



SEGMENTACJA UŻYTKOWNIKÓW SERWISU WWW Z UŻYCIEM METOD STATYSTYCZNYCH I SIECI NEURONOWYCH

Grzegorz Migut, StatSoft Polska Sp. z o.o.

Wstęp

Poznanie klientów jest kluczowym elementem wpływającym na prowadzenie skutecznych działań sprzedażowych. Jedną z możliwości interakcji pomiędzy organizacją a jej klientami jest strona internetowa. Dzięki niej klienci mogą znaleźć interesujące informacje na temat działalności i oferty organizacji dotyczące na przykład nowych produktów, promocji itp. Strona internetowa jest jednak nie tylko źródłem informacji dla klientów na temat organizacji. Również organizacja dzięki niej może lepiej zrozumieć swoich klientów, poznać ich preferencje oraz zainteresowania.

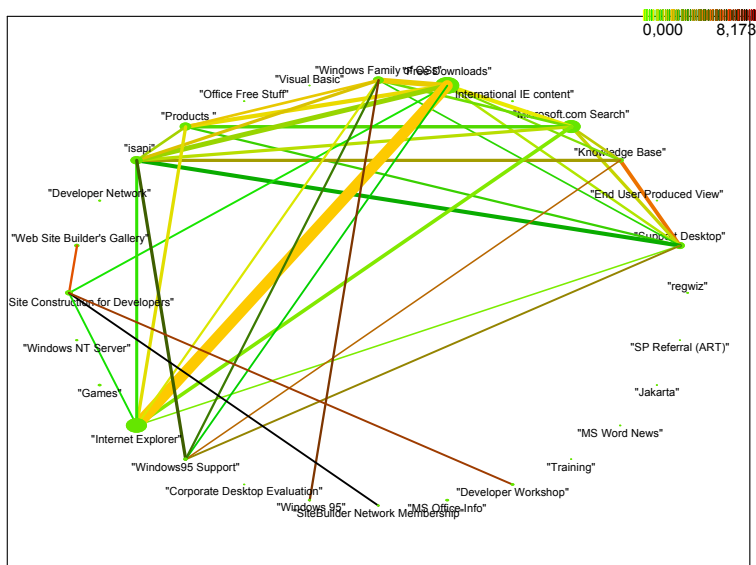
Lepsze zrozumienie zwyczajów i preferencji klientów jest możliwe dzięki analizie ich aktywności i wyborów dokonywanych podczas przeglądania stron internetowych organizacji. Informacje te są zapisywane w logach serwerów WWW. Dane te mogą stanowić nieoceniony zasób wiedzy, pozwalający uzyskać wiele cennych nieznanych wcześniej informacji o obecnych i potencjalnych klientach organizacji. Na ich podstawie możemy lepiej poznać osoby zainteresowane naszą działalnością, ocenić, jakie są ich preferencje i zwyczaje.

Chociaż dane zapisane w logach są zwykle łatwo dostępne, to jednak uzyskanie na ich podstawie interesujących informacji może być kłopotliwe ze względu na format ich zapisu utrudniający ich wygodne przeglądanie oraz zwykle ich bardzo dużą ilość. Dlatego też, by móc uzyskać z tego typu danych cenną wiedzę, warto skorzystać z metod służących do zgłębiania danych (*data mining*). Metody te przeznaczone są właśnie do analizy dużych zbiorów danych i pozwalają na wyszukiwanie w nich interesujących związków oraz nieznanych wcześniej zależności.

Podczas analizy logów internetowych stosujemy zwykle strategie budowy modeli i wyszukiwania reguł zaliczane do nieukierunkowanego *data mining* (uczenie bez nauczyciela). Pierwszy rodzaj analizy, jaki można zastosować w odniesieniu do danych zawartych w logach, polega na wyszukiwaniu ukrytych związków i prezentowaniu ich w formie reguł logicznych, mających formę JEŻELI [*poprzednik*] TO [*następnik*]. Reguły te można również przedstawiać za pomocą wykresów. Przykładowo poniższy wykres przedstawia związki pomiędzy poszczególnymi stronami analizowanego serwisu internetowego. Dzięki tego typu analizie możemy zidentyfikować zwyczaje gości naszych stron oraz zbadać, z jakich miejsc naszego serwisu trafiają do konkretnej podstrony. Tego typu informacja



może nam pomóc w zrozumieniu ich zachowań oraz może pozwolić lepiej dopasować strukturę naszego serwisu do oczekiwań naszych klientów¹.



Drugą grupą analiz, jaką możemy z powodzeniem stosować dla danych z logów internetowych są analizy dzielące osoby odwiedzające stronę na jednorodne grupy, czyli dokonujące ich segmentacji. Dzieląc naszych gości na jednorodne grupy możemy spojrzeć na nich z bardziej ogólnej perspektywy, określić pewne wzorce zachowań, ocenić licznosc poszczególnych grup osób oraz wskazać elementy, które najbardziej różnicują poszczególne segmenty.

Znajomość zainteresowań i preferencji wyróżnionych grup klientów można wykorzystać w przyszłości przy podejmowaniu decyzji marketingowych. Możemy również obserwować zmianę zachowania naszych klientów w czasie, analizując, jak zmienia się charakter naszych grup, ich licznosci oraz główne cechy charakterystyczne. Na tej podstawie możemy wnioskować, jak podejmowane przez nas działania wpływają na poszczególne segmenty, na przykład czy powodują wzrost wizyt danego typu klientów czy też ich spadek².

Metody wykorzystywane w segmentacji

Podczas budowy modelu *data mining* zwykle podajemy na wstępie dane reprezentujące rozpoznane wcześniej wzorce (grupy), które model ma następnie odtwarzać i na tej

¹ Więcej informacji na temat metod umożliwiających budowanie podobnych reguł można znaleźć w artykule „Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej”, który znajduje się w niniejszym opracowaniu.

² Szczegółowe informacje na temat zastosowań technik segmentacji można znaleźć w [5], w artykule „Jak znaleźć grupy podobnych klientów, czyli metody segmentacji”.



podstawie klasyfikować do danej grupy nowe obserwacje. W przypadku segmentacji sytuacja jest inna. Na wstępie analizy nie dysponujemy żadną informacją, jakie segmenty występują w danych, ani też ile jest tych segmentów. Wiedzę tę pragniemy dopiero zdobyć w wyniku analizy. Tak sformułowane zadanie analityczne wymaga zastosowania jednej z metod nieukierunkowanego *data mining*. Najbardziej popularnymi metodami stosowanymi do segmentacji są metody analizy skupień oraz sieci neuronowe Kohonena (SOM).

Analiza skupień – metody hierarchiczne i niehierarchiczne

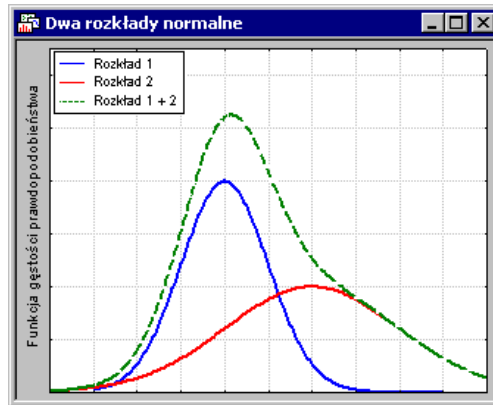
Celem **analizy skupień** (*cluster analysis*) jest wyodrębnienie ze zbioru danych obiektów, które byłyby podobne do siebie, i łączenie ich w grupy. W wyniku działania tej analizy z jednego niejednorodnego zbioru danych otrzymujemy grupę kilku jednorodnych zbiorów. Obiekty znajdujące się w tym samym zbiorze uznawane są za „podobne do siebie”, obiekty z różnych zbiorów traktowane są jako „niepodobne”. Techniki analizy skupień obejmują faktycznie kilka różnych algorytmów, które można najogólniej podzielić na metody hierarchiczne i niehierarchiczne.

Hierarchiczną metodą analizy skupień jest metoda aglomeracyjna. Algorytm aglomeracji służy do grupowania obiektów (np. w naszym przypadku obiektem jest osoba odwiedzająca stronę) w coraz to większe zbiory (skupienia), z zastosowaniem pewnej miary podobieństwa lub odległości. Typowym wynikiem tego typu grupowania jest hierarchiczne drzewo. Na początku tej analizy uznajemy, że każdy element zbioru stanowi oddzielną grupę. Następnie stopniowo osłabiamy kryterium uznawania obiektów za takie same, co powoduje grupowanie się obiektów podobnych. W miarę dalszego osłabiania kryterium wiążemy ze sobą coraz więcej obiektów i agregujemy je w coraz większe skupienia elementów, coraz bardziej różniących się od siebie. W końcu, na ostatnim etapie, wszystkie obiekty zostają ze sobą połączone. Efekty działania tego algorytmu można przedstawić w formie hierarchicznego drzewa, które przedstawia kolejne kroki działania analizy. Tego typu analizę możemy przeprowadzić nie tylko dla przypadków, ale również dla zmiennych, co polega na łączeniu najbardziej podobnych (w sensie odległości, a nie korelacji) w grupy, podobnie jak przedstawiono powyżej. Metoda aglomeracyjna jest rzadko stosowana w segmentacji dużej liczby obiektów, ponieważ wymaga obliczenia macierzy odległości pomiędzy wszystkimi analizowanymi obiektami, co jest bardzo wymagające numerycznie. Jest ona jednak bardzo pomocna podczas ustalania optymalnej liczby skupień, na jaką należy podzielić analizowaną zbiorowość.

Do najważniejszych metod niehierarchicznych należy zaliczyć metodę *k*-średnich oraz EM. Stosowanie metody *k*-średnich wymaga od nas podania liczby grup, na które zostanie podzielony wejściowy zbiór danych. Jedną z wersji tej metody polega na losowym wyborze *k* obiektów z analizowanego zbioru i uznania ich za środki *k* grup. Każdy z pozostałych obiektów jest przypisywany do grupy o najbliższym mu środku. Następnie oblicza się nowe środki każdej podgrupy na podstawie średnich arytmetycznych ze współrzędnych zawartych w nich obiektów. W kolejnym kroku następuje przegrupowanie elementów grup, każdy obiekt jest przesuwany do tej grupy, do której środka ma najbliżej. Procedurę tę powtarzamy do momentu, gdy w danej iteracji żaden z obiektów nie zmieni

swojej podgrupy. Pewną wadą tej metody jest konieczność odgórnego określenia liczby skupień występujących w danych, dlatego też zaleca się powtórzenie procedury dla różnych wartości k i wybranie tej, dla której zbiór danych jest podzielony najlepiej.

Metoda EM jest czasem nazywana analizą skupień bazującą na prawdopodobieństwie lub statystyczną analizą skupień. Program wyznacza skupienia, zakładając różnorodne rozkłady prawdopodobieństwa zmiennych uwzględnianych w analizie. Na początku działania algorytmu, podobnie jak w metodzie k -średnich, musimy podać liczbę skupień, jakie powinny być wyodrębnione ze zbioru wejściowego.

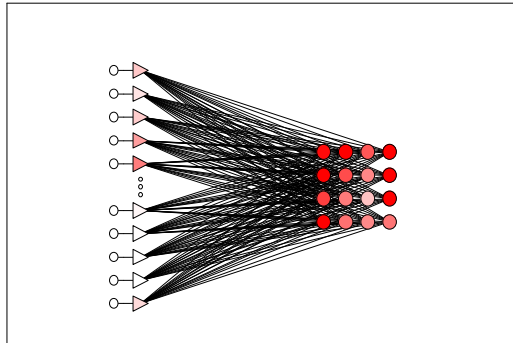


Załóżmy, że przeprowadziliśmy badania w pewnej dużej zbiorowości pod kątem jednej cechy ciągłej. Zaobserwowany rozkład tej cechy był zgodny z funkcją gęstości opisaną linią przerywaną (*Rozkład 1+2*) charakteryzującą się pewną średnią oraz odchyleniem standardowym. Wiemy też, że w zbiorowości tej występują dwa segmenty (na przykład kobiety i mężczyźni) o różnych parametrach funkcji gęstości w swoich segmentach. Algorytm EM ma na celu określenie parametrów rozkładów segmentów na podstawie rozkładu całej grupy oraz przydzielenie poszczególnych obserwacji do najbardziej odpowiadających im segmentów (klasyfikacja następuje na zasadzie prawdopodobieństwa). Na naszym rysunku rozkłady dwóch segmentów zostały przedstawione jako *Rozkład 1* oraz *Rozkład 2*. Po zsumowaniu dają one funkcję rozkładu całej zbiorowości (*Rozkład 1+2*). Algorytm EM dokonuje klasyfikacji nie tylko przy założeniu normalności rozkładu, jak to zaprezentowano na rysunku, wykorzystując go, można również określić inną funkcję gęstości dla badanej cechy (badanych cech).

Sieci neuronowe - sieć Kohonena (Self Organizing Map)

Sieć Kohonena (SOM) została zaprojektowana do uczenia w trybie bez nauczyciela – podczas uczenia ustalanie parametrów sieci nie jest sterowane za pomocą wartości wyjściowych, podczas nauki prezentowane są jedynie dane kierowane na wejścia sieci. Sieć ta posiada dwie warstwy: warstwę wejściową oraz warstwę wyjściową składającą się z neuronów radialnych. Warstwa ta znana jest również jako warstwa tworząca mapę topologiczną, ponieważ takie jest jej najczęstsze zastosowanie. Neurony w warstwie tworzącej mapę topologiczną zwykle wyobrażamy sobie jako węzły dwuwymiarowej

siatki, chociaż możliwe jest również tworzenie jednowymiarowych sieci w postaci długich łańcuchów.



Sieci Kohonena uczone są przy wykorzystaniu algorytmu iteracyjnego. Rozpoczynając od początkowych, wybranych w sposób losowy centrów radialnych, algorytm stopniowo modyfikuje je w taki sposób, aby odzwierciedlić skupienia występujące w danych uczących. Iteracyjna procedura ucząca dodatkowo porządkuje neurony reprezentujące centra położone blisko siebie na mapie topologicznej.

Podstawowy iteracyjny algorytm działa przez dużą liczbę epok (podczas każdej epoki prezentowany jest sieci cały zestaw danych) w następujący sposób [4]:

- ◆ pokazywany jest zestaw danych wejściowych ze zbioru uczącego,
- ◆ wszystkie neurony sieci wyznaczają swoje sygnały wyjściowe, stanowiące odpowiedź na podane wejścia,
- ◆ wybierany jest neuron zwycięski (tzn. ten, który reprezentuje centrum najbardziej zbliżone do prezentowanego na wejściu przypadku),
- ◆ neuron zwycięski modyfikowany jest w taki sposób, aby upodobnić jego wzorec do prezentowanego przypadku. W tym celu wyznaczana jest ważona suma przechowywanego w neuronie centrum oraz przypadku uczącego,
- ◆ wraz ze zwycięskim neuronem w podobny sposób modyfikowane są parametry jego sąsiadów (sąsiedzi wyznaczeni są w oparciu o przyjęty wzór topologii sieci).

Algorytm wykorzystuje zmienny w czasie współczynnik uczenia, który jest wykorzystywany do wyznaczenia ważonej sumy, i powoduje, że zmiany początkowo duże i szybkie stają się coraz bardziej subtelne w trakcie kolejnych epok. Umożliwia to ustalenie centrów w taki sposób, że stanowią one pewien kompromis pomiędzy wieloma przypadkami powodującymi zwycięstwo rozważanego neuronu.

Własność uporządkowania topologicznego jest osiągana przez zastosowanie w algorytmie koncepcji sąsiedztwa. Sąsiedztwo stanowią neurony otaczające neuron zwycięski. Sąsiedztwo, podobnie jak współczynnik uczenia, zmniejszane jest wraz z upływem czasu, tak więc początkowo do sąsiedztwa należy stosunkowo duża liczba neuronów; w końcowych etapach sąsiedztwo ma zerowy zasięg. Ma to istotne znaczenie, ponieważ w algorytmie



Kohonena modyfikacja wag jest w rzeczywistości przeprowadzana nie tylko w odniesieniu do neuronu zwycięskiego, ale również we wszystkich neuronach należących do sąsiedztwa.

Po nauczeniu sieci Kohonena poprawnego rozpoznawania struktury prezentowanych danych można jej użyć jako narzędzia przeprowadzającego wizualizację danych w celu ich lepszego poznania. Ważnym elementem przygotowania sieci Kohonena do bieżącego użytkowania jest właściwe opisanie uformowanej mapy topologicznej. Ustalenie związków pomiędzy skupieniami a znaczeniami wymaga zwykle odwołania się do dziedziny, której dotyczy analiza [4].

Segmentacja użytkowników serwisu WWW

Analiza polegająca na segmentacji gości serwisu internetowego zostanie zaprezentowana na przykładzie danych będących zapisem odwiedzin serwisu www.microsoft.com. Poniższy przykład oparty będzie na przykładzie zamieszczonym w [3].

Przygotowanie danych do analizy

Jednym z najtrudniejszych i na pewno najbardziej pracochłonnych etapów analizy *data mining* jest odpowiednie przygotowanie danych, by w jak najlepszym stopniu spełniały one wymogi stawiane przez metody, jakich pragniemy użyć. W przypadku analizy skupień format danych, jaki musimy uzyskać, powinien mieć kształt płaskiej tabeli danych (taki format danych jest wymagany przez większość metod *data mining*). W przypadku logów internetowych dane zapisane są w pliku tekstowym, w którym występują separatory oddzielające od siebie poszczególne wpisy. By możliwe było użycie danych o takim formacie, konieczne jest uprzednie ich przekształcenie, w tym celu można skorzystać z wbudowanych w program *STATISTICA* procedur wprowadzania danych zewnętrznych. Często jednak format danych zapisanych w logach jest na tyle skomplikowany, że konieczne jest przygotowanie programu lub makra umożliwiającego zmianę ich pierwotnego kształtu.

W naszym przypadku dane, które będą punktem wyjścia dla analizy, zostały już zapisane w dwóch osobnych plikach danych programu *STATISTICA*. W pierwszym o nazwie *MSWebData.sta* przechowywany jest zapis odwiedzin poszczególnych osób na interesujących ich stronach. Plik ten zawiera prawie 100 000 przypadków. Każdy przypadek dotyczy wizyty jednej osoby na jednej z podstron serwisu. Arkusz ten zawiera następujące zmienne:

- ◆ *Visitor ID* – identyfikator osoby odwiedzającej stronę,
- ◆ *Web Area ID* – identyfikator przeglądanej strony,
- ◆ *Time* – kolejność przeglądania danej strony w czasie odwiedzin w serwisie.

	1 Visitor ID	2 Web Area ID	3 Time	
1	10001	1000	0	
2	10001	1001	1	
3	10001	1002	2	
4	10002	1001	0	
5	10002	1003	1	
6	10003	1001	0	
7	10003	1003	1	
8	10003	1004	2	
9	10004	1005	0	
10	10005	1006	0	

Drugi plik pełni funkcję słownika. Dzięki niemu możliwa jest połączenie wartości zmiennej *Web Area ID* z odpowiadającym jej adresem serwisu www.microsoft.com. Plik ten zawiera 294 przypadki, tyle różnych stron odwiedzili internauci, których dotyczy analizowane dane. Informacje te zapisane są w trzech zmiennych:

- ◆ *Web Area ID* – identyfikator przeglądanej strony,
- ◆ *CONTENT* – odnosi się do zawartości danej strony,
- ◆ *URL* – względny adres danej strony.

	1 Web Area ID	2 CONTENT	3 URL (relative to www.microsoft.com)	
1	1287	"International AutoRoute"	"/autoroute"	
2	1288	"library"	"/library"	
3	1289	"Master Chef Product Information"	"/masterchef"	
4	1297	"Central America"	"/centroam"	
5	1215	"For Developers Only Info"	"/developer"	
6	1279	"Multimedia Golf"	"/msgolf"	
7	1239	"Microsoft Consulting"	"/msconsult"	
8	1282	"home"	"/home"	
9	1251	"Reference Support"	"/referencessupport"	
10	1121	"Microsoft Magazine"	"/magazine"	
11	1083	"MS Access Support"	"/msaccesssupport"	
12	1145	"Visual Fox Pro Support"	"/vfoxprosupport"	
13	1276	"Visual Test Support"	"/vtestsupport"	
14	1200	"Benelux Region"	"/benelux"	
15	1259	"controls"	"/controls"	

Chociaż dane zapisane są już w formacie *STATISTICA*, ich obecna forma nie pozwala jeszcze na przeprowadzenie segmentacji gości serwisu. Obecnie jeżeli dana osoba odwiedziła podczas swojej wizyty trzy różne podstrony, to informacja o jej aktywności zapisana została w trzech różnych wierszach. By móc przeprowadzić analizę polegającą na segmentacji, konieczne jest takie przygotowanie danych, by jeden przypadek w arkuszu odpowiadał jednej osobie i był zapisem jej pełnej aktywności podczas całej sesji. Dane w nowym układzie powinny zawierać identyfikator internauty oraz informację, jakie strony odwiedzał.



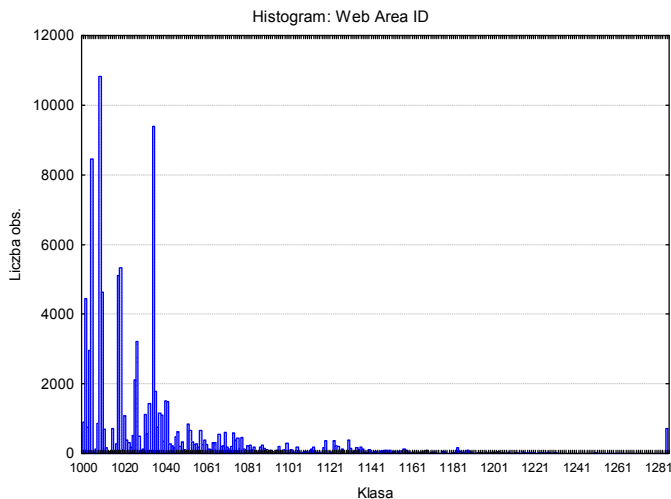
Każdy rodzaj strony powinien być przedstawiony przez osobną zmienną. Wartościami tych zmiennych powinny być liczby odwiedzin danej strony przez danego internautę.

Ponieważ w naszym przypadku Internauci odwiedzili 294 różne strony, dlatego też nowy zbiór danych mógłby zawierać właśnie taką liczbę zmiennych. Warto jednak zastanowić się, czy przyjęty w ten sposób poziom szczegółowości jest prawidłowy, czy liczba rodzajów odwiedzanych stron nie jest zbyt duża.

Kwestię odpowiedniego poziomu szczegółowości danych należałoby rozważyć zarówno z punktu widzenia biznesowego celu analizy, jaki i z punktu widzenia formalnych wymagań technik analitycznych.

Jeśli chodzi o kwestie biznesowe, to zawarty w danych poziom szczegółowości może okazać się za wysoki. Warto zadać sobie pytanie, czy pragniemy odróżniać osoby, które szukały informacji o poszczególnych elementach pakietu MS Office czy różnych wersjach systemu Windows, czy też lepiej będzie, jeśli dla nich utworzymy ogólną kategorię Office oraz Windows.

Oceniając zasadność analizy szczegółowych danych z punktu widzenia formalnych wymagań stawianych przez techniki analityczne, warto przyjrzeć się licznosciom odwiedzin danych stron. Wygodnym narzędziem do oceny tej licznosci jest histogram obliczony dla zmiennej *Web Area ID*. Możemy zauważyć, że kilka stron jest odwiedzanych bardzo często, jednak zdecydowana większość była odwiedzana jedynie przez jednego lub kilku internautów. Utworzona dla takiej podstrony zmienna zawierałaby praktycznie same zera i byłaby nieprzydatna w analizie³.



³ Przyjętą regułą jest, by nie wykorzystywać podczas analizy zmiennych, które przyjmują stałą wartość dla więcej niż 95% przypadków.



Biorąc pod uwagę powyższe argumenty, przed zmianą układu danych należy jeszcze dokonać kategoryzacji poszczególnych rodzajów stron. Odpowiednie strony podzielono na kategorie w sposób analogiczny do przedstawionego w [3]:

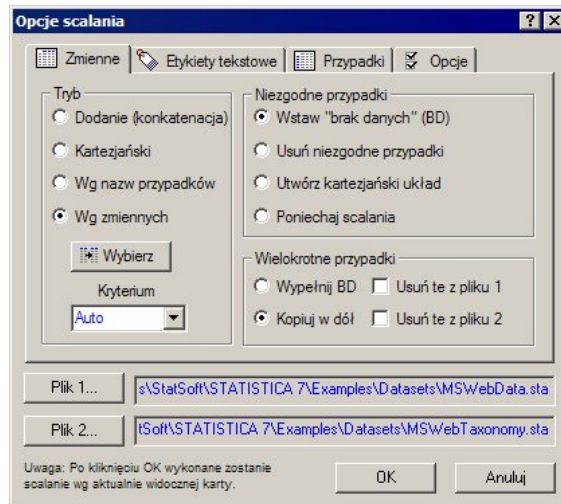
- ◆ *Initial* – obejmuje strony ogólnego dostępu,
- ◆ *Support* – dotyczy wszystkich stron odnoszących się do pomocy i wsparcia technicznego,
- ◆ *Entertainment* – strony dotyczące rozrywki, gier itp.,
- ◆ *Office* – strony zawierające informację o pakiecie Office,
- ◆ *Windows* – informacje o systemach operacyjnych Windows,
- ◆ *Server* – programy serwerowe,
- ◆ *Othersoft* – oprogramowanie różne od wymienionych w powyższych kategoriach,
- ◆ *Download* – zawiera wszystkie strony dotyczące pobierania plików i aktualizacji,
- ◆ *Otherint* – przeznaczone dla osób zajmujących się profesjonalnie branżą IT,
- ◆ *Development* – strony przeznaczone dla osób zajmujących się rozwijaniem oprogramowania,
- ◆ *Business* – strony przeznaczone dla biznesu,
- ◆ *Info* – strony zawierające informacje na temat nowych produktów i usług,
- ◆ *Area* – dotyczy stron dostępnych lokalnie w danym kraju, zawierających informacje w określonym języku.

Informacje o nowych kategoriach zostały zapisane w arkuszu *Web Area ID* w nowej zmiennej o nazwie *Category*.

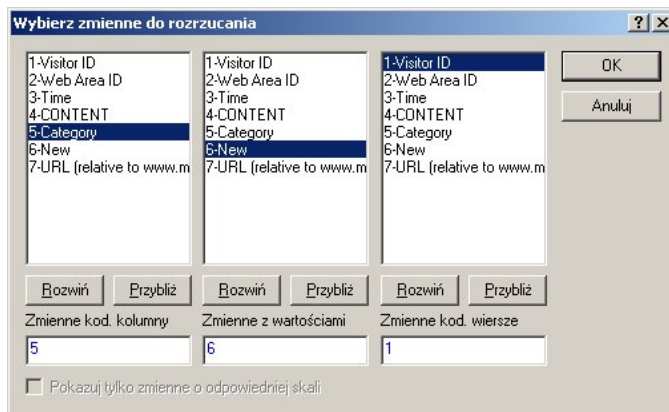
Przed zmianą układu danych konieczne było jeszcze połączenie zbioru zawierającego informacje na temat odwiedzin na stronach ze zbiorem zawierającym opis stron z przeprowadzoną kategoryzacją, tak by w jednym arkuszu znajdował się zarówno identyfikator internauty oraz strony, jaką odwiedził, jak również informacje dotyczące kategorii, do jakiej należała odwiedzana strona.

Połączenie obu zbiorów wykonujemy za pomocą polecenia *Dane-Scal*, wskazując w oknie *Opcje scalania* zmienną *Web Area ID* jako identyfikator obu przypadków oraz w obszarze *Wielokrotne przypadki* opcję *Kopiuj w dół* (zob. rysunek poniżej).

Uzyskany w ten sposób arkusz przekształcamy następnie na arkusz o nowym układzie danych, w którym każdy przypadek reprezentuje jednego internautę, poszczególne zmienne odpowiadają przygotowanym kategoriom stron, natomiast wartości mówią nam, ile razy dany internauta odwiedzał stronę z danej kategorii.



Operację zmiany układu danych wykonujemy za pomocą operacji *Dane-Rozrzucić po zmiennych* (operacja ta jest odpowiednikiem tabel przestawnych w *MS Excel*). Jako zmienną kodującą wiersze wybieramy zmienną *Visitor ID*, natomiast jako zmienną kodującą kolumny wskazujemy zmienną *Category*. Przed wykonaniem analizy konieczne jest dodanie zmiennej, która zawierała będzie liczności odwiedzin na stronie. Ponieważ w przypadku zbioru *MSWebData.sta* każdy przypadek reprezentuje jedną wizytę, dlatego też tę zmienną wypełniamy jedynkami. Nowo utworzoną zmienną wskazujemy jako zmienną z wartościami.



Po wykonaniu powyższych przekształceń otrzymany zbiór danych jest gotowy do przeprowadzenia na jego podstawie segmentacji internautów. Możemy zauważyć, że prawie 100 000 przypadków zapisanych w zbiorze *MSWebData.sta*, było zapisem aktywności 32 711 internautów (patrz rysunek poniżej).

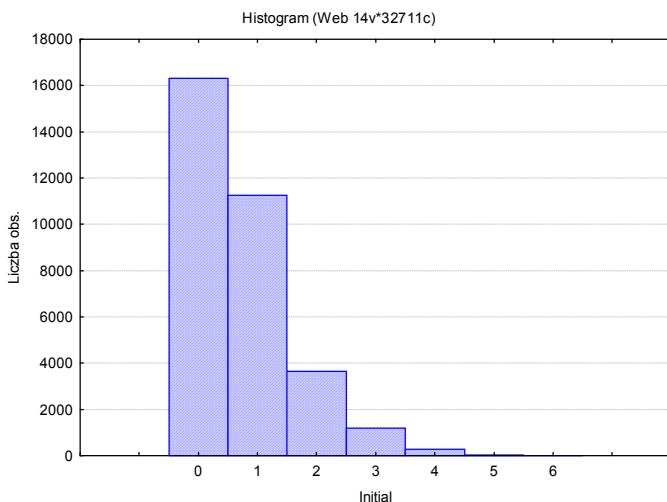
Na każdym etapie przygotowania danych należy sprawdzić, czy proces transformacji przebiegł poprawnie, należy również zbadać dane pod kątem ich poprawności, jedno-



rodności czy też występowania braków danych⁴. Opis tych czynności wykracza poza ramy tego artykułu. My ograniczymy się jedynie do zbadania rozkładu uzyskanych zmiennych.

Dane: Web (14 zm. * 32711 prz.)

	1 Visitor ID	2 Entertainment	3 Initial	4 Support	5 Area	6 Otherint	7 Office	8 Download	9 Windows	10 Development	11 Othersoft	12 Server	13 Business	14 Info
1	10001	0	1	1	1	0	0	0	0	0	0	0	0	0
2	10002	0	1	1	0	0	0	0	0	0	0	0	0	0
3	10003	0	2	1	0	0	0	0	0	0	0	0	0	0
4	10004	0	0	0	1	0	0	0	0	0	0	0	0	0
5	10005	0	0	0	0	1	0	0	0	0	0	0	0	0
6	10006	0	2	0	0	0	0	0	0	0	0	0	0	0
7	10007	0	0	0	0	0	1	0	0	0	0	0	0	0
8	10008	0	1	0	0	0	0	0	0	0	0	0	0	0
9	10009	0	0	0	0	0	0	1	1	0	0	0	0	0
10	10010	0	1	1	0	0	1	0	0	2	1	0	0	0
11	10011	0	2	0	0	0	3	0	0	0	0	0	0	0
12	10012	0	0	0	0	0	0	0	0	1	1	0	0	0
13	10013	0	0	0	0	0	0	1	0	0	0	0	0	0
14	10014	0	0	0	1	0	0	0	0	0	0	0	0	0
15	10015	0	0	0	0	0	0	0	0	0	0	1	0	0
16	10016	0	0	0	0	1	0	0	0	1	0	0	0	0
17	10017	0	1	0	0	0	0	0	0	3	0	0	0	0
18	10018	0	1	0	0	0	0	0	0	0	0	0	0	0
19	10019	3	4	0	0	0	1	1	0	0	0	1	1	0
20	10020	1	0	1	0	0	1	1	0	0	0	0	0	0
21	10021	1	3	1	0	2	1	1	2	2	1	1	1	1
22	10022	0	2	0	0	0	0	1	0	0	0	0	0	0
23	10023	0	0	0	0	0	0	1	0	0	0	0	0	0
24	10024	0	0	0	0	0	0	0	0	1	0	0	0	0
25	10025	0	0	0	0	1	0	0	0	0	0	0	0	0
26	10026	0	0	0	0	0	1	0	0	0	0	0	0	0
27	10027	0	0	1	0	0	1	1	0	0	0	0	0	0
28	10028	0	0	0	0	0	0	0	0	0	0	0	1	0



Jeśli przyjrzeć się uzyskanym zmiennym, możemy zauważyć, że wszystkie mają rozkłady prawostronnie skośne, we wszystkich cechach dominuje wartość zero. Najczęściej odwiedzanymi stronami były strony z kategorii *Initial*, odwiedziło je ponad 50% internautów, najrzadziej odwiedzane były strony zawierające informacje o systemach serwerowych.

⁴ Szczegółowy opis działań związanych z przygotowaniem danych do analizy można znaleźć w artykule „Model *data mining* przewidujący odpowiedź klientów na ofertę” znajdującym się w niniejszym opracowaniu.



Strony te wybrało nieco ponad 6% odwiedzających. Powyższy histogram przedstawia rozkład zmiennej *Initial*.

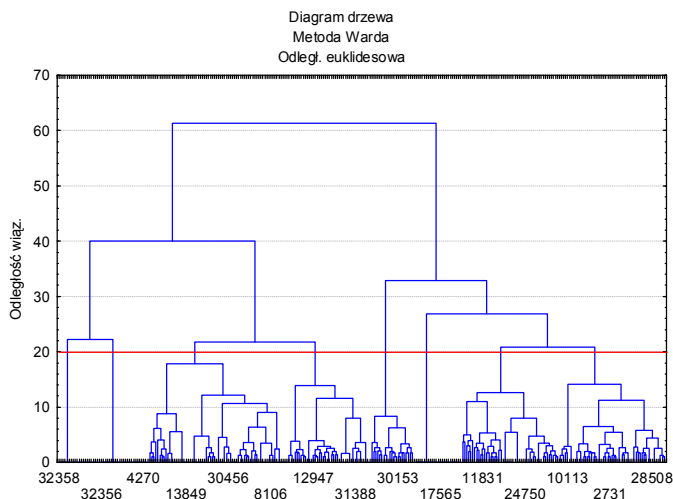
Określenie liczby segmentów

Najczęściej wykorzystywane podczas segmentacji metody (metoda k-średnich, EM oraz sieci Kohonena) wymagają, by przed rozpoczęciem analizy podana została liczba skupień, która ma być wynikiem analizy. Dlatego też po przygotowaniu zbioru danych kolejnym zadaniem jest określenie optymalnej liczby segmentów, na jaką należy podzielić analizowaną grupę internautów.

Decyzja o optymalnej liczbie segmentów jest zwykle podejmowana po uwzględnieniu dwóch aspektów związanych z analizą. Po pierwsze musimy mieć na uwadze praktyczny cel wykonywanej analizy. Najczęściej dążymy do tego, by liczba uzyskanych segmentów była niewielka, tak by obsługa wydzielonych segmentów była jak najmniej kłopotliwa. W praktyce raczej nie spotyka się segmentacji wyróżniających więcej niż dziesięć grup. Drugim czynnikiem decydującym o wyborze liczby grup jest zróżnicowanie badanego zbioru i występowanie w nim naturalnych skupień.

W praktyce stosowanych jest kilka strategii identyfikacji optymalnej liczby skupień. Popularna metoda polega na wykorzystaniu do tego celu metody aglomeracyjnej. W metodzie tej dla wylosowanego podzbioru danych budujemy wykres drzewa (w przypadku budowy drzewa dla całej zbiorowości wykres mógłby być nieczytelny), a następnie na odstawie jego analizy ustalamy odpowiedni punkt odcięcia, ustalając tym samym liczbę skupień. Ponieważ wnioski dotyczące liczby skupień wyciągamy na podstawie jedynie podzbioru danych, operację budowy drzewa powtarzamy wielokrotnie w celu sprawdzenia, czy uzyskana liczba skupień jest stabilna.

Operację tę bardzo wygodnie jest wykonać w przestrzeni roboczej *STATISTICA Data Miner*. Przywołujemy ją poleceniem *Statystyka - Data-mining - Data miner - wszystkie procedury*. W przestrzeni roboczej umieszczamy zbiór *Web.sta*, a następnie z przeglądarki węzłów z grupy *Czyszczenie i filtrowanie danych* wybieramy węzeł *Losowy podzbiór przypadków*. Następnie zmieniamy ustawienia dla węzła, tak by wylosowanych zostało około 300 przypadków z całego zbioru. Do wylosowanego zbioru danych dołączamy węzeł *Aglomeracja*. W oknie parametrów tej metody zaznaczamy, by grupowane były przypadki, a podczas tworzenia drzewa stosowana była metoda Warda. Następnie uruchamiamy powstały projekt. W wyniku analizy otrzymujemy węzeł raportowy zawierający interesujący nas wykres drzewa.



Analiza powyższego dendrogramu pozwala nam ustalić, jaki jest układ skupień w analizowanej zbiorowości, umożliwia nam również określenie najodpowiedniejszej ich liczby. By móc określić tę liczbę, konieczne jest ustalenie punktu odcięcia dendrogramu (odległości, powyżej której poszczególne elementy traktowane są jako różne). Podejmując decyzję o optymalnym poziomie odcięcia, możemy wykorzystać szereg metod numerycznych lub też określić punkt odcięcia w sposób arbitralny, na przykład kierując się względami praktycznymi. W niniejszej analizie punkt odcięcia ustalono na poziomie 20 (pogrubiona linia). Dla takiego poziomu odległości wiązania można wyróżnić osiem różnych grup internautów.

Operację losowania podzbioru i budowy drzewa przeprowadzono jeszcze kilkanaście razy, by wykluczyć możliwość błędnej oceny liczby skupień wynikającej ze złej struktury wylosowanej próbki. Na podstawie uzyskanych wyników stwierdzono, że liczba skupień, jaką należało by wyróżnić podczas analizy, powinna znajdować się w przedziale od 6 do 8.

Inna metoda ustalania optymalnej liczby segmentów została szeroko opisana w [3]. Metoda ta składa się z trzech etapów. W pierwszym etapie budujemy model metodami niehierarchicznymi (np. metodą k-średnich), wskazując by algorytm utworzył model uwzględniający znaczną liczbę segmentów (np. 25).

W drugim kroku określamy średnie wartości dla uzyskanych segmentów i na ich podstawie budujemy drzewo metodą aglomeracyjną. Analiza uzyskanego drzewa pozwala nam ocenić, które z otrzymanych segmentów są do siebie na tyle podobne, by można je było z sobą połączyć. Na podstawie wiedzy uzyskanej dzięki analizie diagramu drzewa określamy optymalną liczbę skupień. Znając optymalną liczbę skupień, budujemy następnie docelowy model, ponownie używając do tego celu metod niehierarchicznych.

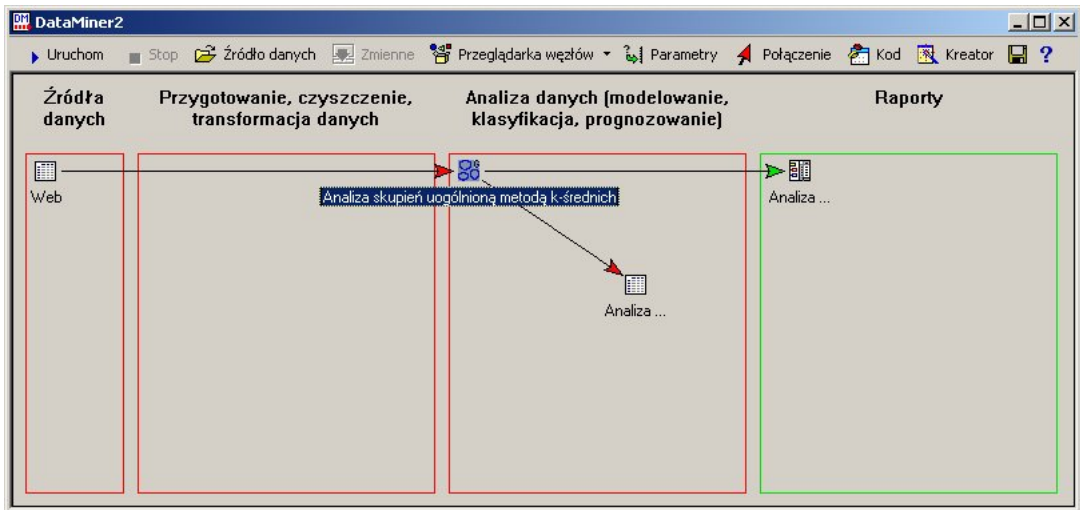
Ponieważ wyniki analizy hierarchicznej nie pozwoliły w sposób jednoznaczny określić liczby segmentów, do jej określenia wykorzystany zostanie zaimplementowany w programie *STATISTICA Data Miner* mechanizm automatycznego wyboru optymalnej



liczby skupień. Mechanizm ten dostępny jest dla metody k -średnich oraz EM i jest oparty na v -krotnym teście krzyżowym. Algorytm działa w ten sposób, że dzieli zbiór wejściowy na kolejno coraz większą liczbę segmentów i ocenia precyzję podziału dla każdego z nich. Jeśli w wyniku kolejnego podziału zbudowany model poprawia się w stosunku do poprzedniego modelu w mniejszym stopniu niż określono to w wartości progowej (domyślnie jest to 5%), algorytm zatrzymuje swoje działanie (dodanie kolejnego segmentu w znaczący sposób nie poprawia wyników modelu). Dla metody k -średnich miarą precyzji podziału jest przeciętna odległość elementów zbioru wejściowego od środka segmentu, w jakim się znajdują, w przypadku metody EM miara ta bazuje na prawdopodobieństwie przynależności do odpowiednich segmentów.

W niniejszym przykładzie do segmentacji wykorzystana zostanie uogólniona metoda k -średnich. W celu wykonania analizy z przeglądarki węzłów wybieramy moduł *Uogólniona metoda k -średnich* i łączymy go ze zbiorem *Web.sta*. W oknie edycji parametrów na zakładce *V-krotny sprawdzian krzyżowy* zaznaczamy, by był on wykonywany. Warto w tym momencie wykorzystać informację o orientacyjnej liczbie segmentów, uzyskaną na podstawie wyników analizy aglomeracyjnej. Dzięki tej informacji możliwe jest znaczne zawężenie przestrzeni poszukiwań. W polu *Minimalna liczba segmentów* wpisujemy 6, natomiast w polu *Maksymalna liczba segmentów* wpisujemy 8.

Następnie uruchamiamy proces segmentacji poleceniem *Uruchom*.



Analiza wyników segmentacji

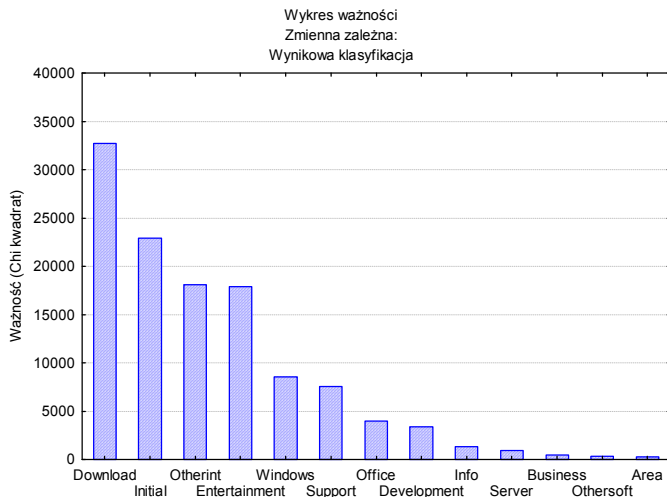
Dla wybranej metody wygenerowane zostały dwa węzły wynikowe (widoczne na rysunku powyżej). Pierwszy z nich zawierał pierwotny arkusz danych z nową zmienną określającą, do którego segmentu przydzieleni zostali poszczególni internauci. Drugi węzeł zawierał raporty przedstawiające szczegółowe wyniki analizy. Poniżej zamieszczono arkusz



z podsumowaniem wykonanej analizy. Na jego podstawie możemy stwierdzić, że algorytm wyróżnił sześć skupień.

Podsumowanie analizy k-średnich (Web)	
Liczba skupień: 6	
Całkowita liczba przypadków uczących: 32711	
Algorytm	k-średnich
Odległość	Euklidesowa
Wstępne środki	Maksymalizuj odległość skupień
BD usuwane przypadkami	Tak
Sprawdzian krzyżowy	3 podzbiorów
Próba testowa	0
Próba ucząca	32711
Błąd w próbie uczącej	0,239731
Liczba skupień	6

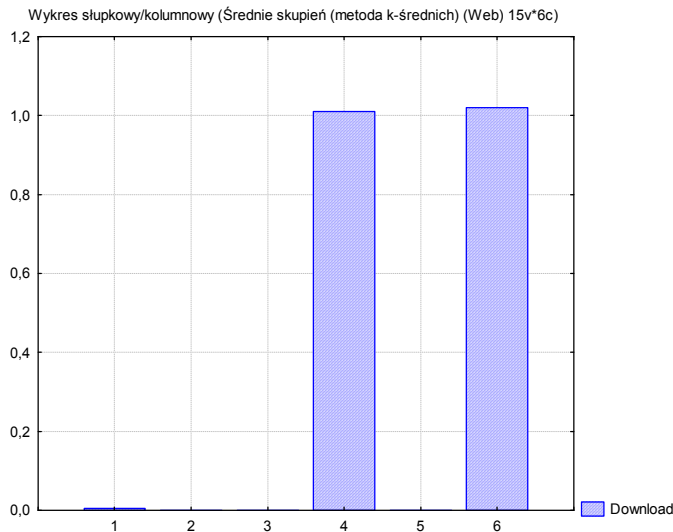
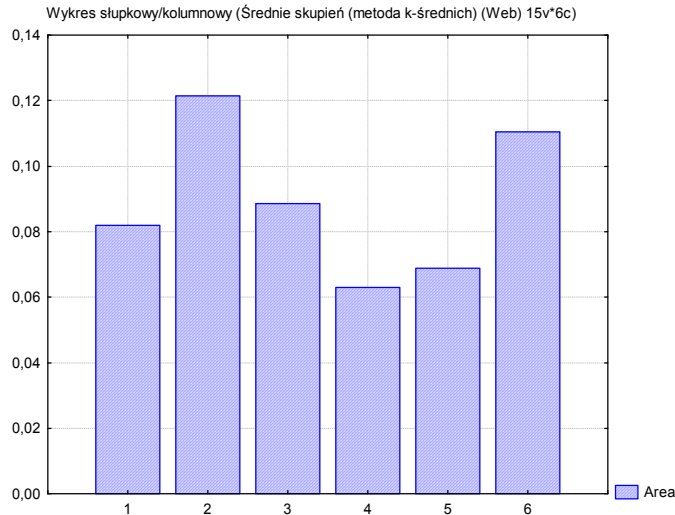
Kluczowym zadaniem na tym etapie analizy jest odpowiednie scharakteryzowanie uzyskanych skupień. Po pierwsze bardzo ważne jest określenie, jakie zmienne miały kluczowy wpływ na proces podziału. Analizę tę możemy wykonać, posługując się węzłem *Dobór zmiennych i analiza przyczyn*, który uruchamiamy dla arkusza wygenerowanego w wyniku analizy. Po jego uruchomieniu otrzymujemy poniższy wykres.



Analizowaną grupę internautów najmocniej różnicowała cecha dotycząca pobierania plików i aktualizacji (*Download*). Kolejne bardzo istotne cechy to: *Initial*, *Otherint* oraz *Entertainment*. Najmniejszy wpływ na proces podziału miały zmienne *Area*, *Othersoft* oraz *Business*. Przedstawione różnice istotności poszczególnych cech mają czytelne odzwierciedlenie w średnich wartościach poszczególnych cech w uzyskanych segmentach.

Poniższe wykresy słupkowe przedstawiają średnie wartości odwiedzin internautów znajdujących się w poszczególnych segmentach na stronach kategorii *Area* oraz *Download*. W przypadku zmiennej *Area* częstości odwiedzin poszczególnych grup są do siebie podobne (zmienna ta nie różnicuje wydzielonych segmentów), w przypadku zmiennej

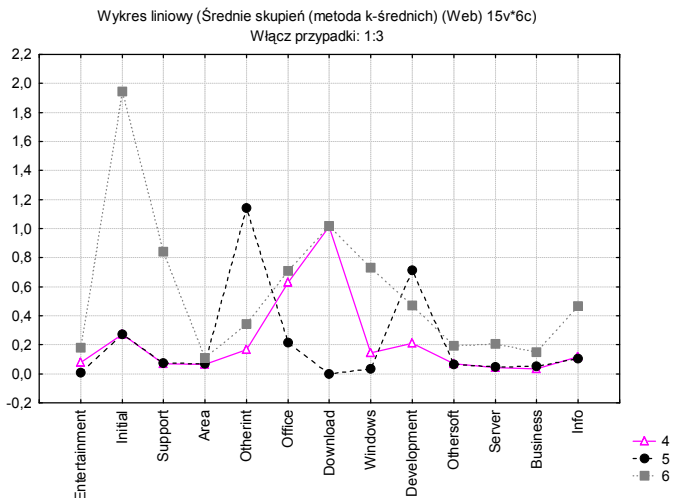
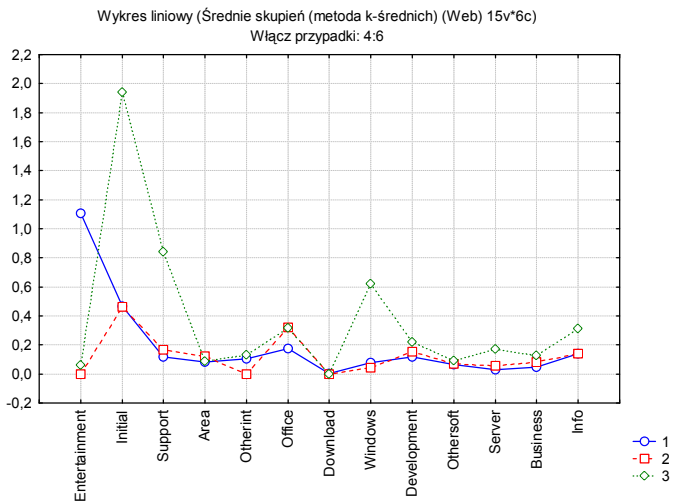
Download możemy zauważyć, że pliki i uaktualnienia były pobierane jedynie przez osoby z segmentu 4 i 6, natomiast osoby z pozostałych segmentów praktycznie nie korzystały z tej możliwości.



Kolejnym zadaniem, jakie należy wykonać, polega na określeniu kluczowych cech uzyskanych segmentów – wskazaniu najbardziej popularnych działań podejmowanych przez osoby znajdujące się w danym segmencie oraz ustaleniu liczności poszczególnych segmentów. Bardzo dobrą praktyką jest nadanie uzyskanym segmentom etykiet umożliwiających ich łatwą identyfikację. Przypisanie nazw może ułatwić komunikację wśród osób korzystających z wyników analizy w przyszłości.



Podczas dokonywania charakterystyki poszczególnych segmentów można skorzystać z wykresów słupkowych analogicznych do przedstawionych powyżej. Innym bardzo wygodnym narzędziem (w przypadku gdy analizowana liczba zmiennych oraz uzyskana liczba skupień nie jest duża) pozwalającym scharakteryzować łącznie wszystkie segmenty jest wykres liniowy przedstawiający środki skupień dla poszczególnych zmiennych. Dla poprawy czytelności wyniki zostaną przedstawione na dwóch wykresach, na każdym z nich przedstawione zostaną trzy segmenty.



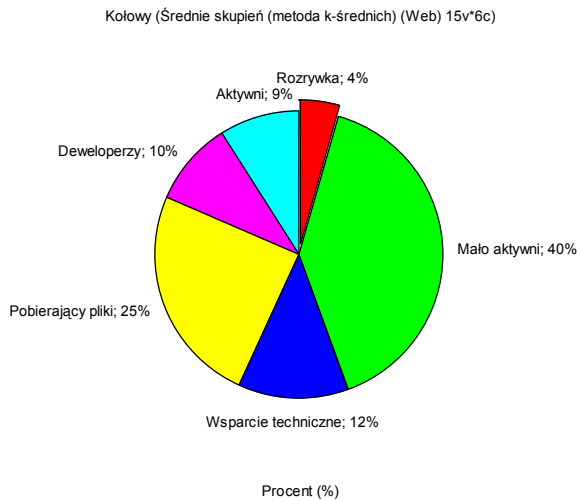
Analizując powyższe segmenty, możemy stwierdzić, że:

- ◆ Osoby należące do pierwszego segmentu odwiedzają jedynie strony związane z rozrywką – grupie tej możemy nadać etykietę *Rozrywka*.



- ◆ Osoby z drugiego segmentu nie wykazują aktywności w żadnej z wyszczególnionych kategorii, odwiedzają jedynie strony startowe i nie wchodzą w głąb serwisu – grupie tej możemy nadać etykietę *Mało aktywni*.
- ◆ Osoby z grupy trzeciej odwiedzają strony z grupy *Initial*, zainteresowani są przede wszystkim stronami dotyczącymi wsparcia technicznego oraz programów *Windows* – grupie nadajemy etykietę *Wsparcie techniczne*.
- ◆ Grupa czwarta to osoby zainteresowane przede wszystkim pobieraniem plików i uaktualnień (*Download*) oraz stron związanych z produktami Office, poza tymi stronami nie odwiedzają już praktycznie żadnych innych stron – grupę tę możemy określić etykietą *Pobierający pliki*.
- ◆ Osoby z grupy piątej odwiedzają przede wszystkim strony przeznaczone dla osób zajmujących się profesjonalnie tematyką IT (*Otherint*) oraz związane z rozwijaniem oprogramowania (*Development*) – grupę tę możemy nazwać *Deweloperzy*.
- ◆ Osoby z grupy szóstej odróżniają się od osób z pozostałych grup poziomem swojej aktywności. Bardzo chętnie odwiedzają strony z kategorii *Initial*, *Support*, *Office*, *Download*, *Windows*, a także jako jedyna grupa strony informacyjne (*Info*) – grupę tę możemy określić jako *Aktywni*.

Za pomocą wykresu kołowego zobaczymy jeszcze, jak wyglądają licznosci w poszczególnych segmentach.



Na podstawie powyższego wykresu możemy stwierdzić, że znaczną grupę (40%) internautów stanowią osoby mało aktywne, osoby z grupy aktywni to jedynie 9% wszystkich gości serwisu. Najmniej liczną grupę (4%) stanowią osoby korzystające ze stron poświęconych rozrywce.



Ocena segmentacji

Ostatni etap analizy polega na interpretacji uzyskanych wyników i ocenie ich praktycznej przydatności. Na tym etapie główną rolę odgrywają osoby mające korzystać z uzyskanych wyników w praktyce, potrafiące stwierdzić, na ile uzyskana wiedza jest interesująca i wartościowa. Przykładowym wnioskiem może być stwierdzenie konieczności skłonienia osób mało aktywnych do szerszego korzystania z zasobów serwisu. Możliwym sposobem osiągnięcia tego celu może być na przykład przebudowa stron z grupy *Initial* (jedynie strony odwiedzane przez segment *Mало aktywni*), by bardziej zachęcały do dalszej eksploracji. Oczywiście po pewnym czasie (np. pół roku) warto przeprowadzić podobną analizę, by móc ocenić, czy podjęte przez nas działania wpłynęły na zachowanie gości naszego serwisu czy też wyróżnione grupy są stabilne.

Literatura:

1. Berry M., Gordon L., *Mastering Data Mining. The Art and Science of Customer Relationship Management*, John Wiley & Sons, Inc, New York 2000.
2. Berson A., Smith S., Thearling K., *Building Data Mining Applications for CRM*, McGraw Hill, New York 2000.
3. Guidici P., *Applied Data Mining Statistical Methods for Business and Industry*, John Wiley & Sons, Inc, 2003.
4. *STATISTICA Neural Networks PL Wprowadzenie do sieci neuronowych*, StatSoft Polska, 2001.
5. *Statystyka i data mining w praktyce*, Materiały z seminariów StatSoft, Warszawa-Kraków 2004.