Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 368 · 2018

Grzegorz Mika

Uniwersytet Ekonomiczny w Katowicach
Wydział Informatyki i Komunikacji
Katedra Inżynierii Wiedzy
grzegorz.mika@edu.uekat.pl

Grzegorz Dziczkowski

Uniwersytet Ekonomiczny w Katowicach
Wydział Informatyki i Komunikacji
Katedra Inżynierii Wiedzy
grzegorz.dziczkowski@ue.katowice.pl

ANALIZA ZACHOWAŃ UŻYTKOWNIKÓW Z WYKORZYSTANIEM TECHNIKI SEGMENTACJI

Streszczenie: W dobie komputeryzacji i informatyzacji niemal każdej sfery naszego życia generowana jest coraz większa ilość danych. Ma to oczywiście swoje plusy, jak również minusy. Dysponując narastającą w ogromnym tempie ilością informacji, człowiek z natury nie ma możliwości przetworzenia danych w wiedzę, która za nimi stoi. Problem ten wymusza gwałtowny rozwój w ostatnich latach, jednej z kluczowych dyscyplin informatyki – eksploracji danych. Głównym celem niniejszej pracy było zebranie i analiza plików logu serwerowego oficjalnej strony Uniwersytetu Ekonomicznej w Katowicach oraz poznanie kontekstu zachowań użytkowników w wyniku segmentacji populacji. W tym celu konieczne było stworzenie projektu, którego celem praktycznym było odkrycie użytecznej wiedzy z udostępnionych logów serwerowych przy wykorzystaniu metody grupowania. W rezultacie pozwoliło to na przeprowadzenie analizy nawigacji użytkowników, ocenę ich zachowań, trendów i przyzwyczajeń.

Słowa kluczowe: segmentacja, profilowanie użytkowników, klastering.

JEL Classification: D83.

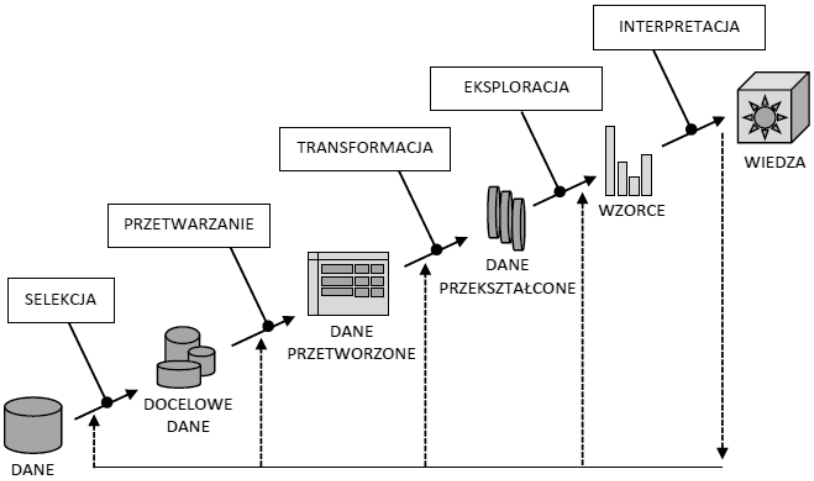
Wprowadzenie

Podejmowany temat pracy – analiza zachowań użytkowników z wykorzystaniem techniki segmentacji – ma na celu zaprezentowanie jednego z podejść odkrywania wiedzy z danych oraz problemów z nim związanych, w rezultacie umożliwiając uzyskanie możliwie najlepszej, najbardziej interesującej wiedzy. Głównym celem niniejszej pracy było zebranie i analiza plików logu serwerowego oficjalnej strony Uniwersytetu Ekonomicznej w Katowicach oraz poznanie kontekstu zachowań użytkowników w wyniku segmentacji populacji. Aby to osiągnąć, konieczne okazało się stworzenie projektu, którego celem praktycznym stało się odkrycie użytecznej wiedzy z udostępnionych logów serwerowych

przy wykorzystaniu metody grupowania. Zdobyte informacje powinny zostać przetworzone do postaci przedstawiającej cechy nawigacji internautów, które będą wykorzystane w budowaniu modeli segmentacji. Po wstępnej analizie wybrano odpowiedni algorytm i narzędzie potrzebne do zrealizowania przedstawionych zadań. W rezultacie pozwoliło to na przeprowadzenie analizy nawigacji użytkowników, ocenę ich zachowań, trendów i przyzwyczajeń.

1. Proces odkrywania wiedzy

Niniejszy rozdział poświęcony jest odkrywaniu wiedzy w bazach danych (ang. *Knowledge Discovery in Databases*). Ze względu na technikę wykorzystaną w naszym projekcie większą uwagę poświęcono uczeniu maszynowemu bez nadzoru. Przebieg iteracyjnego procesu KDD jest złożony, sprowadzający się do przygotowania i przetransformowania danych, ich eksploracji, kończąc na analizie uzyskanych wzorców [Bishop, 2006]. Odkrywanie wiedzy w bazach danych najczęściej składa się z następujących po sobie faz, zaprezentowanych na rysunku 1. Pierwszym etapem jest pozyskanie danych, w przypadku projektu zawartego w niniejszej pracy, obrazujących nawigacje internautów. Następnie w wyniku selekcji, przygotowania i transformacji dane są przefiltrowywane, oczyszczane oraz przekształcane do postaci umożliwiającej dalszą eksplorację. Ostatnie dwa kroki sprowadzają się do uczenia maszynowego, którego rezultaty mogą zostać poddane interpretacji oraz ewaluacji.



Rys. 1. Schemat procesu odkrywania wiedzy w bazach danych

Źródło: Opracowanie własne na podstawie: Fayyad, Piatetsky-Shapiro, Smyth [1996].

Jednym z kluczowych etapów bardziej ogólnego schematu odkrywania wiedzy jest eksploracja danych (ang. *data mining*), powiązana z dyscyplinami sztucznej inteligencji i statystyki. Posługując się odpowiednim algorytmem, można przyjąć, że na główny cel składają się znalezienie powiązań i schematów w dysponowanym zbiorze danych, przedstawiając je w sposób zrozumiały dla człowieka. W efekcie pozwala to zwrócić odpowiedzi na pytania bardziej ogólne. Uzupełniając to, co zostało powiedziane, w niniejszym procesie znajdowane są odpowiedzi, do których nie można ułożyć zwykłych zapytań SQL. Warto podkreślić, że w szerszym gronie sformułowania odkrywanie wiedzy w bazach danych oraz eksploracja danych są nieraz stosowane alternatywnie, lecz pierwsze pojęcie jest traktowane ogólniej jako proces. Natomiast eksploracja danych jest krokiem do stosowania szczególnych algorytmów dla ekstrakcji wzorców z danych, zawartej w tym procesie odkrywania użytecznej wiedzy [Fayyad, Piatetsky-Shapiro, Smyth, 1996].

1.1. Web Mining

Technika ta odwołuje się do automatycznego wykrywania i analizy wzorców z zebranych zapisów powstałych w wyniku interakcji użytkowników na stronach internetowych [Theodoridis, Koutroumbas, 2006]. Warto podkreślić, że metoda ta jest najczęściej wykorzystywana w marketingu ze względu na możliwość przeprowadzenia badania aktywności internautów, trendów, zachowań i sposobów nawigacji, głównie na stronach sieci Web przedsiębiorstw, często w oparciu o dane – logi serwerowe.

1.2. Klastering

Mając na uwadze cel niniejszej pracy, w rozważaniach należy uwzględnić także technikę grupowania (ang. *Clustering*), nazywaną również segmentacją. Polega na przesortowaniu badanego zbioru danych na grupy o różniących się cechach. Otrzymane zróżnicowane zestawy informacji zawierają elementy o podobnych i bardzo zbliżonych do siebie właściwościach [Theodoridis, Koutroumbas, 2006]. Główny problem w opisywanej metodzie uczenia maszynowego bez nadzoru przejawia się w odkryciu pewnych skupień reguł definiujących w sensownych klastrach. Powstałe grupy powinny pozwolić wskazać podobieństwa i różnice zachodzące między poszczególnymi zbiorami w interpretowalny sposób. Przebieg przypisywania obiektów do odpowiednich klastrów może prowadzić do różnych wyników.

W eksploracji korzystania z sieci Web istnieją dwa rodzaje odkrywania interesujących klastrów [Srivastava, Cooley, Deshpande, 2000]:

– Grupowanie użytkowników – segmentacja podatna na tworzenie klastrów użytkowników wykazujących zbliżone wzorce przeglądania stron. Odkryta wiedza ma szczególne zastosowanie w segmentacji rynku e-commerce lub świadczenia spersonalizowanych treści internetowych dla konkretnej grupy użytkowników.
– Grupowanie stron – odkrywa klastry stron posiadających odpowiednią zawartość; przydatny w dużej mierze dla wyszukiwarek internetowych.

Według Theodoridisa i Koutroumbasas [2006] rzeczowe kryterium przyjęte w metodzie grupowania wpływa na otrzymywane rezultaty. Każda grupa charakteryzuje się cechami, które są wspólne dla poszczególnego wystąpienia w danym klastrze. W literaturze proces grupowania przedstawiany jest w poniższych etapach [Theodoridis, Koutroumbas, 2006]:

1. Wybór cech – bardzo ważny jest pierwszy krok, związany z doprowadzeniem danych do minimalnej redundancji. Występowanie w zbiorze danych nieistotnych cech może negatywnie wpłynąć na jakość uczenia, zarazem zużywając więcej pamięci i wydłużając czas całego procesu.
2. Miary sąsiedztwa – w tej części, na etapie wstępnego przetwarzania, miara określa stopień podobieństwa cech dwóch wektorów. W rezultacie pomaga upewnić się, czy wszystkie wybrane atrybuty przyczyniają się do obliczeń miary sąsiedztwa oraz weryfikuje występowania tych dominujących inne.
3. Kryteria grupowania – określenie kryterium oparte jest na uzyskaniu sensownych klastrów znajdujących się w zestawie danych. W głównej mierze jest to zależne od interpretacji ekspertów dziedziny nauki, z której pochodzą eksplorowane dane.
4. Algorytmy grupowania – po przyjęciu miary sąsiedztwa i kryterium segmentacji dokonuje się wyboru konkretnego schematu algorytmicznego, tworząc klastry z zestawu danych. Powstałe grupy pozwalają na zobrazowanie trendów, cech, celu, a nawet predykcji.
5. Interpretacja wyników – ostatni etap dotyczy interpretacji uzyskanych efektów algorytmu grupowania. W celu wyciągnięcia właściwych wniosków należy uwzględnić wyniki klasteringu wraz z innymi dowodami powstałymi w rezultacie eksperymentów i analiz.

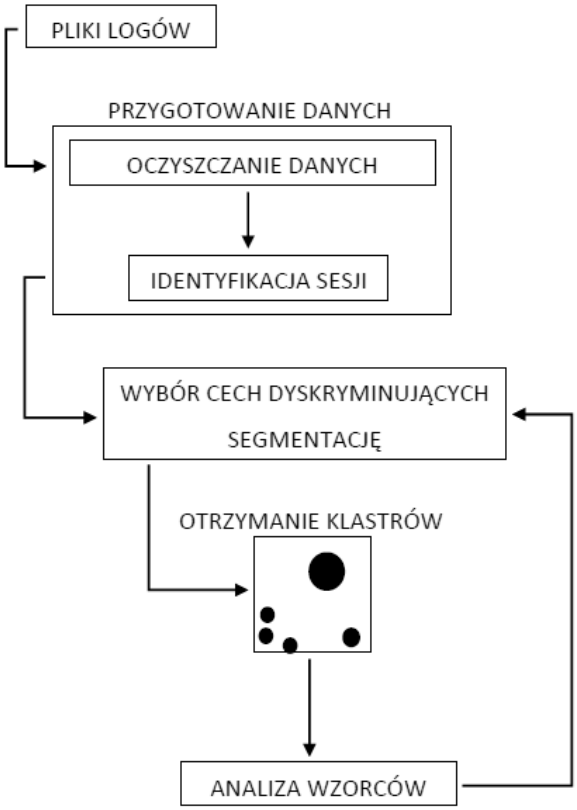
Warto jeszcze raz podkreślić, że cały proces klasteringu jest iteracyjny. Oznacza to, że model może być cały czas dostrajany, aż do osiągnięcia najlepszego oczekiwanego rezultatu. Technika grupowania w sieci Web, z wykorzystaniem odpowiednich algorytmów maszynowego uczenia, dzieli odwiedzających na grupy użytkowników o bardzo zbliżonych do siebie zachowaniach.

1.3. Poznawanie wzorców

W wyniku zastosowanych algorytmów uczenia maszynowego powstają pewne wzorce możliwe do zinterpretowania. Poznawanie i ich interpretacja są ostatnim etapem iteracyjnego dostrajania modelu w procesie odkrywania użytecznej wiedzy. Theodoridis i Koutroumbas [2006] formułują definicję poznawania wzorców (ang. *pattern recognition*) następująco: celem dyscypliny naukowej – rozpoznawania schematów – jest sklasyfikowanie obiektów do pewnych kategorii lub klas. Natomiast Bishop [2006] dodaje, że rozpoznawanie wzorców jest związane z automatycznym wykrywaniem prawidłowości w zbiorze danych, przy wykorzystaniu właściwych algorytmów oraz tych prawidłowości w celu sklasyfikowania danych do różnych kategorii. Podczas analizy plików logu serwerowego w projekcie użytkownicy zostaną przypisani do odpowiednich grup (klastrów) na podstawie podobnych wzorców zachowań.

2. Przetwarzanie informacji z logów serwerowych

Naświetlone w poprzednim rozdziale zagadnienie eksploracji danych w sieci Web zostanie bliżej przedstawione na przykładzie analizy kontekstu zachowań użytkowników oficjalnej strony internetowej Uniwersytetu Ekonomicznego w Katowicach. Proces ten został podzielony na kilka kroków, zgodnie z poniższym rysunkiem (rys. 2). W tym rozdziale opisana została część praktyczna przeprowadzonego przebiegu analizy, składająca się z pozyskania i przygotowania danych oraz wyboru cech do dalszych badań. Przedstawiono proces pozyskiwania sesji użytkowników na podstawie plików logu serwerowego. Opisano również podejście otrzymania większej liczby atrybutów z już istniejących sesji internautów.



Rys. 2. Proces analizy kontekstu zachowań użytkowników

Źródło: Opracowanie własne.

2.1. Proces pozyskiwania cech

Pierwszym krokiem w zrealizowanym projekcie naszej pracy było pozyskanie danych niezbędnych do opisywanej analizy, które obrazują nawigację użytkowników po stronie internetowej. Część dotycząca pozyskania danych jest aspektem etycznie niejednoznacznym, ponieważ zbierane i podlegające analizie są dane osobowe użytkowników, do których należy adres IP. W związku z przekazaniem informacji osobom trzecim pliki logu zostały poddane anonimizacji przez zastąpienie w adresie IP, czwartego oktetu znakiem „X". Czynność pozwoli zapewnić badanym anonimowość i poczucie bezpieczeństwa. W analizie zostały wykorzystane dane w postaci logów serwerowych z czterech tygodni (**3,3Gb**), z okresu od 23 listopada do 16 grudnia 2014 r. włącznie, udostępnione

przez Centrum Informatyczne Uniwersytetu. Do najważniejszych potrzebnych informacji w celu przeprowadzenia badania kontekstu zachowań użytkowników w otrzymanym pliku (tabela 1), pochodzącym z domeny ue.katowice.pl, należą kolejno: adres IP, pełna data (ang. *timestamp*) oraz wywołany adres z pominięciem domeny, z której pochodzi.

Tabela 1. Przykładowy plik logów serwerowych z domeny ue.katowice.pl

1	1.2.3.4 - - [24/Nov/2014:14:03:54 +0100] "GET /no_cache/uczelnia/wydzialy/wydzial-finansow-i-ubezpieczen/komunikaty-studia-stacjonarne/article/zapisy-na-euroclasses-juz-trwaja.html HTTP/1.1" 200 7319 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
2	4.3.2.1 - - [24/Nov/2014:14:03:55 +0100] "GET /no_cache/uczelnia/aktualnosci/article/trwa-rekrutacja-na-studia-podyplomowe-1.html HTTP/1.1" 200 7560 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:29.0) Gecko/20100101 Firefox/29.0"
3	3.2.1.4 - - [24/Nov/2014:14:03:55 +0100] "GET /studenci/dziekanaty/browse/4/article/zaproszenie-na-wyklad-pt-zbuduj-swoja-kariere-wybierz-zawod-przyszlosci.html HTTP/1.1" 200 8180 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
4	3.2.1.4 - - [24/Nov/2014:14:03:55 +0100] "GET /jednostki/rond/kontakt.html HTTP/1.1" 200 7439 "http://www.ue.katowice.pl/jednostki/rond/informacje-dla-studentow/wydzial-finansow-i-ubezpieczen/studia-niestacjonarne.html" "Mozilla/5.0 (Windows NT 6.1; rv:33.0) Gecko/20100101 Firefox/33.0"
5	2.4.2.1 - - [24/Nov/2014:14:03:54 +0100] "GET /jednostki/rond/kontakt.html HTTP/1.1" 200 7439 "http://www.ue.katowice.pl/jednostki/rond/informacje-dla-studentow/wydzial-finansow-i-ubezpieczen/studia-niestacjonarne.html" "Mozilla/5.0 (Windows NT 6.1; rv:33.0) Gecko/20100101 Firefox/33.0"

Źródło: Opracowanie własne.

W początkowej fazie otrzymane dane znajdują się w kilku osobnych plikach. W celu oczyszczenia danych z nieistotnych informacji oraz zebrania ich w jedną całość, przy tak dużej ilości danych, warto przygotować prosty program. Dysponowane dane zostały podzielone na sesje użytkowników. Sesja internautów to logi tego samego adresu IP. Segmentację populacji przeprowadzono po podzieleniu danych na sesje użytkowników, gdzie sesja wygasa po 30 minutach bezczynności użytkownika [Laur, Teisseire, Poncelet, 2003]. W literaturze naukowej wykazano, że wraz z postępem w dziedzinie technik eksploracji danych możliwym jest wyodrębnienie typowego zachowania podczas ustalonego okresu czasu [Laur, Teisseire, Poncelet, 2003]. Założenie polega na rozdzieleniu sesji na dwie, jeżeli od akcji '*n*' do '*n+1*' upłynie 30 minut.

2.2. Oczyszczanie danych

W pierwszym kroku dane znajdujące się w plikach logu zostały wyeksportowane do bazy danych, gdzie przeprowadzono oczyszczanie rekordów. Celem eksploracji danych jest odkrycie standardowych wzorców zachowań na stronach WWW dotyczących internautów. W związku z tym informacje źródłowe powinny przyjmować postać zbiorów lub sekwencji stron internetowych uzyskiwanych z serwera przez użytkowników. Po wstępnym zapoznaniu się z posiadanym zapisem znaczna większość z plików logu pochodziła z operacji wywołanych przez administratorów witryny, również duża część rekordów dotyczyła dokumentów (zdjęcia, dokumenty, arkusze kalkulacyjne). Niewielki fragment wszystkich rekordów zawierał wywołania głównych dokumentów odpowiadających podstronom internetowym domeny ue.katowice.pl. Najczęściej zapisy w logu dotyczą podstawowych dokumentów HTML lub związanymi z obiektami multimedialnymi, zagnieżdżonymi w stronie. Charakter pliku jest rozpoznawalny w dużym stopniu na podstawie jego rozszerzenia (.html, .jpg, .xls, .docx).

Aby dane źródłowe przeznaczone do analizy obejmowały wyłącznie informacje o wywołanych stronach przez internautów, należy poddać rekordy procesowi filtracji. W rezultacie wszystkie zapisy operacji administratorów (wgrywanie plików lub nawigowanie po panelu administracyjnym) oraz obcojęzycznych wersji strony Uniwersytetu Ekonomicznego w Katowicach zostały wykluczone z dalszej analizy. Ostatecznie do następnego etapu naszego badania z **12 mln rekordów** pozostało **ponad 2,5 mln**, które poddano procesowi pozyskiwania cech. W celu zwiększenia liczby cech w naszym projekcie przeprowadzono hierarchizację domeny. Witryna początkowo podzielona została na 7 wybranych kategorii:

- Uczelnia – użytkownik przegląda informacje związane z uczelnią,
- Kandydat – użytkownik przegląda informacje przeznaczone dla kandydata,
- Student – użytkownik przegląda informacje przeznaczone dla studenta,
- Absolwent – użytkownik przegląda informacje przeznaczone dla absolwentów,
- Pracownik – użytkownik przegląda informacje przeznaczone dla pracowników,
- Media – użytkownik przegląda informacje dla mediów,
- Inne – strony, które nie mieszczą się w pozostałych kategoriach.

Cecha może zawierać wartość ze skończonego zakresu lub ze zbioru skończonego (dyskretnego). Gdy skończony zbiór dyskretny ma tylko dwa elementy, jest nazywany binarnym lub dychotomicznym. Z danych logu naszej pracy zebrano różne właściwości w postaci łącznie, aż **47 zmiennych** opisujących zachowanie użytkowników, do których mogą należeć zwykli goście, kandydaci na studia, studenci, absolwenci czy pracownicy uczelni.

Proces segmentacji został wykonany za pomocą analizy statystycznej 47 wybranych zmiennych, do których należą m.in.:

- ID_session – opisuje liczbę sesji dla jednego użytkownika. Sesja jest ograniczona do 30 minut pomiędzy następnymi wywołaniami użytkownika. Identyfikator sesji jest wartością numeryczną, przykład: „32”.
- ID_user – opisuje użytkownika przez adres IP. Informacja ta jest podana w plikach logów serwerowych i jest obecna w każdym rzędzie dzienników. Dane zostały poddane anonimizacji poprzez usunięcie ostatniej części adresu IP i zastąpienie jej znakiem „x”. W rezultacie przykładowa zmienna jest w postaci: „192.168.0.X”.
- All_time – przedstawia całkowity czas trwania danej sesji w sekundach. Jest wartością liczbową, przykład: „120”.
- Start_time – przedstawia czas początku danej sesji w sekundach. Jest wartością liczbową, przykład: „43”.
- End_time – przedstawia czas końca danej sesji w sekundach. Jest wartością liczbową, przykład: „56”.
- Start_date – przedstawia datę początkową danej sesji w formacie dd/mm/rrrr.
- End_date – przedstawia datę końcową danej sesji w formacie dd/mm/rrrr. Atrybut ma zastosowanie tylko w przypadku, gdy sesja użytkownika podczas końca jednego dnia i początku drugiego, przykład: „początek sesji jest o godzinie 23.51 dnia pierwszego, koniec sesji jest o godzinie 00.14 dnia drugiego”.
- Count_actions – przedstawia liczbę wszystkich stron odwiedzonych przez użytkownika w trakcie danej sesji. Jest wartością liczbową, przykład: „15”.
- Count_category – przedstawia liczbę kategorii odwiedzonych przez użytkownika w trakcie danej sesji. Jeżeli użytkownik odwiedził dwa razy kategorię „Student”, to jest liczona jako raz. Jest wartością liczbową, przykład: „2”.
- Time_cat1 – przedstawia czas spędzony przez użytkownika w kategorii „Uczelnia” w sekundach.
- Time_cat2 – przedstawia czas spędzony przez użytkownika w kategorii „Kandydat” w sekundach.
- Time_cat3 – przedstawia czas spędzony przez użytkownika w kategorii „Student” w sekundach.
- Time_cat4 – przedstawia czas spędzony przez użytkownika w kategorii „Absolwent” w sekundach.
- Time_cat5 – przedstawia czas spędzony przez użytkownika w kategorii „Pracownik” w sekundach.
- Time_cat6 – przedstawia czas spędzony przez użytkownika w kategorii „Media” w sekundach.

- Count_scat – przedstawia liczbę podkategorii odwiedzonych przez użytkownika w trakcie danej sesji. Jeżeli użytkownik odwiedził dwa razy podkategorię „Student – plan zajęć", to jest liczona jako raz. Jest wartością liczbową, przykład: „7".
- Time_scat1 – przedstawia czas spędzony przez użytkownika we wszystkich podkategoriach w ramach kategorii „Uczelnia" w sekundach.
- Time_scat2 – przedstawia czas spędzony przez użytkownika we wszystkich podkategoriach w ramach kategorii „Kandydat" w sekundach.
- Time_scat3 – przedstawia czas spędzony przez użytkownika we wszystkich podkategoriach w ramach kategorii „Student" w sekundach.
- Time_scat4 – przedstawia czas spędzony przez użytkownika we wszystkich podkategoriach w ramach kategorii „Absolwent" w sekundach.
- Time_scat5 – przedstawia czas spędzony przez użytkownika we wszystkich podkategoriach w ramach kategorii „Pracownik" w sekundach.
- Time_scat6 – przedstawia czas spędzony przez użytkownika we wszystkich podkategoriach w ramach kategorii „Media" w sekundach.
- KxPx – przedstawia liczbę zmian po nawigacji z kategorii do podkategorii w ramach tej samej kategorii podczas danej sesji użytkownika. Jest wartością liczbową, przykład: „2".
- PxKx – przedstawia liczbę zmian po nawigacji z podkategorii do kategorii w ramach tej samej kategorii podczas danej sesji użytkownika. Jest wartością liczbową, przykład: „3".
- PxPx – przedstawia liczbę zmian po nawigacji z podkategorii do podkategorii w ramach tej samej kategorii podczas danej sesji użytkownika. Jest wartością liczbową, przykład: „1".
- KxKy – przedstawia liczbę zmian po nawigacji z jednej kategorii do innej kategorii podczas danej sesji użytkownika. Jest wartością liczbową, przykład: „2".
- PxPy – przedstawia liczbę zmian po nawigacji z podkategorii jednej kategorii do podkategorii innej kategorii podczas danej sesji użytkownika. Jest wartością liczbową, przykład: „5".
- OthersCat – przedstawia liczbę zmian po nawigacji z kategorii „Inne" do dowolnej kategorii podczas danej sesji użytkownika.
- ScatDocu – przedstawia liczbę zmian po nawigacji z dowolnej podkategorii do strony z dowolnym dokumentem podczas danej sesji użytkownika.
- DocuScat – przedstawia liczbę zmian po nawigacji ze strony z dowolnym dokumentem do dowolnej podkategorii podczas danej sesji użytkownika.

– CatDocu – przedstawia liczbę zmian po nawigacji z dowolnej kategorii do strony z dowolnym dokumentem podczas danej sesji użytkownika.
– DocuCat – przedstawia liczbę zmian po nawigacji ze strony z dowolnym dokumentem do dowolnej kategorii podczas danej sesji użytkownika.

2.3. Selekcja cech

Ważnym krokiem w iteracyjnym modelu analizy kontekstu zachowań internautów jest wybór cech nawigacji strony internetowej z danych w postaci sesji użytkowników. Selekcja atrybutów ma na celu dokonanie wyboru cech zawierających najlepszą wartość informacyjną, czyli teoretycznie służących segmentacji populacji. W prezentowanym projekcie, dysponując tak dużą liczbą aż **47 cech** nawigacji powstałych z **276504 sesji** użytkowników, potrzebny okazał się rozkład danych, który razem z celem klasteringu wskazuje potencjalnie najlepsze cechy. Następnie dokonano obliczenia współczynnika korelacji, co pomogło wykluczyć redundantne atrybuty z dalszego procesu grupowania.

3. Analiza danych

W tym rozdziale zostaną przedstawione wyniki klasteringu, który został przeprowadzony z wykorzystaniem programu na darmowej licencji – WEKA. Na podstawie oficjalnej strony narzędzia opracowanego na Nowozelandzkim Uniwersytecie Waikato program zapewnia wsparcie dla takich technik eksploracji danych jak klasyfikacja, klastering lub reguły asocjacyjne, jednocześnie dając możliwość zaprezentowania w sposób wizualny przetworzone dane. Aby otrzymać zbiór zachowań internautów, przyjęto punkt docelowy – wykonanie klasteringu z wykorzystaniem algorytmu K-średnich. Algorytm centroidów został wybrany do procesu ze względu na swoją złożoność obliczeniową. W przypadku algorytmu EM nie uzyskano rezultatów z powodu niewystarczającej pamięci. Segmentacja przeprowadzona została ze względu na różną porę w ciągu doby różniących się intensywnością akcji na sesję użytkownika. Z racji, że kontekst może wypływać ze zmiany zachowania użytkownika strony, zastosowano ograniczenie na całkowitą liczbę akcji w sesji. Zmniejszony zakres został dobrany na podstawie wglądu do całości danych. Minimalny próg akcji na sesję wynosi 3, ponieważ segmentacja sesji krótszych niż 3 akcje nie była istotna. Natomiast maksymalny próg został ustalony na poziomie 250 akcji. Większa liczba akcji względem niewielkiego czasu wykonywania pokazuje, że były to zachowania niepochodzące bezpośrednio od człowieka.

3.1. Segmentacja użytkowników

Na podstawie stworzonego wcześniej rozkładu danych dla wszystkich atrybutów drogą selekcji wybrano zestaw 15 najbardziej interesujących cech przeznaczonych do pierwszego procesu segmentacji. Warto podkreślić, że proces segmentacji jest iteracyjny (mogący trwać wiele miesięcy), dzięki czemu dostrajany jest model. Należy podkreślić, że w celu otrzymania bliskiego, najlepszego modelu jest potrzebna konsultacja, w tym przypadku z pracownikami uczelni – dla których realizowany jest projekt. Umożliwia to przedstawienie ze strony klienta wymagań oraz informacji, które pomogą dostrajać model w kierunku najbardziej pożądanych rezultatów. W pracy przedstawiono podejście bez wcześniej założonych tez czy celów. W rezultacie analiza skoncentrowana została na pozyskaniu klastrów zgodnych z dysponowanymi danymi, a nie z wcześniej przyjętymi tezami. Z informatycznego punktu widzenia wynikiem są klastry zachowań internautów, które w przyszłym procesie iteracyjnym mogą coraz lepiej obrazować trendy, zachowania, cechy lub predykcje.

3.2. Analiza otrzymanych klastrów

W wyniku przeprowadzenia segmentacji otrzymano wyniki zaprezentowane na rysunku (rys. 3). Pierwszy klaster skupia internautów korzystających średnio blisko 2 minuty w godzinach popołudniowych ze średnią liczbą akcji 8. Użytkownikami mogą być osoby znające dobrze zawartość strony internetowej uczelni, chcący pobrać konkretne pliki z dokumentami, ze zdjęciami znajdujące się w ramach pierwszej kategorii „o uczelni". Drugi klaster, podobnie jak trzeci, dotyczy osób nieco bardziej aktywnych na stronie w porównaniu do pierwszego. Użytkowników przypisanych do drugiego klastra wyróżnia duża aktywność na stronach pod nazwą kategorii „pozostałe". Internauci z trzeciego klastra w szczególności wyróżniają się pod względem czasu spędzonego w trakcie danej sesji, który średnio wynosi 21 minut. Klaster drugi charakteryzuje użytkowników, którzy odwiedzają stronę głównie w pewnym określonym celu. Ostatni klaster, poza dużym podobieństwem do drugiego, wyróżnia się użytkownikami, głównie przeglądającymi dokumenty w podkategorii kategorii pierwszej.

Cluster centroids:

Attribute	Full Data (163436)	Cluster# 0 (88853)	1 (16046)	2 (38387)	3 (20150)
All time	473.5349	112.3247	471.6284	1262.6348	564.5536
Start time	20:30 - 21:00	16:00 - 16:30	15:00 - 15:30	14:30 - 15:00	17:00 - 17:30
Start date	11/17/2014	11/17/2014	11/24/2014	11/20/2014	11/27/2014
Count actions	13.0236	8.1057	12.8401	24.7935	12.4329
Time cat3	10.8674	4.017	13.4103	22.7452	16.4218
Time scat1	139.211	38.9461	90.3765	350.3736	217.9474
Time scat3	47.1557	11.6402	49.3725	116.1342	70.5906
KxPx	0.1866	0.1514	0.2446	0.2328	0.2075
PxPx	2.6524	1.7462	2.3509	4.4325	3.4975
PxPy	0.6251	0.1612	0.3428	1.9482	0.3747
Count Docu	3.3486	2.8436	3.0992	4.6425	3.3087
Time Doku	75.9669	23.3864	53.3838	181.6795	124.419
Time Others	148.0796	23.8906	215.3706	444.2138	77.961
ScatDocu	0.2321	0.12	0.1966	0.4921	0.2596
DocuScat	0.3629	0.2591	0.3064	0.6207	0.3748

Rys. 3. Wyniki procesu segmentacji

Źródło: Opracowanie własne.

Podsumowanie

W pracy przedstawiono pełny proces odkrywania wiedzy z danych, jako ogólne podejście do identyfikacji, modelowania i analizy zachowań użytkowników strony internetowej. Wyszczególnionych zostało **47 cech** nawigacji, na podstawie których w etapie oczyszczania danych zostało wybranych 15 najbardziej interesujących atrybutów. W rezultacie przedstawiona została segmentacja populacji, dla której otrzymano **4 klastry**, których liczba została ustalona przez obliczenie inercji.

Reasumując, wykrywanie profili i zachowań użytkowników może być bardzo ważne dla wielu pracowników uczelni. W celu otrzymania najbardziej optymalnych wyników potrzebna jest współpraca ze środowiskiem akademickim, która pozwoli dostrajać model w kierunku najbardziej pożądanego rezultatu. Ponadto dysponując danymi z atrakcyjniejszego okresu roku, np. w trakcie trwającej sesji, rekrutacji czy innego, równie ważnego wydarzenia na uczelni, można by uzyskać znacznie bardziej atrakcyjne informacje w klastrach. Niemniej jednak dane zaprezentowane w powyższej pracy w postaci wykresów mogą okazać się dla pracowników źródłem przydatnej wiedzy po stosownej analizie.

Głównym celem niniejszej pracy było zebranie i analiza plików logu serwerowego oficjalnej strony Uniwersytetu Ekonomicznej w Katowicach oraz poznanie kontekstu zachowań użytkowników w wyniku segmentacji populacji. Po bardzo czasochłonnym procesie eksploracji danych z wykorzystaniem techniki segmentacji otrzymano sensowne klastry, bazując na wszystkich danych zawartych w wybranych 15 najbardziej interesujących cechach. Należy jednak mieć na uwadze fakt, że proces odkrywania wiedzy w bazach danych jest powtarzającym się – iteracyjnym – procesem. Jego długość jest zależna przede wszystkim od rezultatów, które klient chce otrzymać.

W projekcie przedstawiono przykładowe podejście do tworzenia modelu, którego dokładność może zostać rozbudowana o wiele dodatkowych czynników. Dalsze prace związane z wykonanym procesem mogą dążyć do wykonania na tych samych danych badania z wykorzystaniem innych metod i algorytmów eksploracji korzystania z sieci Web. Pozwoli to wykazać doświadczalnie różnice w poszczególnych technikach oraz rodzaj odkrytej wiedzy. Warto również rozwinąć kwestie wyboru cech o bardziej zaawansowane metody statystyczne, co w efekcie powinno pozwolić na uzyskanie jeszcze bardziej dokładnych rezultatów.

Literatura

Bishop C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, Singapore.

Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996), *The KDD Process for Extracting Useful Knowledge from Volumes of Data*, "Communication of the ACM", Vol. 39, No. 11, s. 27-34.

Laur P.-A., Teisseire M., Poncelet P. (2003), *Web Usage Mining: Extraction, Maintenance and Behavior Trends* [w:] Proceedings of the 1st Indian International Conference on Artificial Intelligence, IICAI 2003, Hyderabad, India, December 18-20, s. 1-14.

Srivastava J., Cooley R., Deshpande M. (2000), *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, "SIGKDD Explorations", Vol. 1, Iss. 2, s. 12-23.

Theodoridis S., Koutroumbas K. (2006), *Pattern Recognition*, 3rd Edition, Academic Press, Cambridge.

ANALYSIS OF USER BEHAVIOUR WITH THE USE OF SEGMENTATION TECHNIQUE

Summary: In the era of computerization, almost every sphere of our lives is generated by an increasing amount of data. Of course, it has its advantages as well as disadvantages. Due to the fact that the amount of information is growing really quickly, man is not naturally able to transform data into the knowledge behind them. This problem requires rapid development in recent years, one of the key areas of computer science - data mining. The main purpose of this research was to collect and analyze server log files from the official website of the University of Economics in Katowice. Then learning the context of user behavior as a result of segmentation of the population. For this purpose, it was necessary to create a project with the practical purpose of discovering useful knowledge from the server logs provided using the grouping method. As a result, users' navigation was analyzed and their behavior, trends and habits were assessed.

Keywords: segmentation, user profiling, clustering.