission-sts-63.html	0	0	0	0	0	0	0	0	0

Rysunek 8: Wartości środków skupienia klastrów dla flag stron podczas analizy sesji (8 klastrów).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	622	21%
1	294	10%
2	148	5%
3	1120	37%
4	207	7%
5	324	11%
6	264	9%
7	8	0%

Rysunek 9: Rozkład wszystkich danych na poszczególne klastry (8 klastrów).

Po przeanalizowaniu powyższych wyników pierwszym wnioskiem jest to, że 2 klastry to niewystarczająca liczba na wyciągnięcie większej liczby informacji z analizy skupień dla atrybutów dotyczących przeglądanych stron. Liczba sesji przydzielona do każdego z dwóch klastrów była dość podobna (różnica rzędu kilku procent). Dla dwóch klastrów można jedynie stwierdzić, że jeśli odwiedzono stronę `/shuttle/countdown/`, to często w ramach tej samej sesji odwiedzono również `/shuttle/missions/sts-71/mission-sts-71.html`, drugi klaster informuje o tym, że żadna z najpopularniejszych stron nie została odwiedzona. Dla 5 klastrów rozkład zdecydowanie nie jest równomierny, jednak warto zauważyć, że dwa najliczniejsze klastry dotyczą tych samych grup co w przypadku 2 klastrów. Pozostałe klastry pokazują między innymi, że w ramach jednej sesji często przeglądano strony

/history/apollo/apollo.html, /history/apollo/apollo-13/apollo-13.html oraz /history/apollo/apollo-13/apollo-13-info.html. W przypadku jednego z klastrów często odwiedzaną stroną był adres /ksc.html. Dla 8 klastrów rozkład również nie był równoliczny, dominował klaster o numerze, który pojawił się już w poprzednich przedstawionych analizach skupień. Pojawił się również jeden bardzo mało liczny klaster (jedynie 8 obserwacji), według którego w ramach tej samej sesji przeglądane były adresy /shuttle/missions/sts-71/mission-sts-71.html oraz /facilities/lc39a.html.

3.2. Analiza skupień dla sesji (wartości numeryczne)

Algorytm dla 2 klastrów zatrzymał się po 9 iteracjach, a suma błędów kwadratowych jest równa (w przybliżeniu do 2 miejsca po przecinku) 105.17.

		Cluster#	
Attribute	Full Data	0	1
	(2987.0)	(674.0)	(2313.0)
session_time	433.0144	1256.8234	192.9594
visited_sites	4.92	8.9125	3.7566
avg_time_per_site	89.4662	222.3592	50.7416

Rysunek 10: Wartości środków skupienia klastrów dla wartości statystycznych podczas analizy sesji (2 klastry).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	674	23%
1	2313	77%

Rysunek 11: Rozkład wszystkich danych na poszczególne klastry (2 klastry).

Algorytm dla 5 klastrów zatrzymał się po 49 iteracjach, a suma błędów kwadratowych jest równa (w przybliżeniu do 2 miejsca po przecinku) 42.61.

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(2987.0)	(173.0)	(586.0)	(212.0)	(1669.0)	(347.0)
session_time	433.0144	1487.1908	419.1092	1510.2925	97.0683	888.5965
visited_sites	4.92	3.896	6.0085	15.6462	2.9509	6.5101
avg_time_per_site	89.4662	437.0631	88.0586	115.5897	33.3333	172.5732

Rysunek 12: Wartości środków skupienia klastrów dla wartości statystycznych podczas analizy sesji (5 klastrów).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	173	6%
1	586	20%
2	212	7%
3	1669	56%
4	347	12%

Rysunek 13: Rozkład wszystkich danych na poszczególne klastry (5 klastry).

Algorytm dla 8 klastrów zatrzymał się po 84 iteracjach, a suma błędów kwadratowych jest równa (w przybliżeniu do 2 miejsca po przecinku) 27.06.

		Cluster#							
Attribute	Full Data	0	1	2	3	4	5	6	7
	(2987.0)	(185.0)	(369.0)	(69.0)	(180.0)	(148.0)	(96.0)	(637.0)	(1303.0)
session_time	433.0144	1546.0703	541.8428	1504.913	982.8111	881.9797	1555.1458	247.1476	68.6462
visited_sites	4.92	7.5351	6.4607	2.6232	10.4056	3.3784	21.6563	4.4867	2.6301
avg_time_per_site	89.4662	231.5289	103.329	599.8599	105.9941	280.0436	78.5077	65.9968	26.6936

Rysunek 14: Wartości środków skupienia klastrów dla wartości statystycznych podczas analizy sesji (8 klastrów).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	185	6%
1	369	12%
2	69	2%
3	180	6%
4	148	5%
5	96	3%
6	637	21%
7	1303	44%

Rysunek 15: Rozkład wszystkich danych na poszczególne klastry (8 klastrów).

W przypadku 2 klastrów uzyskujemy informacje o pewnej tendencji. Aż 77% wszystkich obserwacji dotyczy sesji o mniejszych wartościach atrybutów czasu oraz liczby odwiedzanych stron, a także średniego czasu poświęconego na jedną stronę. Dzięki temu można wysnuć wniosek, że większość użytkowników miała krótsze sesje, jednak większa liczba klastrów dodatkowo może potwierdzić powyższe spostrzeżenie. Dla 5 klastrów ponad połowa (56%) próbek zostało przyporządkowanych klastrowi z najmniejszym czasem sesji, najmniejszą liczbą odwiedzonych stron oraz najmniejszym średnim czasem na pojedynczą stronę. Najdłuższe sesje rozłożyły się na dwa klastry, które różniły się liczbą odwiedzanych stron (około 4 strony dla klastra 0 oraz około 16 stron dla klastra 2), co bezpośrednio przełożyło się na średni czas

na pojedynczą stronę. Analiza skupień przy użyciu 8 klastrów potwierdziła przypuszczenie, że sesje użytkowników trwały najczęściej stosunkowo krótko (około 69 sekund dla klastra, który zawiera aż 44% wszystkich obserwacji). Następny pod względem liczebności klaster był również drugim klastrem o najkrótszym czasie. Jedną z różnic w porównaniu z analizą skupień dla odwiedzonych stron jest to, że w przypadku wartości numerycznych wraz ze wzrostem liczby klastrów rośnie liczba wykonanych iteracji, ale maleje suma błędów kwadratowych.

3.3. Analiza skupień dla użytkowników (odwiedzone strony)

Algorytm dla 2 klastrów zatrzymał się po 3 iteracjach, a suma błędów kwadratowych jest równa 8176.

Attribute	Full Data (3782.0)	Cluster# 0 (3094.0)	1 (688.0)
/shuttle/countdown/	0	0	0
/shuttle/missions/sts-71/images/images.html	0	0	0
/shuttle/missions/sts-71/mission-sts-71.html	0	0	0
/	0	0	1
/ksc.html	0	0	0
/shuttle/countdown/liftoff.html	0	0	0
/shuttle/missions/missions.html	0	0	0
/htbin/cdt_main.pl	0	0	0
/shuttle/missions/sts-71/movies/movies.html	0	0	0
/history/apollo/apollo.html	0	0	0
/history/apollo/apollo-13/apollo-13.html	0	0	0
/history/history.html	0	0	0
/shuttle/countdown/countdown.html	0	0	0
/shuttle/technology/sts-newsref/stsref-toc.html	0	0	0
/shuttle/resources/orbiters/atlantis.html	0	0	0
/htbin/cdt_clock.pl	0	0	0
/software/winvn/winvn.html	0	0	0
/shuttle/countdown/lps/fr.html	0	0	0
/shuttle/missions/sts-67/mission-sts-67.html	0	0	0
/history/apollo/apollo-13/apollo-13-info.html	0	0	0
/history/apollo/apollo-11/apollo-11.html	0	0	0
/facilities/lc39a.html	0	0	0
/shuttle/missions/sts-70/mission-sts-70.html	0	0	0
/shuttle/countdown/tour.html	0	0	0
/history/apollo/apollo-13/	0	0	0
/history/apollo/apollo-13/images/	0	0	0
/shuttle/missions/sts-71/news/	0	0	0
/shuttle/missions/sts-63/mission-sts-63.html	0	0	0

Rysunek 16: Wartości środków skupienia dla flag stron podczas analizy użytkowników (2 klastry).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	3094	82%
1	688	18%

Rysunek 17: Rozkład wszystkich danych na poszczególne klastry (2 klastry).

Algorytm dla 5 klastrów zatrzymał się po 4 iteracjach, a suma błędów kwadratowych jest równa 6364.

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(3782.0)	(2189.0)	(538.0)	(571.0)	(347.0)	(137.0)
/shuttle/countdown/	0	0	0	1	1	0
/shuttle/missions/sts-71/images/images.html	0	0	0	1	0	0
/shuttle/missions/sts-71/mission-sts-71.html	0	0	0	1	1	0
/	0	0	1	0	0	0
/ksc.html	0	0	0	0	0	0
/shuttle/countdown/liftoff.html	0	0	0	0	0	0
/shuttle/missions/missions.html	0	0	0	0	0	0
/htbin/cdt_main.pl	0	0	0	0	0	0
/shuttle/missions/sts-71/movies/movies.html	0	0	0	0	0	0
/history/apollo/apollo.html	0	0	0	0	0	0
/history/apollo/apollo-13/apollo-13.html	0	0	0	0	0	0
/history/history.html	0	0	0	0	0	0
/shuttle/countdown/countdown.html	0	0	0	0	0	0
/shuttle/technology/sts-newsref/stsref-toc.html	0	0	0	0	0	0
/shuttle/resources/orbiters/atlantis.html	0	0	0	0	0	0
/htbin/cdt_clock.pl	0	0	0	0	0	0
/software/winvn/winvn.html	0	0	0	0	0	1
/shuttle/countdown/lps/fr.html	0	0	0	0	0	0
/shuttle/missions/sts-67/mission-sts-67.html	0	0	0	0	0	0
/history/apollo/apollo-13/apollo-13-info.html	0	0	0	0	0	0
/history/apollo/apollo-11/apollo-11.html	0	0	0	0	0	0
/facilities/lc39a.html	0	0	0	0	0	0
/shuttle/missions/sts-70/mission-sts-70.html	0	0	0	0	0	0
/shuttle/countdown/tour.html	0	0	0	0	0	0
/history/apollo/apollo-13/	0	0	0	0	0	0
/history/apollo/apollo-13/images/	0	0	0	0	0	0
/shuttle/missions/sts-71/news/	0	0	0	0	0	0
/shuttle/missions/sts-63/mission-sts-63.html	0	0	0	0	0	0

Rysunek 18: Wartości środków skupienia dla flag stron podczas analizy użytkowników (5 klastrów).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	2189	58%
1	538	14%
2	571	15%
3	347	9%
4	137	4%

Rysunek 19: Rozkład wszystkich danych na poszczególne klastry (5 klastrów).

Algorytm dla 8 klastrów zatrzymał się po 3 iteracjach, a suma błędów kwadratowych jest równa 6662.

		Cluster#						
Attribute	Full Data	0	1	2	3	4	5	6	7
	(3782.0)	(2418.0)	(378.0)	(198.0)	(23.0)	(131.0)	(299.0)	(146.0)	(189.0)
/shuttle/countdown/	0	1	0	1	1	0	0	1	0
/shuttle/missions/sts-71/images/images.html	0	0	0	0	0	0	1	1	0
/shuttle/missions/sts-71/mission-sts-71.html	0	0	0	1	1	0	0	1	0
/	0	0	1	1	0	0	0	0	0
/ksc.html	0	0	0	0	0	0	0	1	0
/shuttle/countdown/liftoff.html	0	0	0	0	0	0	0	0	1
/shuttle/missions/missions.html	0	0	0	0	0	0	0	1	0
/htbin/cdt_main.pl	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-71/movies/movies.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/apollo-13.html	0	0	0	0	0	0	0	0	0
/history/history.html	0	0	0	0	0	0	0	0	0
/shuttle/countdown/countdown.html	0	0	0	0	0	0	0	0	0
/shuttle/technology/sts-newsref/stsref-toc.html	0	0	0	0	0	0	0	0	0
/shuttle/resources/orbiters/atlantis.html	0	0	0	1	0	0	0	0	0
/htbin/cdt_clock.pl	0	0	0	0	0	0	0	0	0
/software/winvn/winvn.html	0	0	0	0	0	1	0	0	0
/shuttle/countdown/lps/fr.html	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-67/mission-sts-67.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/apollo-13-info.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-11/apollo-11.html	0	0	0	0	0	0	0	0	0
/facilities/lc39a.html	0	0	0	0	1	0	0	0	0
/shuttle/missions/sts-70/mission-sts-70.html	0	0	0	0	0	0	0	0	0
/shuttle/countdown/tour.html	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/	0	0	0	0	0	0	0	0	0
/history/apollo/apollo-13/images/	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-71/news/	0	0	0	0	0	0	0	0	0
/shuttle/missions/sts-63/mission-sts-63.html	0	0	0	0	0	0	0	0	0

Rysunek 20: Wartości środków skupienia dla flag stron podczas analizy użytkowników (8 klastrów).

Numer klastra	Liczba przypisanych próbek	Procent przypisanych próbek
0	2418	64%
1	378	10%
2	198	5%
3	23	1%
4	131	3%
5	299	8%
6	146	4%
7	189	5%

Rysunek 21: Rozkład wszystkich danych na poszczególne klastry (8 klastrów).

W przypadku dwóch klastrów zdecydowanie przeważał klaster (aż 82%), według którego przyporządkowani do niego użytkownicy nie weszli na żadną z najpopularniejszych stron. Dla 5 klastrów nadal większość (58%) użytkowników została przyporządkowana do klastra, który stanowił 82% podczas eksperymentu z 2 klastrami. Pozostałe klastry były o wiele mniej liczne, jednak zaczęły przedstawiać pewne zależności. Zgodnie z klastrem 2 (15% obserwacji) użytkownicy przeglądali strony `/shuttle/countdown/`, `/shuttle/missions/sts-71/images/images.html` oraz `/shuttle/missions/sts-71/mission-sts-71.html`, co może sugerować, że strony są ze sobą powiązane lub obejrzenie jednej z nich zachęca czy odnosi się do drugiej. W przypadku 8 klastrów liczba iteracji spadł o 1, natomiast wzrosła suma błędów kwadratowych. Dodatkowo można zauważyć, że ponownie ponad

połowa obserwacji (64%) dotyczy sytuacji, kiedy użytkownik nie zajrzał na żadną z najpopularniejszych stron. Pozostałe klastry w porównaniu z najliczniejszym stanowią zazwyczaj zaledwie kilka procent obserwacji w każdej z nich. Wśród nich można zauważyć podobne zestawienia stron co w przypadku analizy skupień dla sesji – mowa tu chociażby o 3 klastrze, według którego użytkownicy odwiedzali `/shuttle/countdown/`, `/shuttle/missions/sts-71/mission-sts-71.html` oraz `/facilities/lc39a.html`.

3.4. Wyznaczanie reguł asocjacyjnych dla sesji

Do wyznaczenia reguł asocjacyjnych skorzystano z zaimplementowanego w programie Weka algorytmu Apriori. Opiera się on na dwóch etapach, pierwszym jest generowanie zbiorów częstych, natomiast drugim jest budowanie reguł asocjacyjnych z wygenerowanych wcześniej zbiorów częstych. Zbiór częsty to taki zbiór, który występuje w danej bazie w procencie większym od (ustalonego przez program lub użytkownika) progu wsparcia. Algorytm Apriori wraz z kolejnymi iteracjami generuje coraz dłuższe zbiory częste (wpierw jednoelementowe, następnie dwuelementowe i tak dalej), aż do pewnego momentu, kiedy nie będzie można wygenerować już zbiorów częstych z większą liczbą elementów [1].

Przy pomocy programu Weka oraz algorytmu Apriori zbudowano najlepsze reguły asocjacyjne dla pliku sesji, oddzielnie dla atrybutów odpowiadających stronom (10 reguł) oraz atrybutów typu kategorycznego (8 reguł).

```
Best rules found:

 1. /history/apollo/apollo-13/apollo-13-info.html=0 /history/apollo/apollo-13/=0 2859 ==> /history/apollo/apollo-13/images/=0 2855    <conf:(1)> lift:(1.03) lev:(0.02) [69] conv:(14.74)
 2. /history/apollo/apollo-13/apollo-13-info.html=0 /history/apollo/apollo-13/images/=0 2863 ==> /history/apollo/apollo-13/=0 2855    <conf:(1)> lift:(1.02) lev:(0.01) [43] conv:(5.75)
 3. /history/apollo/apollo-13/apollo-13-info.html=0 2873 ==> /history/apollo/apollo-13/images/=0 2863    <conf:(1)> lift:(1.02) lev:(0.02) [64] conv:(6.73)
 4. /software/winvn/winvn.html=0 /history/apollo/apollo-13/apollo-13-info.html=0 2848 ==> /history/apollo/apollo-13/images/=0 2838    <conf:(1)> lift:(1.02) lev:(0.02) [63] conv:(6.67)
 5. /history/apollo/apollo-13/apollo-13-info.html=0 2873 ==> /history/apollo/apollo-13/=0 2859    <conf:(1)> lift:(1.01) lev:(0.01) [37] conv:(3.46)
 6. /history/apollo/apollo-13/apollo-13-info.html=0 2873 ==> /history/apollo/apollo-13/=0 /history/apollo/apollo-13/images/=0 2855    <conf:(0.99)> lift:(1.03) lev:(0.03) [75] conv:(4.91)
 7. /history/apollo/apollo-13/images/=0 2910 ==> /history/apollo/apollo-13/=0 2890    <conf:(0.99)> lift:(1.01) lev:(0.01) [32] conv:(2.51)
 8. /software/winvn/winvn.html=0 /history/apollo/apollo-13/images/=0 2885 ==> /history/apollo/apollo-13/=0 2865    <conf:(0.99)> lift:(1.01) lev:(0.01) [32] conv:(2.48)
 9. /shuttle/missions/sts-67/mission-sts-67.html=0 2866 ==> /software/winvn/winvn.html=0 2842    <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.96)
10. /history/apollo/apollo-13/=0 2933 ==> /software/winvn/winvn.html=0 2908    <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.94)
```

Rysunek 22: Znalezione reguły asocjacyjne dla atrybutów stron.

```
Best rules found:

 1. session_time_categories=t<2min 1124 ==> session_average_time_per_site_categories=t<2min 1124    <conf:(1)> lift:(1.27) lev:(0.08) [237] conv:(237.44)
 2. session_time_categories=t<2min done_things_categories=x<=4 1072 ==> session_average_time_per_site_categories=t<2min 1072    <conf:(1)> lift:(1.27) lev:(0.08) [226] conv:(226.46)
 3. session_time_categories=2min<=t<5min 661 ==> session_average_time_per_site_categories=t<2min 641    <conf:(0.97)> lift:(1.23) lev:(0.04) [119] conv:(6.65)
 4. session_average_time_per_site_categories=2min<=t<5min 473 ==> session_time_categories=5min<=t<30min 453    <conf:(0.96)> lift:(2.38) lev:(0.09) [262] conv:(13.46)
 5. session_time_categories=2min<=t<5min done_things_categories=x<=4 448 ==> session_average_time_per_site_categories=t<2min 428    <conf:(0.96)> lift:(1.21) lev:(0.02) [74] conv:(4.51)
 6. session_time_categories=t<2min 1124 ==> done_things_categories=x<=4 1072    <conf:(0.95)> lift:(1.48) lev:(0.12) [349] conv:(7.58)
 7. session_time_categories=t<2min session_average_time_per_site_categories=t<2min 1124 ==> done_things_categories=x<=4 1072    <conf:(0.95)> lift:(1.48) lev:(0.12) [349] conv:(7.58)
 8. session_time_categories=t<2min 1124 ==> session_average_time_per_site_categories=t<2min done_things_categories=x<=4 1072    <conf:(0.95)> lift:(1.81) lev:(0.16) [480] conv:(10.05)
```

Rysunek 23: Znalezione reguły asocjacyjne dla atrybutów typu kategorycznego.

Po przeanalizowaniu powyższych reguł asocjacyjnych można zauważyć, że reguły asocjacyjne dla atrybutów związanych z odwiedzeniem stron zawsze dotyczą stron nieodwiedzonych, czyli przyjmują format „Jeśli podczas sesji nie odwiedzono strony <adres1>, to nie odwiedzono także strony <adres2>”. Implikacje tego rodzaju nie niosą zbyt wielu istotnych informacji. W każdym przypadku współczynnik ufności był równy lub bliski wartości 1, co jest najwyższą możliwą wartością dla tego czynnika i oznacza, że jest prawdziwa dla 100% sesji, których dotyczy dana reguła. Można również zauważyć, że w większości znalezionych reguł mowa jest o stronie `/history/apollo/apollo-13/apollo-13-info.html`.

W przypadku reguł związanych z wartościami typu kategorycznego można zauważyć, że większość dotyczy całkowitego czasu sesji lub średniego czasu na jedną stronę. Ze względu na to, że atrybut średniego czasu na jedną stronę jest zależny od pozostałych dwóch atrybutów, część przedstawionych reguł jest intuicyjna. Przykładowo, jeśli czas sesji wyniósł poniżej 2 minut, to średni czas na stronę także wyniesie poniżej 2 minut. Także w tym przypadku nie wszystkie wartości ufności były równe 1, jednak każda z przedstawionych reguł miała współczynnik równy lub większy od wartości 0.95, co jest również bardzo wysokim wynikiem.

Bibliografia

[1] Reguły asocjacyjne, algorytm Apriori, S. H. Nguyen, https://edu.pjwstk.edu.pl/wyklady/adn/scb/rW12.htm, dostęp: 17.11.2020.