



ANALIZA SPOSOBU PORUSZANIA SIĘ UŻYTKOWNIKÓW PO PORTALU INTERNETOWYM

Radosław Kita, Onet.pl S.A. Grupa ITI

W poniższych przykładach zostaną wykorzystane dane pochodzące z systemów Onet.pl.

Wprowadzenie

Onet.pl, jak każdy portal internetowy, jest zainteresowany jak najlepszym poznaniem użytkowników odwiedzających jego strony. Jednym z najważniejszych zagadnień jest tutaj analiza wzorców przemieszczania się internautów pomiędzy serwisami. Pozwala ona dostosować rozmieszczenie linków nawigacyjnych do upodobań użytkowników i tym samym ułatwić poruszanie się po portalu. Możemy również myśleć o przedłużeniu ścieżek użytkowników poprzez dodanie odsyłaczy do serwisów zawierających treści, które mogą być potencjalnie dla nich interesujące, a które zostały pominięte.

Zajmiemy się dwoma klasycznymi problemami: analizą ścieżek oraz przewidywaniem, na podstawie stron odwiedzonych wcześniej, serwisu, na który przejdzie internauta.

Analiza ścieżek

Cele analizy

Chcemy poznać sposób, w jaki użytkownicy poruszają się po portalu Onet.pl. Ze zrozumiałych względów ograniczymy się do analizy na poziomie serwisów (nie będzie nas interesował ruch pomiędzy pojedynczymi stronami WWW).

Opis danych

Plik danych wykorzystywany w opisywanym przykładzie zawiera następujące zmienne:

- ◆ **Nazwy przypadków** - identyfikatory użytkowników (wzięte z „cookie”);
- ◆ **Wizyta** - kolejny numer wizyty;
- ◆ **Serwis1, Serwis2, ... Serwis14** - nazwy serwisów według kolejności odwiedzin w trakcie sesji;



Z analizy wyłączono użytkowników, którym nie udało się nadać „cookie”.

Jeden użytkownik może odbyć kilka wizyt na portalu – stąd rozróżnienie pomiędzy identyfikatorem użytkownika a identyfikatorem wizyty.

Poniżej przedstawiono fragment arkusza danych w programie *STATISTICA*, który zawiera opisywane dane.

	1 Wizyta	2 SERWIS1	3 SERWIS2	4 SERWIS3	5 SERWIS4
.2632001000	1	www	film	www	pogoda
.2632001000	2	www	waluty	www	polityka
.2632001000	3	www			
.2633014000	4	www	info	www	info
.2633014000	5	www			
.2633014000	6	www	info	www	waluty
.2638027000	7	www	infoseek		
.2641034000	8	www	ssl	poczta	
.2646047000	9	www	sport	www	info
.2646047000	10	www	sport	www	wycieczki
.2646047000	11	wycieczki			
.2646047000	12	www	info	www	info
.2646050000	13	www	ssl	poczta	
.2652063000	14	www	wiem	www	adresy
.2652063000	15	www	info	www	info
.2652063000	16	www	infoseek	www	info
.2652063000	17	info	www		
.2652063000	18	infoseek	www	info	www

Analiza danych

Zacznijmy od analizy rozkładu liczby serwisów odwiedzanych w trakcie jednej wizyty. Rozkład jest zdecydowanie prawoskośny: średnia liczba odwiedzonych serwisów w trakcie jednej wizyty wynosi: 6,33, a mediana 4,0. W trakcie najdłuższej wizyty odwiedzono 1017 serwisów. Chcemy poznać zachowanie „przeciętnego” użytkownika – nie ma więc potrzeby analizowania wszystkich, dowolnie długich ścieżek. Odrzucimy więc górne 10% przypadków (90. percentyl wynosi 14) i ograniczymy się do 14 serwisów odwiedzonych jako pierwsze.

Do analizy wykorzystamy moduł *Reguły asocjacji (Association Rules)*. Pozwala on określić współwystępowanie sekwencji (asocjacji) wartości zmiennych.



Wyniki uzyskujemy w postaci tabeli:

Summary of association rules (sciezki.sta)						
Min. support = 1,0%, Min. confidence = 1,0%, Min. correlation = 1,0%						
Max. size of body = 10, Max. size of head = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
1	www\	==>	infoseek	30,97492	37,2748	58,80103
2	www\	==>	polityka	2,61003	3,1409	16,40787
3	www\	==>	info	17,93756	21,5858	45,32701
4	www\	==>	wiem	1,74002	2,0939	12,11570
5	www\	==>	tygodnikforum	1,67605	2,0169	13,64073
6	www\	==>	muzyka	3,03224	3,6490	18,67381
7	www\	==>	czat	7,80450	9,3918	28,98332
8	www\	==>	ssl	17,80962	21,4319	44,36087
9	www\	==>	sport	13,68987	16,4742	40,03113
10	www\	==>	katalog	7,12641	8,5758	28,12206
11	www\	==>	altavista	2,14944	2,5866	15,36777
12	www\	==>	kiosk	2,40532	2,8945	16,70517
13	www\	==>	aukcje	2,25179	2,7098	15,40370
14	www\	==>	technopolis	2,57165	3,0947	17,46190
15	www\	==>	poczta	15,68577	18,8761	41,99852
16	www\	==>	republika_portal	4,66991	5,6197	17,65592
17	www\	==>	biznes	1,56090	1,8784	12,15918
18	www\	==>	gielda	1,57369	1,8938	12,29859
19	www\	==>	kartki	2,03429	2,4480	12,78848
20	www\	==>	waluty	3,30092	3,9723	18,39115
21	www\	==>	czytelnia	1,42016	1,7090	12,89974
22	www\	==>	rozrywka	3,97902	4,7883	20,77597
23	www\	==>	film	2,57165	3,0947	16,30416
24	www\	==>	zakupy	1,52252	1,8322	13,26009
25	www\	==>	pogoda	2,26459	2,7252	15,60814
26	www\	==>	wycieczki	1,38178	1,6628	12,33653
27	www\	==>	infoseek, info	2,82753	3,4026	18,32225
28	www\	==>	infoseek, wiem	1,33060	1,6012	12,24848

Pozwala ona wyrazić zależności pomiędzy wartościami zmiennych za pomocą trzech parametrów:

- ♦ **Support (wsparcie)** – określa, jaki procent wszystkich asocjacji stanowi analizowana asocjacja, np. przejścia ze strony głównej do serwisu „Infoseek” (zapisane jako: „www => infoseek”) stanowią 30,97%.
- ♦ **Confidence** – określa, jaki procent asocjacji, które zaczynają się od danego poprzednika (*Head*), stanowi rozważana asocjacja, np. w naszym przypadku sekwencje prowadzące ze strony głównej do serwisu „Infoseek” stanowią 37,27% procent wszystkich ścieżek zaczynających się od strony głównej.
- ♦ **Correlation** - jest to wartość wsparcia dla całej reguły podzielona przez pierwiastek kwadratowy iloczynu wsparcia dla poprzednika (*Head*) oraz wsparcia następnika (*Body*).

Posortujmy teraz wyniki malejąco według wartości *Support*:

Summary of association rules (sciezki.sta)						
Min. support = 1,0%, Min. confidence = 1,0%, Min. correlation = 1,0%						
Max. size of body = 10, Max. size of head = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
1	www\	==>	infoseek	30,97492	37,2748	58,80103
55	infoseek	==>	www\	30,97492	92,7586	58,80103
3	www\	==>	info	17,93756	21,5858	45,32701
81	info	==>	www\	17,93756	95,1799	45,32701
8	www\	==>	ssl	17,80962	21,4319	44,36087
118	ssl	==>	www\	17,80962	91,8206	44,36087
123	ssl	==>	poczta	16,53019	85,2243	91,61066
169	poczta	==>	ssl	16,53019	98,4756	91,61066
15	www\	==>	poczta	15,68577	18,8761	41,99852
165	poczta	==>	www\	15,68577	93,4451	41,99852
47	www\	==>	ssl, poczta	15,53224	18,6913	41,90809
128	ssl	==>	www, poczt	15,53224	80,0792	89,04793
235	www, s:	==>	poczta	15,53224	87,2126	89,83222
174	poczta	==>	www, s:	15,53224	92,5305	89,83222
302	ssl, poczta	==>	www\	15,53224	93,9628	41,90809
254	www, poczt	==>	ssl	15,53224	99,0212	89,04793
9	www\	==>	sport	13,68987	16,4742	40,03113
137	sport	==>	www\	13,68987	97,2727	40,03113
7	www\	==>	czat	7,80450	9,3918	28,98332
109	czat	==>	www\	7,80450	89,4428	28,98332
10	www\	==>	katalog	7,12641	8,5758	28,12206
148	katalog	==>	www\	7,12641	92,2185	28,12206
86	info	==>	sport	6,32037	33,5370	38,80871
139	sport	==>	info	6,32037	44,9091	38,80871

W ten sposób uzyskaliśmy informacje o głównych ścieżkach internautów. Najczęściej wybierana ścieżka prowadzi ze strony głównej do serwisu „Infoseek” - 30,97%. Potem kolejno: ze strony głównej do serwisu „info” (zawierającego wiadomości) - 17,94%, do serwisu służącego do logowania „ssl” - 17,81%, a z serwisu „ssl” do serwisu pocztowego - 16,53% itd.

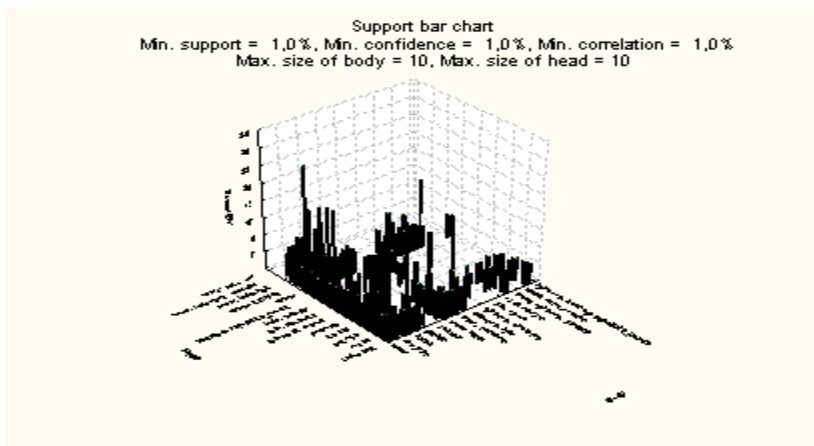
Bardzo przydatnym raportem dostępnym w *STATISTICA Data Miner* jest macierz ilustrująca *Support* dla poszczególnych asocjacji (zob. poniżej).

Dla przykładu: jeśli przyjrzymy się ostatniemu wierszowi zamieszczonego tutaj fragmentu macierzy, zobaczymy, że 2,83% wizyt prowadziło ze strony głównej do serwisu „Infoseek”, a potem do serwisu „info”.



		Support rule matrix (sciezki.sta)									
		Min. support = 1,0%, Min. confidence = 1,0%, Min. correlation = 1,0%									
		Max. size of body = 10, Max. size of head = 10									
Body/Head		infoseek	polityka	info	wiem	tygodnikforum	muzyka	czat	ssl	sport	katalog
www		30,97492	2,610031	17,93756	1,740020	1,676049	3,032242	7,804504	17,80962	13,68967	7,12644
infoseek				2,86592	1,420164			1,125896	2,39253	2,26459	6,06444
polityka				1,57369							
info		2,86592	1,573695				1,215455		1,77840	6,32037	
wiem		1,42016									
tygodnikforum											
muzyka				1,21546							
czat		1,12590							1,95752		
ssl		2,39253		1,77840				1,957523		1,21546	
sport		2,26459		6,32037							
katalog		6,06448									
altavista		1,59928									
kiosk				1,31781							
aukcje											
technopolis				1,29222							
poczta		1,98311		1,56090				1,049130	16,53019	1,07472	
republika_portal		3,74872									1,4073
biznes											
gielda											
kartki											
waluty				1,24104							
czytelnia											
rozrywka											
film											
zakupy											
pogoda											
wycieczki											
www, infoseek				2,82753	1,330604			1,113101	2,37973	2,26459	5,7958

Możemy też sporządzić wykres pozwalający ogarnąć „jednym rzutem oka” całość sytuacji:



Podobną analizę przeprowadzimy dla *Confidence*. Ten parametr będziemy interpretować jako prawdopodobieństwo, że użytkownik, który już odwiedził określone serwisy, zdecyduje się odwiedzić następnym.

Spójrzmy na macierz pokazującą nasz parametr dla poszczególnych asocjacji. Tym razem posortowaliśmy malejąco macierz względem zawartości pierwszej kolumny:



		Confidence rule matrix (ścieżki)						
		Min. support = 1,0%, Min. confidence = 1,0%, Min. correlation = 1,0%						
		Max. size of body = 10, Max. size of head = 10						
Body\Head		infoseek	polityka	info	wiem	tygodnikforum	muzyka	czat
www, katalog, republika_portal	84,3137%							
katalog, republika_portal	81,8181%							
www, katalog	81,3285%							
katalog	78,4768%							
www, wiem	76,4705%							
www, republika_portal	75,0684%							
www, altavista	70,8333%							
altavista	67,9347%							
wiem	57,2164%							
republika_portal	44,5288%							
www	37,2748%	3,14087%		21,5858%	2,09391%		2,01693%	3,64896%
www, sport	16,5420%			45,9813%				
sport	16,0909%			44,9090%				
www, info	15,7632%	8,63052%					6,77603%	
info	15,2070%	8,35030%					6,44942%	
www, czat	14,2623%							
www, ssl	13,3620%			9,9856%				10,41%
czat	12,9032%							
www, ssl, czat	12,7677%			9,9856%				10,41%

Widzimy, że aż 84,31% użytkowników, którzy ze strony głównej przeszli do serwisu „katalog”, by potem odwiedzić strony serwisu „republika_portal”, skierowało się do „Infoseeka”. Okazuje się, że w przypadku dużej części wizyt internauci przechodzą z serwisu „republika_portal” (związanego ze stronami darmowymi) do serwisu „Infoseek”. Jest to o tyle zaskakujące, że nie ma bezpośredniego linku pomiędzy tym częściami portalu. Tak więc trzeba tam umieścić link. Podobnie link do serwisu wyszukiwawczego powinien być umieszczony na stronach serwisów: „katalog”, „wiem”, „sport”, „info” oraz „czat”. Możemy powtórzyć naszą analizę dla każdego z serwisów, sortując macierz względem kolejnych kolumn. Dla przykładu: w zamieszczonym powyżej fragmencie macierzy możemy zauważyć silny związek pomiędzy serwisami „sport” oraz „info”.

Wychwycenie podobnych związków w naturalny sposób budzi chęć przeprowadzenia analizy predykcyjnej prowadzącej do personalizacji stron pokazywanych internaucie w zależności od stron odwiedzonych wcześniej. Taką analizę pokażemy w następnym przykładzie.

Pozostał nam jeszcze do omówienia ostatni z parametrów - *Correlation*. Ma on trochę bardziej złożoną interpretację. Można go traktować jako informację o tym, jaka część dostępnych „zdarzeń” została „wykorzystana” w danej asocjacji. Dla przykładu: *Correlation* równe 100% dla $A \Rightarrow B$ oznaczałoby, że zdarzenia A oraz B pojawiają się jedynie w wymienionej asocjacji.

Jeśli posortujemy teraz tabelę, od której zaczęliśmy analizę, malejąco właśnie według wartości *Correlation*, to zobaczymy, że asocjacja serwis pocztowy – serwis „ssl” prawie całkowicie wyczerpuje wszystkie wystąpienia tych serwisów w ścieżkach internautów.



Summary of association rules (ściezki)						
Min. support = 1,0%, Min. confidence = 1,0%, Min. correlation = 1,0%						
Max. size of body = 10, Max. size of head = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
169	poczta	==>	ssl	16,5301%	98,475%	91,6106%
123	ssl	==>	poczta	16,5301%	85,224%	91,6106%
236	www, ssl	==>	poczta	15,5322%	87,212%	89,8322%
174	poczta	==>	www, ssl	15,5322%	92,530%	89,8322%
128	ssl	==>	www, poczta	15,5322%	80,079%	89,0479%
254	www, poczta	==>	ssl	15,5322%	99,021%	89,0479%
1	www	==>	infoseek	30,9749%	37,274%	58,8010%
55	infoseek	==>	www	30,9749%	92,758%	58,8010%
81	info	==>	www	17,9375%	95,179%	45,3270%
3	www	==>	info	17,9375%	21,585%	45,3270%
8	www	==>	ssl	17,8096%	21,431%	44,3608%
118	ssl	==>	www	17,8096%	91,820%	44,3608%
15	www	==>	poczta	15,6857%	18,876%	41,9985%
165	poczta	==>	www	15,6857%	93,445%	41,9985%
302	ssl, poczta	==>	www	15,5322%	93,962%	41,9080%
47	www	==>	ssl, poczta	15,5322%	18,691%	41,9080%
9	www	==>	sport	13,6898%	16,474%	40,0311%
137	sport	==>	www	13,6898%	97,272%	40,0311%
140	ssl, poczta	==>	www, info	8,2047%	44,737%	20,5109%

Równie silny związek widzimy w przypadku ścieżek prowadzących ze strony głównej poprzez serwis „poczta” do serwisu „ssl”. Tak silne połączenie pomiędzy serwisem „ssl” oraz „poczta” jest wymuszone przez konieczność logowania się użytkownika przed przeczytaniem poczty.

Na uwagę zasługuje *Correlation* większa niż 50% w przypadku strony głównej i serwisu wyszukiwawczego „Infoseek”. Jeśli przypomnimy sobie wnioski z analizy parametru *Confidence*, to możemy przyjąć, że dodanie linku do serwisu „Infoseek” na stronach wymienionych wyżej serwisów znacznie podniosłoby ruch na serwisie wyszukiwawczym.

Przewidywanie następnego serwisu w ścieżce

Cele analizy

Chcemy zbudować model pozwalający jak najlepiej przewidywać, na jaki kolejny serwis skieruje się użytkownik na podstawie trzech ostatnio odwiedzonych serwisów. Model może zostać wykorzystany do dynamicznej personalizacji stron – umieszczania linków nawigacyjnych odpowiednich do przewidywanego następnego kroku internauty.

Opis danych

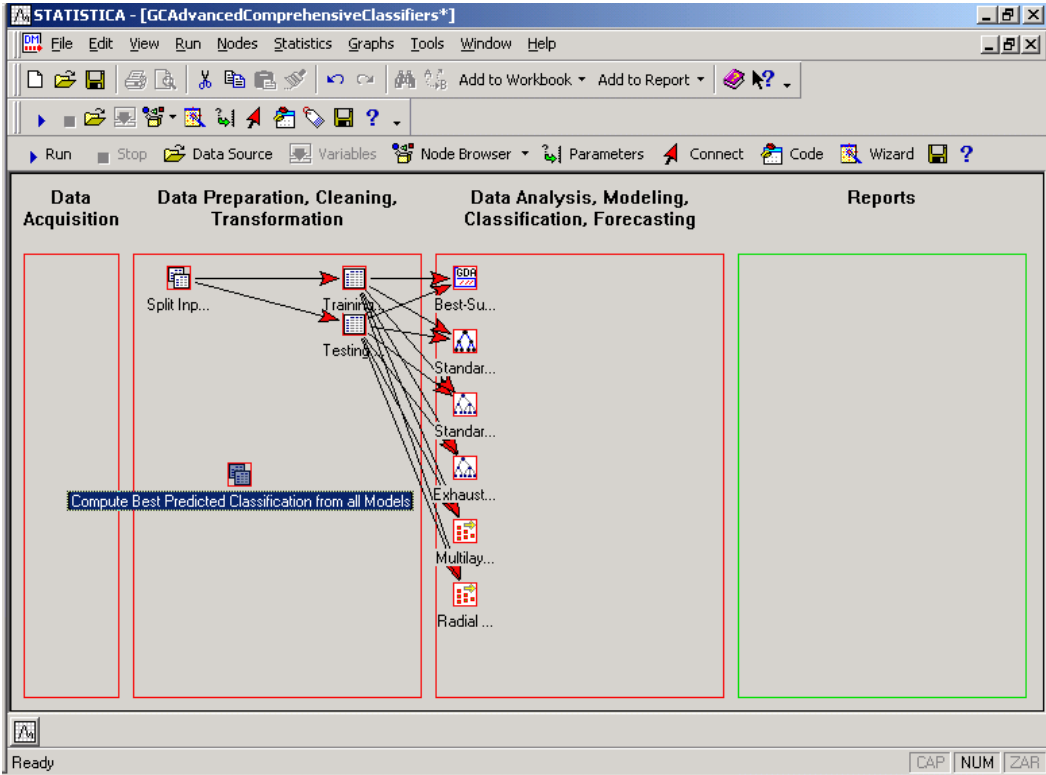
Plik danych zawiera tylko cztery zmienne:

- ♦ **Serwis1, Serwis2, ... Serwis4.** Są to nazwy serwisów według kolejności odwiedzin w trakcie sesji.



Analiza danych

W analizie wykorzystamy gotowy projekt załączony do *STATISTICA Data Miner*:



Pozwala on, po wskazaniu źródła danych, przygotować i porównać ze sobą przewidywania modeli zbudowanych w oparciu o: *Best-Subset and Stepwise GDA ANCOVA*, drzewa decyzyjne budowane metodami: CART, CHAID i Exhaustive CHAID oraz dwie sztuczne sieci neuronowe: perceptron wielowarstwowy i sieci o radialnych funkcjach bazowych.

Po przygotowaniu modeli łączymy dane testowe przygotowane przez każdy z modeli z węzłem „Goodness of Fit for Multiple Inputs”. Pozwala on porównać wartości przewidywane przez poszczególne modele z rzeczywistymi. Uzyskaliśmy szereg raportów porównujących stopień dopasowania modeli. Dla nas największe znaczenie będzie miało zestawienie:

Summary Goodness of Fit (Spreadsheet407)			
Observed variable: SERVMS4			
	1	2	3
	Chi-square statistic	G-square statistic	Percent disagreement
Testing_GDA5(Predicted 1)	5,14444444	10,0257846	0,346153846
Testing_CTrees6(Predicted 1)	4,36904762	6,99803746	0,269230769
Testing_CCHAID7(Predicted 1)	7,78321678	10,704226	0,458333333
Testing_CECHAID8(Predicted 1)	3,04329004	10,2337116	0,291666667
Testing_CMLP9(SERVMS4.1)	2,60769231	10,8990267	0,269230769
Testing_CRBF10(SERVMS4.1)	2,66666667	12,9594043	0,307692308



Najlepiej dopasowany okazał się model oparty na drzewach decyzyjnych zbudowanych metodą CART. Przewidywania oparte na nim są błędnie tylko w 27% przypadków. Godne uwagi są również modele zbudowane przez sztuczne sieci neuronowe.

Teraz wystarczy wygenerować kod opisujący wybrany model i zintegrować go z systemem podawania stron.

Literatura

1. Mattison R., 1999, Web Warehousing and Knowledge Management, McGraw-Hill, New York.
2. Witten I.H., 2002, Data Mining, Morgan Kaufman, San Francisco California.