Informatyka stosowana, studia dzienne, II st. semestr II

Eksploracja danych internetowych **2020/2021**
Prowadzący: dr inż. Krzysztof Myszkorowski poniedziałek, 10:00

Ćwiczenie 1: Eksploracja użycia na podstawie pliku logów

Autorzy:

Paweł Jeziorski 234066
Karol Podlewski 234106

Spis treści

1. Cel 3

2. Przygotowanie danych 3

2.1. Opis pliku logów 3

2.2. Przetworzenie pliku 3

3. Opis przeprowadzonych badań 6

3.1. Analiza skupień 6

3.2. Analiza podobieństw 6

4. Wyniki analizy klastrowej 7

4.1. Sesje - 3 klastry 7

4.2. Sesje - 6 klastrów 8

4.3. Użytkownicy - 3 klastry 10

4.4. Użytkownicy - 6 klastrów 12

5. Wyniki analizy koszykowej 13

Literatura 15

1. Cel

Celem zadania było przeprowadzenie analizy pliku logów. Plik logów w formacie Common Log Format należało przygotować do analizy w programie Weka poprzez wyodrębnienie i grupowanie użytkowników oraz sesji. Kolejnym krokiem było zbudowanie reguł asocjacyjnych dla tych grup.

W ramach tego ćwiczenia skorzystano z pliku logów `access_log_Aug95`, który jest dostępny do pobrania pod następującym adresem: `ftp://ita.ee.lbl.gov/traces/NASA_access_log_Aug95.gz`.

2. Przygotowanie danych

2.1. Opis pliku logów

Plik `access_log_Aug95` jest typowym plikiem logów stworzonym w formacie Common Log Format.

```
in24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0" 200 1839
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 304 0
slppp6.intermind.net - - [01/Aug/1995:00:00:10 -0400] "GET /history/skylab/skylab.html HTTP/1.0" 200 1687
piweba4y.prodigy.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/launchmedium.gif HTTP/1.0" 200 11853
slppp6.intermind.net - - [01/Aug/1995:00:00:11 -0400] "GET /history/skylab/skylab-small.gif HTTP/1.0" 200 9202
```

Rysunek 1. Pierwsze 10 linii pliku logów

Plik logów zawiera informacje o hoście, nazwie użytkownika podczas uwierzytelniania oraz w systemie klienta, znacznik czasu, pierwszy wiersz nagłówka żądania HTTP, status obsługi tego żądania oraz rozmiar odpowiedzi w bajtach.

Bardzo częstym zjawiskiem jest brak informacji w 2 oraz 3 kolumnie - są one powiązane z uwierzytelnianiem do systemu, co nie jest zawsze wymagane. Właśnie taka sytuacja ma miejsce dla pliku `access_log_Aug95`.

2.2. Przetworzenie pliku

W celu przetworzenia pliku logów stworzono skrypt `preprocessing.py` w języku python, w wersji 3.7.9, z wykorzystaniem biblioteki pandas.

Wczytano 50 000 linii z pominięciem 2 oraz 3 kolumny, które nie zawierały żadnych informacji. Następnie przetworzono wszystkie zmienne wczytane jako ciąg znaków w celu wyodrębnienia z nich kluczowych informacji, takich jak metodę i url żądania HTTP oraz znacznik czasu. Następnie rekordy

ograniczono do tych, które zawierały metodę GET, status odpowiedzi wynosił 200 oraz zapytanie nie było związane z plikami graficznymi (formaty jpg, gif, bmp, xbm, png oraz jpeg). Stworzono też listę najczęściej odwiedzanych stron - liczba odwiedzin musiała być większa niż 0,5%.

Kolejnym krokiem była identyfikacji użytkowników oraz ich w sesji. W przypadku użytkowników skupiono się wyłącznie na liczbie odwiedzonych stron - przygotowany plik w formacie `arff` widoczny jest na rysunku 2.2. Przygotowano także plik bez wartości numerycznych (id użytkownika czy liczby zapytań).

```
@relation .\access_log_Aug95
@attribute user integer
@attribute requests_count integer
@attribute /ksc.html {True, False}
@attribute / {True, False}
@attribute /shuttle/missions/missions.html {True, False}
@attribute /shuttle/countdown/ {True, False}
@attribute /shuttle/missions/sts-69/mission-sts-69.html {True, False}
@attribute /shuttle/missions/sts-70/mission-sts-70.html {True, False}
@attribute /history/history.html {True, False}
@attribute /history/apollo/apollo.html {True, False}
@attribute /history/apollo/apollo-13/apollo-13.html {True, False}
@attribute /images/ {True, False}
@attribute /shuttle/technology/sts-newsref/stsref-toc.html {True, False}
@attribute /software/winvn/winvn.html {True, False}
@attribute /htbin/cdt_main.pl {True, False}
@attribute /shuttle/countdown/liftoff.html {True, False}
@attribute /shuttle/missions/sts-71/mission-sts-71.html {True, False}
@attribute /shuttle/missions/sts-70/images/images.html {True, False}
@attribute /shuttle/technology/sts-newsref/sts_asm.html {True, False}
@attribute /shuttle/countdown/countdown.html {True, False}
@attribute /facilities/lc39a.html {True, False}
@attribute /shuttle/missions/sts-71/images/images.html {True, False}
@attribute /elv/elvpage.htm {True, False}
@attribute /history/apollo/apollo-11/apollo-11.html {True, False}
@attribute /shuttle/missions/sts-70/movies/movies.html {True, False}
@attribute /htbin/wais.pl {True, False}
@attribute /shuttle/missions/sts-71/movies/movies.html {True, False}
@attribute /shuttle/resources/orbiters/endeavour.html {True, False}
@attribute /whats-new.html {True, False}
@attribute /htbin/cdt_clock.pl {True, False}
@data
0,42,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
1,3,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
2,11,True,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
3,1,False,True,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
4,1,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
5,1,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
```

Rysunek 2. Wyodrębnieni użytkownicy

Równolegle wyodrębniono też sesje, które zawierały następujące flagi:

— Czas sesji w sekundach
— Liczba działań w czasie sesji
— Przeciętny czas na stronę
— Zmienne flagowe dla najpopularniejszych stron

Sesja była uznawana za zakończoną, jeżeli po 10 minutach od ostatniego zapytania nie pojawiło się kolejne. Przeciętny czas spędzony na jednej stronie został obliczony ze wzoru:

$$\text{Przeciętny czas na stronę} = \frac{\text{czas sesji}}{\text{liczba działań w czasie sesji} - 1}$$

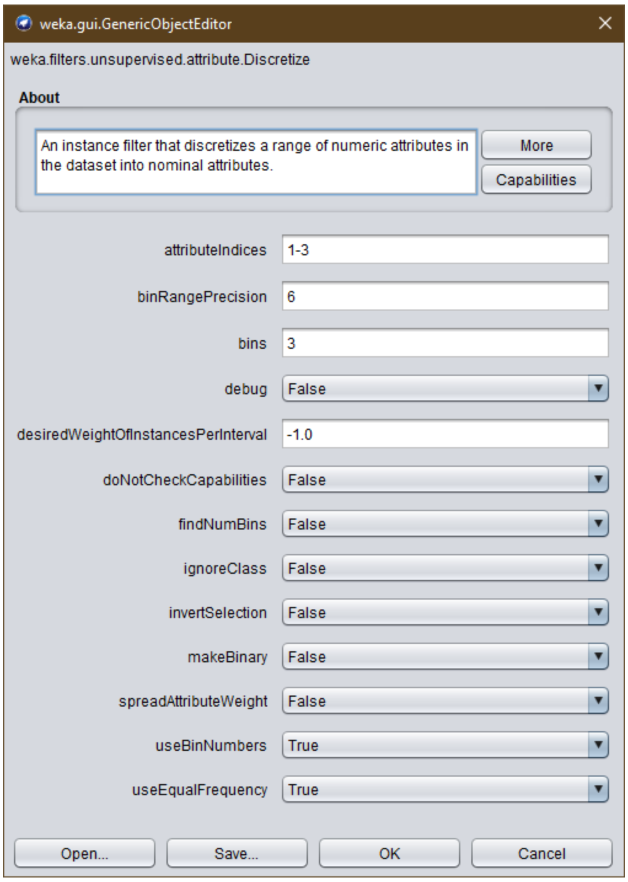
Nie wiadomo jak dużo czasu spędził na ostatniej stronie, dlatego nie jest ona uwzględniania przy obliczaniu czasu.

Przygotowany plik `arff` przedstawiono na rysunku 2.2. Przygotowano też wersję, która zawierała tylko i wyłącznie flagi dla odwiedzonych stron.

```
@relation .\access_log_Aug95
@attribute duration real
@attribute requests_count integer
@attribute avg_request_duration real
@attribute /ksc.html {True, False}
@attribute / {True, False}
@attribute /shuttle/missions/missions.html {True, False}
@attribute /shuttle/countdown/ {True, False}
@attribute /shuttle/missions/sts-69/mission-sts-69.html {True, False}
@attribute /shuttle/missions/sts-70/mission-sts-70.html {True, False}
@attribute /history/history.html {True, False}
@attribute /history/apollo/apollo.html {True, False}
@attribute /history/apollo/apollo-13/apollo-13.html {True, False}
@attribute /images/ {True, False}
@attribute /shuttle/technology/sts-newsref/stsref-toc.html {True, False}
@attribute /software/winvn/winvn.html {True, False}
@attribute /htbin/cdt_main.pl {True, False}
@attribute /shuttle/countdown/liftoff.html {True, False}
@attribute /shuttle/missions/sts-71/mission-sts-71.html {True, False}
@attribute /shuttle/missions/sts-70/images/images.html {True, False}
@attribute /shuttle/technology/sts-newsref/sts_asm.html {True, False}
@attribute /shuttle/countdown/countdown.html {True, False}
@attribute /facilities/lc39a.html {True, False}
@attribute /shuttle/missions/sts-71/images/images.html {True, False}
@attribute /elv/elvpage.htm {True, False}
@attribute /history/apollo/apollo-11/apollo-11.html {True, False}
@attribute /shuttle/missions/sts-70/movies/movies.html {True, False}
@attribute /htbin/wais.pl {True, False}
@attribute /shuttle/missions/sts-71/movies/movies.html {True, False}
@attribute /shuttle/resources/orbiters/endeavour.html {True, False}
@attribute /whats-new.html {True, False}
@attribute /htbin/cdt_clock.pl {True, False}
@data
821.0,16,54.733333333333334,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
938.0,24,40.78260869565217,False,False,True,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
1245.0,9,155.625,False,False,False,False,True,False,False,True,True,False,False,False,False,False,False,False,False,False,True,False,False,False,False,False,False,True,False,False
1112.0,18,65.41176470588235,False,False,False,False,False,True,False,False,False,False,True,False,False,False,True,False,True,False,True,True,False,False,False,False,False,True,False,False
366.0,3,183.0,False,False,False,False,False,False,False,False,False,True,False,False,True,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False
```

Rysunek 3. Wyodrębnione sesje

Dyskretyzacji sesji dokonano w programie Weka, za pomocą odpowiedniego filtru - parametry przedstawiono na rysunku 2.2. W celu stworzenia równolicznych grup wartość parametru `useEqualFrequency` musi być ustawiona na wartość `True`, w przypadku grup nierównolicznych (domyślna metoda) wartość ta musi wynosić `False`.



Rysunek 4. Parametry filtru dokonującego dyksretyzacji

3. Opis przeprowadzonych badań

3.1. Analiza skupień

Analizę skupień przeprowadzono przy pomocy metody k-średnich. Tworzy ona `k` klastrów, w losowy sposób przypisując do nich po jednej obserwacji. Następnie do każdego klastra przypisuje kolejne obserwacje na podstawie ich odległości od środka skupienia, po czym wybierane jest nowe skupienie, którego współrzędne znajdują się najbliższej średniej arytmetycznej współrzędnych wszystkich punktów należących do tego skupienia - jest to nowy środek klastra. Algorytm działa tak długo aż dojdzie do określonej liczby iteracji bądź między kolejnymi iteracjami nie będzie zmian wśród skupień.

Algorytm uruchomiono z 3 oraz 6 klastrami dla plików sesji z atrybutami numerycznymi oraz flagami odwiedzonych stron oraz dla użytkowników z flagami odwiedzonych stron.

3.2. Analiza podobieństw

W celu odnalezienia reguł asocjacyjnych wykorzystano algorytm Apriori. Jest to algorytm iteracyjny, który w kolejnych krokach znajduje zbiory częste o rosnących rozmiarach. Na początku algorytm wyodrębnia wszystkie zbiory jednoelementowe o odpowiednim wsparciu. Następnie, w oparciu o te zbiory tworzy zbiory kandydujące dwuelementowe, sprawdzając ich wsparcie - te o odpowiednio wysokiej wartości wykorzystane będą przy tworzeniu zbiorów kandydujących trójelementowych. Każdy kolejny krok generuje zbiory kandydujące o rozmiarze o jeden większym tak długo, aż nie da się wygenerować kolejnych zbiorów kandydujących [1].

Algorytm uruchomiono dla równolicznych zbiorów oraz nierównolicznych zbiorów przygotowanych przez dyskretyzację na pliku sesji.

4. Wyniki analizy klastrowej

4.1. Sesje - 3 klastry

Dla danych z wyłącznie flagami stron algorytm potrzebował 3 iteracje, a suma błędów kwadratowych wyniosła 1246.

Tabela 1. Wybrane centroidy dla 3 klastrów z flagami stron podczas analizy sesji

Attribute	Full Data	0	1	2
/ksc.html	False	False	False	False
/	False	False	False	False
/shuttle/missions/missions.html	False	False	False	False
/shuttle/countdown/	False	False	False	**True**
/shuttle/missions/sts-69/mission-sts-69.html	False	False	False	False
/shuttle/missions/sts-70/mission-sts-70.html	False	False	False	False
/history/history.html	False	**True**	False	False
/history/apollo/apollo.html	False	**True**	False	False
/history/apollo/apollo-13/apollo-13.html	False	False	False	False
/images/	False	False	False	False
/shuttle/technology/sts-newsref/stsref-toc.html	False	False	False	False
/software/winvn/winvn.html	False	False	False	False
/htbin/cdt_main.pl	False	False	False	False
/shuttle/countdown/liftoff.html	False	False	False	False
/shuttle/missions/sts-71/mission-sts-71.html	False	False	False	False
/shuttle/missions/sts-70/images/images.html	False	False	False	False
/shuttle/technology/sts-newsref/sts_asm.html	False	False	False	False
/shuttle/countdown/countdown.html	False	False	False	False
/facilities/lc39a.html	False	False	False	False
/shuttle/missions/sts-71/images/images.html	False	False	False	False
/elv/elvpage.htm	False	False	False	False
/history/apollo/apollo-11/apollo-11.html	False	False	False	False
/shuttle/missions/sts-70/movies/movies.html	False	False	False	False
/htbin/wais.pl	False	False	False	False
/shuttle/missions/sts-71/movies/movies.html	False	False	False	False
/shuttle/resources/orbiters/endeavour.html	False	False	False	False
/whats-new.html	False	False	False	False
/htbin/cdt_clock.pl	False	False	False	False

Tabela 2. Przypisanie obserwacji dla 3 klastrów z flagami stron podczas analizy sesji

Klaster	Liczba obserwacji	Procent obserwacji
0	82	11%
1	**560**	**75%**
2	102	14%

Dla atrybutów numerycznych algorytm potrzebował 12 iteracje, a suma błędów kwadratowych wyniosła 11.72.

Tabela 3. Wybrane centroidy dla 3 klastrów z atrybutami numerycznymi podczas analizy sesji

Attribute	Full Data	0	1	2
duration	437.4543	179.6325	868.1154	517.1538
requests_count	5.8159	4.7279	9.0726	2.4505
avg_request_duration	123.9246	49.8463	158.3098	376.5912

Tabela 4. Przypisanie obserwacji dla 3 klastrów z atrybutami numerycznymi podczas analizy sesji

Klaster	Liczba obserwacji	Procent obserwacji
0	**419**	**56%**
1	234	31%
2	91	12%

Podział na klastry wyraźnie różnił się - w przypadku danych z flagami odwiedzonych stron powstał jeden klaster, który zbierał aż $\frac{3}{4}$ wszystkich sesji. W przypadku atrybutów numerycznych największy klaster był dużo mniej liczby, ale wciąż zawierał w sobie ponad 50% wszystkich sesji. Niezależnie od wybranego zbioru atrybutów do analizy skupień, najmniejszy klaster zbierał trochę więcej niż 10% badanej grupy.

Z analizy flag wynika, że kiedy w trakcie sesji odwiedzano stronę `/history/history.html`, odwiedzono też stronę `/history/apollo/apollo.html`.

Dużo większe zróżnicowanie widoczne jest przy analizie atrybutów numerycznych, gdzie każdy środek klastra różnił się w znaczącym stopniu od innych centroidów. Najwięcej sesji zostało przypisanych do 1 klastra, który cechował się najkrótszą długością sesji oraz najmniejszym średnim czasem spędzonym na stronie, ale już nie najmniejszą liczbą zapytań - wciąż była to jednak wartość poniżej średniej dla całego zbioru.

4.2. Sesje - 6 klastrów

Dla 6 klastrów z flagami stron potrzeba było 3 iteracji, a suma błędów kwadratowych wyniosła 1017.

Tabela 5. Wybrane centroidy dla 6 klastrów z flagami stron podczas analizy sesji

Attribute	Full Data	0	1	2	3	4	5
/ksc.html	False	False	False	False	**True**	False	**True**
/	False	False	False	False	False	False	False
/shuttle/missions/missions.html	False	False	False	False	False	False	False
/shuttle/countdown/	False	False	False	**True**	False	False	False
/shuttle/missions/sts-69/mission-sts-69.html	False	False	False	False	False	False	False
/shuttle/missions/sts-70/mission-sts-70.html	False	False	False	False	False	False	False
/history/history.html	False	**True**	False	False	False	False	False
/history/apollo/apollo.html	False	**True**	False	False	False	False	False
/history/apollo/apollo-13/apollo-13.html	False	False	False	False	False	False	False
/images/	False	False	False	False	False	False	False
/shuttle/technology/sts-newsref/stsref-toc.html	False	False	False	False	False	False	False
/software/winvn/winvn.html	False	False	False	False	False	False	False
/htbin/cdt_main.pl	False	False	False	False	False	**True**	False
/shuttle/countdown/liftoff.html	False	False	False	False	False	False	False
/shuttle/missions/sts-71/mission-sts-71.html	False	False	False	False	False	False	False
/shuttle/missions/sts-70/images/images.html	False	False	False	False	False	False	False
/shuttle/technology/sts-newsref/sts_asm.html	False	False	False	False	False	False	False
/shuttle/countdown/countdown.html	False	False	False	False	False	False	False
/facilities/lc39a.html	False	False	False	False	False	False	False
/shuttle/missions/sts-71/images/images.html	False	False	False	False	False	False	False
/elv/elvpage.htm	False	False	False	False	False	False	False
/history/apollo/apollo-11/apollo-11.html	False	False	False	False	False	False	False
/shuttle/missions/sts-70/movies/movies.html	False	False	False	False	False	False	False
/htbin/wais.pl	False	False	False	False	False	False	**True**
/shuttle/missions/sts-71/movies/movies.html	False	False	False	False	False	False	False
/shuttle/resources/orbiters/endeavour.html	False	False	False	False	False	False	False
/whats-new.html	False	False	False	False	False	False	False
/htbin/cdt_clock.pl	False	False	False	False	False	False	False

Tabela 6. Przypisanie obserwacji dla 6 klastrów z flagami stron podczas analizy sesji

Klaster	Liczba obserwacji	Procent obserwacji
0	72	10%
1	**355**	**48%**
2	99	13%
3	191	26%
4	13	2%
5	14	2%

Dla atrybutów numerycznych algorytm dokonał klasteryzacji w 16 iteracji, a suma błędów kwadratowych wyniosła 5.73.

Tabela 7. Wybrane centroidy dla 6 klastrów z flagami numerycznymi podczas analizy sesji

Attribute	Full Data	0	1	2	3	4	5
duration	437.4543	466.3737	2519.069	571.4189	578.5113	499.8718	89.8925
requests_count	5.8159	7.1105	35.4828	3	4.4662	2.1026	3.7599
avg_request_duration	123.9246	89.7212	118.6991	294.6956	171.0568	458.8205	33.1848

Tabela 8. Przypisanie obserwacji dla 6 klastrów z atrybutami numerycznymi podczas analizy sesji

Klaster	Liczba obserwacji	Procent obserwacji
0	190	26%
1	29	4%
2	74	10%
3	133	18%
4	39	5%
5	**279**	**38%**

Podobnie jak dla 3 klastrów, algorytm dla 6 klastrów przy różnych danych nie wyznaczył podobnych liczbowo skupień - większy klaster został utworzony dla danych z flagami stron, skupiał on prawie połowę badanych sesji. Dla atrybutów numerycznych najmniejsze klastry były jednak ponad dwukrotnie większe od najmniejszych klastrów dla danych zawierających flagi sesji.

Wszystkie wyznaczone klastry przez algorytm dla 3 skupień pojawiają się w grupie zawierającej 6 skupień. W aż dwóch nowych skupieniach pojawia się najpopularniejsza strona `ksc.html`.

Ponownie najliczniejszym skupieniem dla atrybutów numerycznych jest to, którego długość sesji oraz średni czas na stronie są najmniejsze. Do klastra 1 algorytm przypisał sesje które cechowały się bardzo wysokimi wartościami - jest to prawie 30 rekordów, które znacznie podnoszą średnią wartość wszystkich atrybutów numerycznych - dzięki tej obserwacji wysoka liczność wśród klastra zawierającego dużo niższe od średnich wartości w ogóle nie dziwi.

4.3. Użytkownicy - 3 klastry

Do wyznaczenia 3 klastrów z flagami stron dla użytkowników algorytm potrzebował 9 iteracji, a suma błędów kwadratowych wyniosła 4610,76.

Tabela 9. Wybrane centroidy dla 3 klastrów z flagami stron podczas analizy użytkowników

Attribute	Full Data	0	1	2
/ksc.html	False	False	False	False
/	False	False	False	False
/shuttle/missions/missions.html	False	False	**True**	False
/shuttle/countdown/	False	False	False	**True**
/shuttle/missions/sts-69/mission-sts-69.html	False	False	False	False
/shuttle/missions/sts-70/mission-sts-70.html	False	False	False	False
/history/history.html	False	False	False	False
/history/apollo/apollo.html	False	False	False	False
/history/apollo/apollo-13/apollo-13.html	False	False	False	False
/images/	False	False	False	False
/shuttle/technology/sts-newsref/stsref-toc.html	False	False	False	False
/software/winvn/winvn.html	False	False	False	False
/htbin/cdt_main.pl	False	False	False	False
/shuttle/countdown/liftoff.html	False	False	False	False
/shuttle/missions/sts-71/mission-sts-71.html	False	False	False	False
/shuttle/missions/sts-70/images/images.html	False	False	False	False
/shuttle/technology/sts-newsref/sts_asm.html	False	False	False	False
/shuttle/countdown/countdown.html	False	False	False	False
/facilities/lc39a.html	False	False	False	False
/shuttle/missions/sts-71/images/images.html	False	False	False	False
/elv/elvpage.htm	False	False	False	False
/history/apollo/apollo-11/apollo-11.html	False	False	False	False
/shuttle/missions/sts-70/movies/movies.html	False	False	False	False
/htbin/wais.pl	False	False	False	False
/shuttle/missions/sts-71/movies/movies.html	False	False	False	False
/shuttle/resources/orbiters/endeavour.html	False	False	False	False
/whats-new.html	False	False	False	False
/htbin/cdt_clock.pl	False	False	False	False

Tabela 10. Przypisanie obserwacji dla 3 klastrów z flagami stron podczas analizy użytkowników

Klaster	Liczba obserwacji	Procent obserwacji
0	**2810**	**79%**
1	406	11%
2	326	9%

Użytkownicy różnili się od siebie dużo bardziej niż sesje, na co wskazuje ponad 3 razy większy średni błąd kwadratowy, co będzie widoczne także dla 6 klastrów. Większość użytkowników, bo prawie $\frac{4}{5}$, zostało przypisanych do klastra, który nie cechuje się odwiedzeniem żadnej z najpopularniejszych stron. Klastry te różnią się także od tych utworzonych na podstawie sesji, co może wskazywać na większą popularność konkretnych stron dla różnej grupy użytkowników.

4.4. Użytkownicy - 6 klastrów

W przypadku 6 klastrów potrzeba było 3 iteracji, a suma błędów kwadratowych wyniosła 3493.

Tabela 11. Wybrane centroidy dla 6 klastrów z flagami stron podczas analizy użytkowników

Attribute	Full Data	0	1	2	3	4	5
/ksc.html	False	False	False	False	False	False	**True**
/	False	False	False	False	False	False	False
/shuttle/missions/missions.html	False	False	**True**	False	**True**	False	False
/shuttle/countdown/	False	False	False	**True**	False	False	False
/shuttle/missions/sts-69/mission-sts-69.html	False	False	False	False	False	False	False
/shuttle/missions/sts-70/mission-sts-70.html	False	False	False	False	False	False	False
/history/history.html	False	False	False	False	False	False	False
/history/apollo/apollo.html	False	False	False	False	False	False	False
/history/apollo/apollo-13/apollo-13.html	False	False	False	False	False	False	False
/images/	False	False	False	False	False	False	False
/shuttle/technology/sts-newsref/stsref-toc.html	False	False	False	False	False	False	False
/software/winvn/winvn.html	False	False	False	False	False	False	False
/htbin/cdt_main.pl	False	False	False	False	False	False	False
/shuttle/countdown/liftoff.html	False	False	False	False	False	False	False
/shuttle/missions/sts-71/mission-sts-71.html	False	False	**True**	False	False	False	False
/shuttle/missions/sts-70/images/images.html	False	False	False	False	False	False	False
/shuttle/technology/sts-newsref/sts_asm.html	False	False	False	False	False	False	False
/shuttle/countdown/countdown.html	False	False	False	False	False	False	False
/facilities/lc39a.html	False	False	False	False	False	False	False
/shuttle/missions/sts-71/images/images.html	False	False	False	False	False	False	False
/elv/elvpage.htm	False	False	False	False	False	**True**	False
/history/apollo/apollo-11/apollo-11.html	False	False	False	False	False	False	False
/shuttle/missions/sts-70/movies/movies.html	False	False	False	False	False	False	False
/htbin/wais.pl	False	False	False	False	False	False	False
/shuttle/missions/sts-71/movies/movies.html	False	False	False	False	False	False	False
/shuttle/resources/orbiters/endeavour.html	False	False	False	False	False	False	False
/whats-new.html	False	False	False	False	False	False	False
/htbin/cdt_clock.pl	False	False	False	False	False	False	False

Tabela 12. Przypisanie obserwacji dla 6 klastrów z flagami stron podczas analizy użytkowników

Klaster	Liczba obserwacji	Procent obserwacji
0	**2051**	**58%**
1	31	1%
2	364	10%
3	337	10%
4	48	1%
5	711	20%

Podobnie jak przy 3 klastrach, tak w przypadku 6 klastrów największy klaster cechuje się nie odwiedzeniem żadnej z najpopularniejszych stron. Aż 20% użytkowników zostało przypisanych do klastra, który cechuje się odwiedzeniem strony `ksc.html` - taki klaster jednak nie powstał dla 3 skupień.

5. Wyniki analizy koszykowej

Tabela 13. Parametry przebiegu algorytmu Apriori dla atrybutów numerycznych

Paramter	Wartość
Minimum support	0.1 (74 instances)
Minimum metric	0.9
Number of cycles performed	18
Size of set of large itemsets L(1)	4
Size of set of large itemsets L(2)	5
Size of set of large itemsets L(3)	2

Tabela 14. Liczba obserwacji przypisanych do koszyków dla atrybutów numerycznych

Atrybut	B1of3	B2of3	B3of3
Czas sesji	726	15	3
Liczba odwiedzonych stron	741	2	1
Średni czas na stronę	602	110	32

Tabela 15. Parametry przebiegu algorytmu Apriori dla flag stron

Paramter	Wartość
Minimum support	0.95 (707 instances)
Minimum metric	0.9
Number of cycles performed	1
Size of set of large itemsets L(1)	18
Size of set of large itemsets L(2)	49
Size of set of large itemsets L(3)	20

Wyznaczone reguły asocjacyjne są widoczne w tabeli 16.

Wykryte reguły asocjacyjne są w większości powiązane z najliczniejszymi koszykami - szczególnie w przypadku analizowania czasu sesji oraz liczby odwiedzonych stron, gdzie największe koszyki pokrywają ponad 97% wszystkich sesji. Algorytm odnalazł kilka reguł powiązanych z "przeciętnym" średnim czasem poświęconym na stronę - takie rekordy w większości wciąż będą przypisane do koszyka zawierającego najmniejsze wartości w atrybutach powiązanych z długością sesji czy liczbą stron. Najmniejszy wskaźnik pewności wyniósł `0,97`.

Reguły wyznaczone dla stron zawsze dotyczą nie odwiedzenia danych stron, na przykład "*Użytkownik który nie odwiedził w danej sesji strony X, nie odwiedził także strony Y*". Wskaźnik pewności waha się w przedziale `<0,99;1>`, co jest bardzo wysoką wartością, ale sama przydatność tych reguł jest znikoma. W dodatku aż 9 na 10 reguł dotyczy nieodwiedzenia strony `/software/winvn/winvn.html`.

Tabela 16. Znalezione reguły asocjacyjne

Jeżeli	To	Parametry
	Atrybuty numeryczne	
avg_request_duration='B2of3' 110	requests_count='B1of3' 110	conf:(1) lift:(1) lev:(0) [0] conv:(0.44)
duration='B1of3' avg_request_duration='B2of3' 109	requests_count='B1of3' 109	conf:(1) lift:(1) lev:(0) [0] conv:(0.44)
duration='B1of3' 726	requests_count='B1of3' 725	conf:(1) lift:(1) lev:(0) [1] conv:(1.46)
duration='B1of3' avg_request_duration='B1of3' 585	requests_count='B1of3' 584	conf:(1) lift:(1) lev:(0) [1] conv:(1.18)
avg_request_duration='B1of3' 602	requests_count='B1of3' 599	conf:(1) lift:(1) lev:(-0) [0] conv:(0.61)
avg_request_duration='B2of3' 110	duration='B1of3' 109	conf:(0.99) lift:(1.02) lev:(0) [1] conv:(1.33)
requests_count='B1of3' avg_request_duration='B2of3' 110	duration='B1of3' 109	conf:(0.99) lift:(1.02) lev:(0) [1] conv:(1.33)
avg_request_duration='B2of3' 110	duration='B1of3' requests_count='B1of3' 109	conf:(0.99) lift:(1.02) lev:(0) [1] conv:(1.4)
requests_count='B1of3' 741	duration='B1of3' 725	conf:(0.98) lift:(1) lev:(0) [1] conv:(1.05)
requests_count='B1of3' avg_request_duration='B1of3' 599	duration='B1of3' 584	conf:(0.97) lift:(1) lev:(-0) [0] conv:(0.91)
	Flagi stron	
/shuttle/countdown/liftoff.html=False 709	/htbin/cdt_clock.pl=False 707	conf:(1) lift:(1.01) lev:(0.01) [6] conv:(2.86)
/htbin/cdt_clock.pl=False 735	/software/winvn/winvn.html=False 729	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.85)
/shuttle/resources/orbiters/endeavour.html=False 732	/software/winvn/winvn.html=False 726	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.84)
/whats-new.html=False 730	/software/winvn/winvn.html=False 724	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.84)
/history/apollo/apollo-11/apollo-11.html=False 729	/software/winvn/winvn.html=False 723	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.84)
/shuttle/countdown/countdown.html=False 728	/software/winvn/winvn.html=False 722	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.84)
/shuttle/missions/sts-71/movies/movies.html=False 727	/software/winvn/winvn.html=False 721	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.84)
/images/=False 724	/software/winvn/winvn.html=False 718	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.83)
/elv/elvpage.htm=False 723	/software/winvn/winvn.html=False 717	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.83)
/shuttle/resources/orbiters/endeavour.html=False /htbin/cdt_clock.pl=False 723	/software/winvn/winvn.html=False 717	conf:(0.99) lift:(1) lev:(-0) [0] conv:(0.83)

Literatura

[1] *Algorytmy odkrywania binarnych reguł asocjacyjnych,* A. Starczewski, A. Krzyżak, `http://wazniak.mimuw.edu.pl/images/c/c3/ED-4.2-m03-1.0.pdf` [dostęp: 05.11.2020]