Informatyka stosowana, studia dzienne, II st. semestr II

Eksploracja danych internetowych **2020/2021**

Prowadzący: dr inż. Krzysztof Myszkorowski poniedziałek, 10:00

Ćwiczenie 3: System rekomendacji

Autorzy:

Paweł Jeziorski 234066

Karol Podlewski 234106

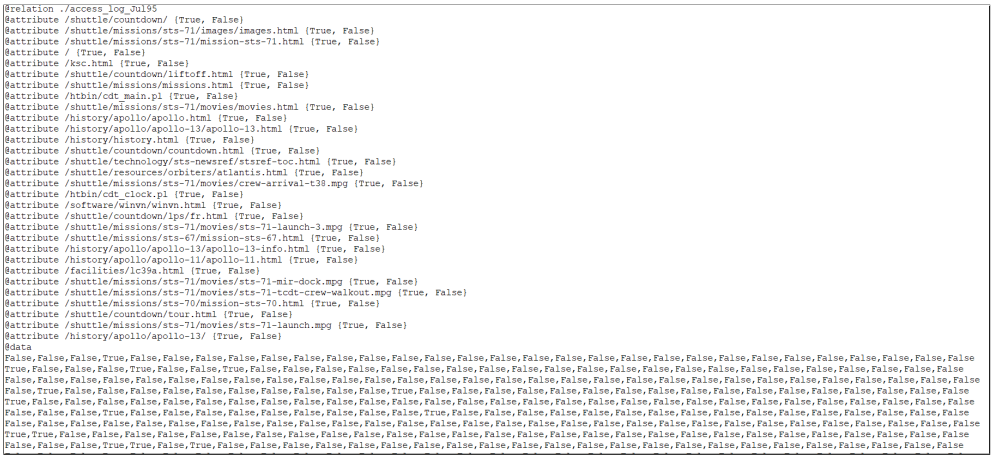
Spis treści

1. Cel

Celem zadania było zarekomendowanie losowo utworzonemu użytkownikowi stron, które powinien odwiedzić. System znajduje najbardziej podobną mu grupę na podstawie aktywności innych użytkowników. Następnie powinien wskazać strony na które wchodziła większość członków tej grupy.

2. Opis metody

Do realizacji zadania wymagany jest zbiór danych zawierający użytkowników i odwiedzone przez nich strony (Źródło danych: `ftp://ita.ee.lbl.gov/traces/NASA_access_log_Jul95.gz`). Taka reprezentacja została przygotowana w ramach realizacji zadanie pierwszego i będzie stanowiła podstawę dla systemu rekomendacji.



```
@relation ./access_log_Jul95
@attribute /shuttle/countdown/ {True, False}
@attribute /shuttle/missions/sts-71/images/images.html {True, False}
@attribute /shuttle/missions/sts-71/mission-sts-71.html {True, False}
@attribute / {True, False}
@attribute /ksc.html {True, False}
@attribute /shuttle/countdown/liftoff.html {True, False}
@attribute /shuttle/missions/missions.html {True, False}
@attribute /htbin/cdt_main.pl {True, False}
@attribute /shuttle/missions/sts-71/movies/movies.html {True, False}
@attribute /history/apollo/apollo.html {True, False}
@attribute /history/apollo/apollo-13/apollo-13.html {True, False}
@attribute /history/history.html {True, False}
@attribute /shuttle/countdown/countdown.html {True, False}
@attribute /shuttle/technology/sts-newsref/stsref-toc.html {True, False}
@attribute /shuttle/resources/orbiters/atlantis.html {True, False}
@attribute /shuttle/missions/sts-71/movies/crew-arrival-t38.mpg {True, False}
@attribute /htbin/cdt_clock.pl {True, False}
@attribute /software/winvn/winvn.html {True, False}
@attribute /shuttle/countdown/lps/fr.html {True, False}
@attribute /shuttle/missions/sts-71/movies/sts-71-launch-3.mpg {True, False}
@attribute /shuttle/missions/sts-67/mission-sts-67.html {True, False}
@attribute /history/apollo/apollo-13/apollo-13-info.html {True, False}
@attribute /history/apollo/apollo-11/apollo-11.html {True, False}
@attribute /facilities/lc39a.html {True, False}
@attribute /shuttle/missions/sts-71/movies/sts-71-mir-dock.mpg {True, False}
@attribute /shuttle/missions/sts-71/movies/sts-71-tcdt-crew-walkout.mpg {True, False}
@attribute /shuttle/missions/sts-70/mission-sts-70.html {True, False}
@attribute /shuttle/countdown/tour.html {True, False}
@attribute /shuttle/missions/sts-71/movies/sts-71-launch.mpg {True, False}
@attribute /history/apollo/apollo-13/ {True, False}
@data
False,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,False,False,False
True,False,False,False,True,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,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,False,False,False
False,True,False,False,False,False,False,False,False,False,False,False,True,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,False,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,False,False,False,False,False,False
False,False,False,True,False,False,False,False,False,False,False,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,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
True,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,False,False,False
False,False,False,True,True,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
```

Rysunek 1. Fragment pliku arff zawierającego użytkowników i odwiedzone przez nich strony

Dane o użytkownikach dzielone są na klastry według odwiedzonych przez nich stron. Do klasteryzacji wykorzystany zostanie program Weka, a algorytmem odpowiedzialnym za analizę skupień będzie metoda *k średnich* [2]. Wyniki pracy programu stanowią podstawę do wyznaczenia podobieństw i w konsekwencji rekomendacji.

/shuttle/countdown/	False	False	False	False	True	True
/shuttle/missions/sts-71/images/images.html	False	False	False	False	False	False
/shuttle/missions/sts-71/mission-sts-71.html	False	False	False	False	False	True
/	False	False	False	False	False	False
/ksc.html	False	False	False	False	False	False
/shuttle/countdown/liftoff.html	False	False	False	False	False	False
/shuttle/missions/missions.html	False	False	False	False	False	False
/htbin/cdt_main.pl	False	True	False	False	False	False
/shuttle/missions/sts-71/movies/movies.html	False	False	False	False	False	False
/history/apollo/apollo.html	False	False	False	False	False	False
/history/apollo/apollo-13/apollo-13.html	False	False	False	False	False	False
/history/history.html	False	False	False	False	False	False
/shuttle/countdown/countdown.html	False	False	False	False	False	False
/shuttle/technology/sts-newsref/stsref-toc.html	False	False	False	False	False	False
/shuttle/resources/orbiters/atlantis.html	False	False	False	False	False	False
/shuttle/missions/sts-71/movies/crew-arrival-t38.mpg	False	False	False	False	False	False
/htbin/cdt_clock.pl	False	False	False	False	False	False
/software/winvn/winvn.html	False	False	False	True	False	False
/shuttle/countdown/lps/fr.html	False	False	False	False	False	False
/shuttle/missions/sts-71/movies/sts-71-launch-3.mpg	False	False	False	False	False	False
/shuttle/missions/sts-67/mission-sts-67.html	False	False	False	False	False	False
/history/apollo/apollo-13/apollo-13-info.html	False	False	False	False	False	False
/history/apollo/apollo-11/apollo-11.html	False	False	False	False	False	False
/facilities/lc39a.html	False	False	False	False	False	False
/shuttle/missions/sts-71/movies/sts-71-mir-dock.mpg	False	False	False	False	False	False
/shuttle/missions/sts-71/movies/sts-71-tcdt-crew-walkout.mpg	False	False	False	False	False	False
/shuttle/missions/sts-70/mission-sts-70.html	False	False	False	False	False	False
/shuttle/countdown/tour.html	False	False	False	False	False	False
/shuttle/missions/sts-71/movies/sts-71-launch.mpg	False	False	False	False	False	False
/history/apollo/apollo-13/	False	False	False	False	False	False

Rysunek 2. Fragment pliku zawierający wyniki klasteryzacji metodą k średnich za pomocą programu Weka (*numCluster*=5). Stanowi to element wejściowy dla programu rekomendacyjnego.

Nowy użytkownik tworzony jest jako wektor zawierający wartości binarne, które reprezentują odwiedzone przez niego poszczególne strony. Każdemu atrybutowi odpowiadającemu jednej stronie przypisana jest losowo wartość *True* lub *False*.

Podobieństwo użytkownika do poszczególnych, utworzonych wcześniej grup, określane jest na podstawie *współczynnika podobieństwa Jaccarda* [1]. Zdefiniowany jest on jako iloraz liczby elementów występujących w obu zbiorach oraz liczby wszystkich elementów:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

gdzie w przypadku proponowanego rozwiązania A to zbiór zawierający strony odwiedzone przez losowego użytkownika, a B to zbiór zawierający strony odwiedzone przez większość użytkowników danego klastra. Im wyższa wartość współczynnika, tym zbiory są bardziej podobne (wartość równa 1 oznacza, że zbiory są identyczne).

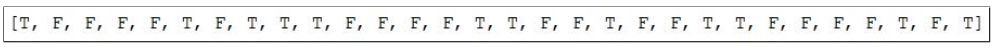
Posiadając wartości podobieństwa użytkowania do wszystkich badanych klastrów, należy wybrać ten z najwyższą wartością, czyli najbardziej podobny, i wskazać go jako wzór do rekomendacji.

Na jego podstawie wszystkie nieodwiedzone przez użytkownika strony są mu przekazywane jako rekomendacje.

3. Wyniki

Rezultaty zostały podzielone na kilka przypadków, które różnią się liczbą klastrów użytych w algorytmie klasteryzacji. Każdy z nich opisany jest za pomocą tabeli centroidów reprezentujących klaster oraz jego charakterystykę.

W celu zapewnienia spójności i możliwości porównania poszczególnych przypadków, wygenerowano losowego użytkownika, który będzie brał udział w każdym badanym eksperymencie.



Rysunek 3. Losowo wygenerowany użytkownik

3.1. Liczba klastrów - 5

3.1.1. Informacje dotyczące klastrów

Tabela 1. Wynik analizy skupień dla pięciu klastrów

Klaster	Wartość
1	160 (4%)
2	2252 (58%)
3	133 (3%)
4	856 (22%)
5	483 (12%)

Tabela 2. Wybrane centroidy dla 5 klastrów z flagami stron użytkowników

Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5
False	False	False	**True**	**True**
False	False	False	False	False
False	False	False	False	**True**
False	False	False	False	False
False	False	False	False	False
False	False	False	False	False
False	False	False	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	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	**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
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	False	False
False	False	False	False	False

3.1.2. Rekomendacja

Dla wymienionych wyżej klastrów dokonano porównania ich z nowo wygenerowanym użytkownikiem. Najlepiej dopasowane okazały się klastry numer 1 i 4, z wartością współczynnika podobieństwa Jaccarda równą 0.0833. W takich sytuacjach algorytm wybiera pierwszy znaleziony klaster, a więc do kolejnego etapu zostaje wybrany klaster 1.

Tabela 3. Wartości współczynnika podobieństwa Jaccarda dla nowego użytkownika i każdego klastra

Klaster	Wartość J
1	**0.0833**
2	0.0000
3	0.0000
4	0.0833
5	0.0769

Dla tej konfiguracji algorytm nie zarekomendował żadnej strony.

3.2. Liczba klastrów - 8

3.2.1. Informacje dotyczące klastrów

Tabela 4. Wynik analizy skupień dla ośmiu klastrów.

Klaster	Wartość
1	59 (2%)
2	61 (2%)
3	133 (3%)
4	887 (23%)
5	481 (12%)
6	289 (7%)
7	177 (5%)
8	1797 (46%)

Tabela 5. Wybrane centroidy dla 8 klastrów z flagami stron użytkowników

Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5	Klaster 6	Klaster 7	Klaster 8
False	False	False	**True**	**True**	False	False	False
False	False	False	False	False	False	False	False
False	False	False	False	**True**	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	False	False	False	False	**True**	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	False	False	False	False	False	False
False	False	False	False	False	False	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	False	False	False	False	False	False
False	False	False	False	False	False	False	False
False	False	**True**	False	False	False	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	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	False	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	False	False	False	False	False
False	False	False	False	False	False	False	False

3.2.2. Rekomendacja

W przypadku ośmiu klastrów najlepszą grupą okazał się klaster numer **1**. Wartość współczynnika podobieństwa Jaccarda jaką przyjmuje to **0.1538**.

Tabela 6. Wartości współczynnika podobieństwa Jaccarda dla nowego użytkownika i każdego klastra

Klaster	Wartość J
1	**0.1538**
2	0.0000
3	0.0000
4	0.0833
5	0.0769
6	0.0000
7	0.0000
8	0.0000

Dla tej konfiguracji algorytm zarekomendował:

— /shuttle/countdown/countdown.html

3.3. Liczba klastrów - 13

3.3.1. Informacje dotyczące klastrów

Tabela 7. Wynik analizy skupień dla trzynastu klastrów.

Klaster	Wartość
1	58 (1%)
2	61 (2%)
3	133 (3%)
4	834 (21%)
5	389 (10)
6	273 (7%)
7	169 (4%)
8	1330 (34%)
9	49 (1%)
10	41 (1%)
11	217 (6%)
12	250 (6%)
13	80 (2%)

Tabela 8. Wybrane centroidy dla 13 klastrów z flagami stron użytkowników

Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5	Klaster 6	Klaster 7	Klaster 8	Klaster 9	Klaster 10	Klaster 11	Klaster 12	Klaster 13
False	False	False	**True**	**True**	False	False	False	**True**	**True**	False	False	**True**
False	False	False	False	False	False	False	False	**True**	False	False	False	**True**
False	False	False	False	**True**	False	**True**	False	**True**	False	False	False	**True**
False	False	False	False	False	False	False	False	False	False	False	**True**	**True**
False	False	False	False	False	False	False	False	False	**True**	**True**	False	False
False	False	False	False	False	False	False	False	**True**	False	False	False	False
False	False	False	False	False	**True**	False	False	False	**True**	False	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
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	False	False	False	False	False	False	False	False	False	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
False	False	False	False	False	False	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	False	False	False	False	False
False	False	**True**	False	False	False	False	False	False	False	False	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
False	False	False	False	False	False	False	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	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	False	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	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	False	False	False	False	False	False	False	False	False	False

3.3.2. Rekomendacja

Dla trzynastu klastrów wartości współczynnika podobieństwa Jaccarda są nieco wyższe, a najlepszy z nich wynosi **0.2143** i jest reprezentowany przez klaster numer **9**.

Tabela 9. Wartość współczynnika podobieństwa Jaccarda

Klaster	Wartość J
1	0.1538
2	0.0000
3	0.0000
4	0.0833
5	0.0769
6	0.0000
7	0.0000
8	0.0000
9	**0.2143**
10	0.0000
11	0.0000
12	0.0000
13	0.0667

Algortym zasugerował dwie strony jako rekomendacje dla użytkownika:

— /shuttle/missions/sts-71/images/images.html

— /shuttle/missions/sts-71/mission-sts-71.html

4. Podsumowanie

Dla każdego badanego przypadku największy klaster zawsze zawierał wyłącznie wartości *False*. Oznacza to, że większość jego użytkowników nie odwiedziła żadnej ze stron. Dla pozostałych klastrów dało się zauważyć większą różnorodność, szczególne dla tych z małą liczbą przypadków. W tych klastrach występowało kilka stron, które zostały odwiedzone przez reprezentantów tej grupy. Niemniej jednak były to wciąż pojedyncze przypadki odwiedzin.

Mała różnorodność spowodowała niskie wartości współczynnika podobieństwa Jaccarda. Zgodnie ze wzorem, część wspólna badanych klastrów i dowolnego nowego użytkownika, w najlepszym przypadku (Tabela 8, klaster nr 9) stanowi mało liczny zbiór, ponieważ każdy klaster zawiera niewielką liczbę odwiedzonych stron. W konsekwencji otrzymane wyniki podobieństwa Jaccarda dla wszystkich przypadków są niskie (od 0.00 do 0.2143).

Zwiększenie liczby klastrów skutkowało podziałem na małe grupy, w których można było znaleźć więcej odwiedzonych stron. W każdym badanym przypadku nowy użytkownik był przypisywany do mało licznych zbiorów (4%, 2% i 1% wszystkich obserwacji). Przypisane użytkownika do takiego zbioru sugeruje, że odwiedzone przez niego strony cieszą się szczątkową, ale zauważalną popularnością. Istnieje duża szansa, że przy podobnych podziałach wielkościowych klastrów, w przypadku badania obejmującego dużo więcej stron, zaproponowane przez system rekomendacje faktycznie mogą zainteresować użytkownika.

W przypadku bardziej różnorodnej reprezentacji użytkowników efekty działania rekomendacji mogłyby okazać się bogatsze. Jeżeli baza zawiera użytkowników o wysokiej aktywności, odwiedzających różne strony, wówczas klastry reprezentowałyby tę aktywność. Konsekwencją tego byłoby więcej rekomendowanych stron dla nowych użytkowników.

Literatura

[1] *Jaccard Index/Similarity Coefficient,* Stephanie Glen, `https://www.statisticshowto.com/jaccard-index/` [dostęp: 02.12.2016]

[2] *Understanding K-means Clustering in Machine Learning,* Dr. Michael J. Garbade, `https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1` [dostęp: 12.09.2018]