DRZEWA KLASYFIKACYJNE W IDENTYFIKACJI PREFERENCJI KLIENTÓW E-HANDLU

TOMASZ ZDZIEBKO, PIOTR SULIKOWSKI

Streszczenie

Celem artykułu jest prezentacja pełnych wyników badań identyfikacji preferencji użytkowników serwisów e-handlu z wykorzystaniem metody drzew klasyfikacyjnych. Badania te doprowadziły do opracowania modeli klasyfikacyjnych, zbudowanych w oparciu o wskaźniki pozyskane z wykorzystaniem pośredniej informacji zwrotnej. Artykuł stanowi punkt wyjścia do dalszych rozważań nad problematyką systemów rekomendacyjnych.

Słowa kluczowe: preferencje klientów, e-handel, drzewa klasyfikacyjne, systemy rekomendacyjne

Wprowadzenie

Dynamiczny rozwój e-handlu, jaki obserwujemy od początku jego narodzin wynika z wielu jego zalet w stosunku do handlu tradycyjnego. Jedną z nich jest szersza oferta e-sklepów, lecz stanowi ona jednocześnie poważne wyzwanie. Niestety wybór spośród nierzadko dziesiątek tysięcy towarów zmusza klientów do zwiększonego wysiłku, aby mogli odnaleźć towary spełniające ich indywidualne potrzeby. W miejsce tradycyjnego sprzedawcy służącego poradą w sklepach internetowych coraz częściej wykorzystuje się tzw. *systemy rekomendacyjne*, których zadaniem jest dostarczanie użytkownikom rekomendacji produktów, informacji i usług. Najlepsze efekty dają rekomendacje dopasowane do indywidualnych preferencji i potrzeb poszczególnych klientów. Nim jednak dostarczenie takich rekomendacji stanie się możliwe, konieczna jest jak najdokładniejsza identyfikacja i charakterystyka klientów. W pracy preferencje rozumiane są jako zbiór wartości określających postawy klientów wobec towarów, które wzbudziły ich zainteresowanie. Postawy te mierzone są poziomem zainteresowania wyznaczanym dla każdego z towarów oglądanych przez poszczególnych użytkowników.

Problemem tym zajmuje się dziedzina określana terminem *modelowanie użytkowników* (ang. *user modeling*). Według R. B. Allena model użytkownika stanowi jego opis, stworzony lub wyselekcjonowany przez system w celu ułatwienia interakcji między systemem a użytkownikiem [1, 511–543]. G. Fischer twierdzi, że jest to model, który system posiada nt. użytkownika w środowisku informatycznym [5]. Z kolei M. Próchnicka definiuje modelowanie użytkownika – jako „proces prowadzący do stworzenia obrazu użytkownika" [14]. Cyfrową reprezentację modelu użytkownika stanowi profil użytkownika, który odzwierciedla jego upodobania, zainteresowania i preferencje [12]. Proces modelowania preferencji korzysta z dorobku technik informacji zwrotnej (ang. *relevance feedback*), które mogą być realizowane w sposób bezpośredni (ang. *explicit feedback*) lub pośredni (ang. *implicit feedback*).

Bezpośrednie pozyskiwanie preferencji polega na pytaniu użytkowników o nie. Jak zauważa J. Nielsen, to właśnie „użytkownicy z reguły wiedzą najlepiej o własnych preferencjach, potrzebach i celach" [13]. Bezpośrednie pozyskiwanie preferencji e-klientów może być realizowane np. poprzez formularze do oceny satysfakcji z produktu w skali kilkustopniowej bądź zbieranie opinii o produkcie w formie tekstowych komentarzy. Niestety w wielu przypadkach pytanie użytkowników o ich

preferencje jest niemożliwe lub niepożądane. Sytuacja taka występuje z reguły w serwisach internetowych, których użytkownicy zwykle chcą od razu przystąpić do realizacji swoich celów i nie są zainteresowani dodatkowymi, zajmującymi czas czynnościami. Jak dowiodły badania, pytanie użytkowników o preferencje zakłóca realizację zadań, a nawet ich irytuje [10, 168–175]. Obserwacja zachowań użytkowników prowadzi do wniosku, że są oni niechętni do podejmowania dodatkowych działań, jeśli w ich mniemaniu nie przyniosą im one korzyści [7, 35–92]. Prowadzi to niestety bardzo często do całkowitej rezygnacji z wyrażania swoich opinii o oglądanych towarach [2, 40–88].

Ze względu na powyższe problemy, bezpośrednie pozyskanie preferencji użytkowników w praktyce nie przynosi zadowalających efektów. W literaturze przedmiotu proponuje się zastosowanie techniki *pośredniej informacji zwrotnej* (ang. *implicit feedback*). Technika ta polega na ukrytej obserwacji zachowań użytkowników w trakcie ich interakcji z systemem – stroną internetową. Dane zebrane podczas obserwacji wykorzystywane są w procesie wnioskowania o zainteresowaniach i preferencjach użytkowników. Mimo tego, że technika ta jest z reguły mniej dokładna [16, 55–60], jest ona pozbawiona istotnych wad techniki bezpośredniej, ponieważ odkrywanie preferencji odbywa się w sposób niewidoczny dla użytkownika, nierozpraszający i niewymagający od niego żadnych dodatkowych aktywności. W naukowych badaniach poznawczych obserwacja uczestników odbywa się często z wykorzystaniem specjalistycznej aparatury badawczej umożliwiającej szczegółowe śledzenia zachowania i odczuć użytkowników w trakcie interakcji ze stroną. Wykorzystuje się w tym celu m.in. eye-trackery czy encefalografy. Jednakże, mimo że urządzenia takie dostarczają dodatkową wartość poznawczą, ich zastosowanie ograniczone jest jedynie do badań laboratoryjnych. W związku z tym wnioski płynące z tych badań, odnoszące się np. do koncentracji użytkowników na określonych obszarach strony, nie zawsze mogą być zastosowane w praktyce ze względu na brak możliwości nieskomplikowanego monitoringu tego aspektu interakcji. Wymóg stosowalności uzyskanych wyników sprawia, że badania w obszarze pośredniej informacji zwrotnej opierają się najczęściej na wykorzystaniu jedynie technik monitoringu zachowania użytkowników możliwych do zastosowania w przeglądarkach internetowych.

1. Badania w obszarze identyfikacji preferencji e-klientów

Dane rejestrowane podczas obserwacji zachowań internautów wykorzystywane mogą być do wyznaczania różnorodnych wskaźników opisujących wielorakie aspekty interakcji ze stroną internetową. Analiza literatury wskazuje na pewien zbiór wskaźników, który pozwala na wnioskowanie o zainteresowaniu użytkowników produktem, usługą lub informacją. Jednakże wśród badaczy nie ma konsensusu co do konkretnego znaczenia przypisywanego poszczególnym parametrom zachowań. Niektórzy badacze np. uważają, że zmierzony dystans przewijania strony jest pozytywnie skorelowany z poziomem zainteresowania, podczas gdy inne badania nie potwierdzają takiej zależności. Rozbieżności tego typu wynikać mogą z faktu, że omawiane zjawisko jest na tyle złożone, iż nie ma mowy o jednej, najlepszej metodzie modelowania preferencji użytkowników z wykorzystaniem techniki pośredniej informacji zwrotnej. Jak zauważa D. Kelly, wciąż istnieje spory niedostatek badań w tym obszarze [11, 169–186]. Problem ten zauważają również inni badacze, wśród których należy wymienić: J. Goecksa i J. Shavlika [6] oraz G. Velayathana i S. Yamadę [15].

Sytuacja ta jest po części efektem stosowania różnych metod w zakresie: zbierania danych, obliczania wskaźników zachowania oraz modelowania preferencji. Nie wszystkie wskaźniki mogą świadczyć jednakowo o zainteresowaniu. Ponadto Demski podnosi, iż w praktyce w modelowaniu

zwykle stosuje się wskaźniki obliczone wskutek przekształceń danych źródłowych [4, 53–57]. Dlatego niektóre ze wskaźników mogą być użyteczne dopiero w kombinacji z innymi wskaźnikami.

Istotny problem badaczy stanowi również pozyskanie rzeczywistych danych o zachowaniu użytkowników, gdyż są one jednym z najpilniej strzeżonych zasobów firm prowadzących działalność handlową w sieci.

Szczególnie duży niedostatek badań występuje w obszarze identyfikacji preferencji użytkowników serwisów handlu elektronicznego. Znaleźć można natomiast pewien zbiór wartościowych prac [3], w których badacze monitorowali zachowanie użytkowników podczas wizyt w różnych rodzajach serwisów internetowych, niekoniecznie handlowych. Badacze ci prezentowali różnorodne podejścia do modelowania preferencji, polegające na analizie korelacji albo wykorzystaniu drzew klasyfikacyjnych. Wyniki tych badań dostarczają wartościowej wiedzy, nie ujmują jednak specyfiki serwisów handlu elektronicznego.

W systemach rekomendacyjnych, podobnie jak w innych systemach realizujących zadania eksploracji danych czy też analizy *stricte* statystyczne, stosuje się wiele metod modelowania, m.in. algorytm *k*-najbliższych sąsiadów, algorytmy genetyczne, drzewa decyzyjne, grupowanie hierarchiczne i metodą *k*-średnich, reguły asocjacyjne, regresję liniową, regresję logistyczną, sieci bayesowskie, sieci neuronowe, zbiory przybliżone itp. Nie można jednak wskazać najlepszej z nich. W każdym przypadku wybór powinien zależeć od szczegółów związanych z problemem. Dla problemu klasyfikacji, rozważanego w opisywanych dalej badaniach, wzięto pod uwagę cel klasyfikacji, a także w szczególności strukturę danych, wykorzystywane charakterystyki oraz zakres, w jakim można rozdzielić klasy [8, 523–541].

2. Procedura badawcza

W celu wypełnienia opisanej powyżej luki badawczej na potrzeby przeprowadzenia badania opracowano autorskie rozszerzenie ECPM (ang. E-commerce Customer Preference Monitor) dla przeglądarki Mozilla Firefox. Wybór tej technologii został podyktowany wysoką popularnością przeglądarki, możliwością prowadzenia badań w dowolnie wybranych serwisach e-handlu oraz dobrze udokumentowaną procedurą tworzenia rozszerzeń. Ważny był również aspekt praktycznej możliwości wykorzystania uzyskanych wyników. Do monitorowania zachowania użytkowników ECPM wykorzystuje pierwszy poziom obiektowego modelu dokumentu(ang. DOM Level 1). Mechanizm ten jest zaimplementowany w prawie wszystkich obecnych przeglądarkach internetowych. Umożliwia to wykorzystanie metod pozyskiwania pośredniej informacji zwrotnej zaimplementowanych w ECPM do monitorowania dowolnych stron z sektora e-commerce i nie tylko. Uczestnicy przystępujący do badania musieli jedynie zainstalować odpowiednie rozszerzenie. Należy zauważyć, że w celu ochrony prywatności każdy uczestnik mógł wyłączyć rozszerzenie lub odinstalować je w dowolnym momencie.

W trakcie oglądania stron z oferowanymi produktami rozszerzenie monitorowało aktywność uczestników badania w obrębie następujących pięciu polskich serwisów e-handlu: agito.pl, komputronik.pl, electro.pl. morele.net, merlin.pl. Wybór tych serwisów wynikał z ich dużej popularności oraz konsekwentnego interfejsu pozwalającego na intuicyjną implementację monitorowania zachowań użytkowników. Argumentem przemawiającym za wyborem tych serwisów było też umożliwienie uczestnikom badania odnalezienie towarów zaspokajających ich różnorodne potrzeby. Należy zauważyć, że rozszerzenie ECPM monitorowało aktywność uczestników badania tylko w obrębie tych sklepów (z wyłączeniem innych odwiedzanych stron internetowych).

Na podstawie zgromadzonych danych rozszerzenie następnie wyliczało kilkadziesiąt wskaźników opisujących zachowania użytkowników na danej stronie, jak również parametry charakteryzujące poszczególne odwiedzane strony. W momencie opuszczania strony zawierającej informacje o produkcie respondenci byli pytani jawnie o ocenę zainteresowania produktem znajdującym się na tej stronie – w pięciostopniowej skali, gdzie wartości 1 przypisano znaczenie „nieinteresujący" a wartości 5 – „bardzo interesujący". Uczestnik badania mógł również określić to, czy znał wcześniej (przed badaniem) oglądany produkt. Na bazie zarejestrowanych parametrów wyznaczane były relatywne wskaźniki zachowań, mające na celu lepsze odzwierciedlenie aktywności internautów w stosunku do zawartości strony. Pełna lista wskaźników zachowań oraz parametrów rejestrowanych dla każdej odwiedzonej przez użytkownika strony została przedstawiona w Tabeli 1.

Tabela 1. Wskaźniki zachowań i parametry stron rejestrowane przez rozszerzenie

Parametr	Objaśnienie
mark	ocena zainteresowania produktem
familiar	wcześniejsza znajomość produktu
Parametry określające cechy strony produktu	
document_length	liczba znaków tekstu zawartego na stronie
desc_length	liczba znaków opisu produktu zawartego na stronie
review_length	liczba znaków opinii o produkcie zawartego na stronie produktu
recommend_length	liczba znaków tekstów dotyczących rekomendowanych produktów
image_number	liczba zdjęć produktu zawartych na stronie
page_height	wysokość zawartości strony w pikselach
Parametry określające czasy interakcji	
page_time	czas otwarcia strony
tab_activ_time	czas aktywności karty
user_activ_time	czas aktywności użytkownika
prod_desc_time	czas przebywania kursora w obrębie opisu produktu
prod_recommend_time	czas przebywania kursora w obrębie rekomendacji
prod_review_time	czas przebywania kursora w obrębie opinii o produkcie
prod_image_time	czas przebywania kursora w obrębie zdjęć produktu
prod_other_time	czas przebywania kursora w obrębie pozostałego obszaru
Parametry określające zachowanie użytkowników	
mouse_distance	dystans kursora o jaki został przesunięty kursor myszy
vertical_scroll	dystans przewijania strony w pionie
horizontal_scroll	dystans przewijania strony w poziomie
mouse_clicks	liczba kliknięć myszy
lb_mouse_clicks	liczba kliknięć lewego klawisza myszy
rb_mouse_clicks	liczba kliknięć prawego klawisza myszy
mb_mouse_clicks	liczba kliknięć środkowego klawisza myszy
copycut_action	liczba zdarzeń kopiowania/wycinania

Parametr	Objaśnienie
select_action	liczba zdarzeń zaznaczania
select_text_size	liczba znaków zaznaczonego tekstu
keydown_single	liczba zdarzeń wielokrotnego wciśnięcia klawisza
keydown_repeatable	liczba zdarzeń pojedynczego wciśnięcia klawiszy
find_action	liczba akcji wyszukiwania
print_action	liczba akcji drukowania
bookmark_action	liczba akcji tworzenia zakładki
save_action	liczba akcji zapisu
resize_action	liczba akcji modyfikowania rozmiaru wyświetlanego dokumentu
search_referral	produkt odszukany poprzez wyszukiwarkę
Relatywne parametry określające zachowanie użytkowników	
rel_page_time	relatywny czas otwarcia
rel_user_activ_time	relatywny czas aktywności użytkownika
rel_tab_active_time	relatywny czas aktywności karty
rel_prod_desc_time	relatywny czas przebywania kursora w obrębie opisu produktu
rel_prod_recommend_time	relatywny czas przebywania kursora w obrębie rekomendacji
rel_prod_review_time	relatywny czas przebywania kursora w obrębie opinii o produkcie
rel_prod_image_time	relatywny czas przebywania kursora w obrębie zdjęć produktu
rel_mouse_distance	relatywny dystans kursora
rel_vertical_scroll	relatywny dystans przewijania strony w pionie
rel_horizontal_scroll	relatywny dystans przewijania strony w poziomie

Źródło: opracowanie własne.

3. Wyniki badania

3.1. Analiza przeglądowa

Badanie poznawcze z udziałem użytkowników prowadzone było w sposób ciągły przez okres 7 miesięcy. Dobór próby do badania został przeprowadzony w sposób uznaniowy (ang. convenience sample) ze względu na ograniczone możliwości. Aby pozyskać możliwie dużą próbę o charakterystyce zbliżonej do populacji, propozycja udziału w badaniu została skierowana do szerokiego i zróżnicowanego kręgu potencjalnych kandydatów. W badaniu prowadzonym na zasadzie dobrowolności udziału uczestniczyło ostatecznie 85 osób. Internauci odwiedzili i ocenili w sumie 1396 produktów w 5 e-sklepach. Minimalna liczba towarów ocenionych przez jednego uczestnika wyniosła 1, a maksymalna 116. Świadczy to o dużej dysproporcji w liczbie ocenianych towarów przez jednego uczestnika. Średnia liczba ocenionych towarów przez jednego respondenta wyniosła 16,42, a odchylenie standardowe – 16. Jedna czwarta uczestników oceniła poniżej 7 towarów, podczas gdy górny kwartyl ocenił więcej niż 20 towarów. Wartość rozstępu międzykwartylowego wyniosła 13.

W związku z istotnymi różnicami w zaangażowaniu uczestników badania podjęto decyzję, że proces budowy modeli zostanie przeprowadzony na dwóch zbiorach danych. Pierwszy z nich stanowiła pełna próba – dane zebrane od wszystkich uczestników. Drugi zbiór danych zaś stanowiły dane zgromadzone od grupy najbardziej aktywnych internautów, którzy w trakcie badania ocenili przynajmniej 30 towarów Do tej grupy zakwalifikowano 10 uczestników, którzy ocenili w sumie 494 towary.

Rozkład częstości bezpośrednich ocen poziomu zainteresowania towarami dla całej badanej populacji został przedstawiony w Tabeli 2. Uczestnicy badania najczęściej przyznawali najwyższą ocenę – 5, a najrzadziej ocenę najniższą – 1.

Tabela 2. Rozkład częstości bezpośrednich ocen poziomu zainteresowania towarami

Ocena	Liczba ocen
1	130
2	180
3	325
4	346
5	415

Źródło: opracowanie własne.

Ze względów obliczeniowych w algorytmach rekomendacyjnych zainteresowanie często wyrażane jest w skali binarnej, dlatego na potrzeby prowadzonych badań dokonano również dodatkowych obliczeń, aby optymalnie przetransponować wyrażenie parametru z 5-stopniowej skali nominalnej do skali binarnej. Wartościom 1 i 2 przyporządkowano nową wartość 0, oznaczającą niewielkie zainteresowanie lub jego brak, natomiast wartościom 3, 4 i 5 – wartość 1, która oznacza zainteresowanie produktem.

3.2. Procedura budowy modeli drzew klasyfikacyjnych

Dobór zmiennych do modelu został przeprowadzony w oparciu o wyniki testu Kruskala-Wallisa oraz analizę współliniowości zmiennych w oparciu o współczynnik tau Kendalla. Na tej podstawie do modelu klasyfikacyjnego włączono piętnaście kluczowych zmiennych objaśniających: *desc_length*, *keydown_single*, *lb_mouse_clicks*, *mouse_clicks*, *mouse_distance*, *page_height*, *page_time*, *prod_desc_time*, *prod_other_time*, *prod_recommend_time*, *rel_prod_review_time*, *search_refferal*, *tab_activ_time*, *user_activ_time*, *vertical_scroll*.

Do budowy modelu klasyfikacyjnego preferencji e-klientów wykorzystano program *SAS Enterprise Miner 6.2* (w skrócie: SAS EM). Jako kryterium optymalizacyjne algorytmu poszukującego najlepszych modeli wybrano błąd klasyfikacji. Oznacza to, że procedura budowy drzewa wybiera ten model, który posiada najniższy błąd klasyfikacji. Jednocześnie określone zostały następujące kryteria stopu, mające na celu zapobiegnięcie nadmiernemu dopasowaniu modeli do danych: minimalny rozmiar liścia = 20, maksymalna liczba gałęzi = 3.

W budowie modeli drzew dla wszystkich uczestników badania selekcję przypadków do próby przeprowadzono metodą doboru zbioru równoważonego (równa liczba przypadków dla każdej z klas zmiennej objaśnianej) w sposób losowy. Taki dobór próby pozwala na lepsze oszacowanie zdolności predykcyjnych modelu dla każdej z klas – budowę modelu bardziej uniwersalnego. Dla zbioru przypadków, przy poziomie zainteresowania wyrażonym w skali 5-stopniowej, wylosowano próbę składającą się z 650 przypadków (po 130 na każdą pierwotną klasę).

3.3. Modele drzew klasyfikacyjnych dla pełnej populacji

W wyniku przeprowadzenia opisanej powyżej procedury postępowania dla danych zarejestrowanych dla całej populacji uzyskano model drzewa klasyfikacyjnego, który cechuje się błędem klasyfikacji na poziomie 59,2 proc., co jest wartością istotnie niższą w stosunku do modelu losowego (80 proc.) Analizując macierz błędnych klasyfikacji (Tabela 3), można zauważyć, że największą trafność predykcji uzyskano kolejno dla klas zainteresowania: 5, 2, 1, 4 i 3. Szczególnie dobrze klasyfikowane były dwie pierwsze klasy: 5 i 2. Stosunkowo dobra trafność klasyfikacji dla obiektów najbardziej interesujących świadczy dobrze o zdolności modelu do predykcji produktów o najwyższym poziomie zainteresowania. Największa liczba przypadków została zakwalifikowana kolejno do klas: 2, 5, 1, 3 i 4.

Przedstawiony model cechuje względnie dobra zdolność klasyfikacji, co potwierdzają przeprowadzone testy. Analizując odległości pomiędzy przewidywanym poziomem zainteresowania, a rzeczywistym zainteresowaniem, należy zauważyć, iż model myli się zwykle w niewielkim stopniu. Przy założeniu, że zainteresowanie może być wyrażone dowolną liczbą rzeczywistą z przedziału <1, 5>, można wyznaczyć skalę pomyłki uzyskanego klasyfikatora. Nominalna średnia wartość błędu predykcji zainteresowania wynosi ±1,22. Najbardziej istotne zmienne objaśniające zawarte w modelu to *vertical_scroll*, *prod_other_time*, *page_height*, *mouse_distance*, *prod_desc_time* I *tab_active_time*.

Tabela 3. Macierz błędnych klasyfikacji modelu drzewa klasyfikacyjnego przy zainteresowaniu wyrażonym w 5-stopniowej skali nominalnej

		Przewidywane zainteresowanie					Trafność klasyfikacji
		1	2	3	4	5	
Rzeczywiste zainteresowanie	1	54	30	11	11	24	41,5%
	2	21	69	15	14	11	53,1%
	3	16	40	34	14	26	26,2%
	4	18	31	15	38	28	29,2%
	5	7	29	15	9	70	53,8%
Suma		116	199	90	86	159	

Źródło: opracowanie własne.

W przypadku modelu zbudowanego dla zainteresowania wyrażonego w skali binarnej, błąd klasyfikacji wyniósł 31,1 proc., co potwierdza jego zdolności predykcyjne. Pole powierzchni pod krzywą AUC wyniosło 0,735. Analizując macierz błędnych klasyfikacji (Tabela 4) oraz wartości współczynników czułości i specyficzności należy zauważyć, że model charakteryzuje się wysoką czułością predysponującą go do predykcji faktycznego zainteresowania. Najbardziej istotne zmienne wchodące w skład tego modelu to: *vertical_scroll*, *page_time*, *user_active_time*, *search_referral* and *tab_active_time*

Tabela 4. Ocena klasyfikacji modelu drzewa przy zainteresowaniu wyrażonym w skali dwuwartościowej

Fałszywie negatywne	Prawdziwie negatywne	Fałszywie pozytywne	Prawdziwie pozytywne	Czułość	Specyficzność
80	197	113	230	0,742	0,636

Źródło: opracowanie własne.

3.4. Modele drzew klasyfikacyjnych dla najbardziej aktywnych uczestników

W procesie budowy modeli dla danych zarejestrowanych dla najbardziej aktywnych uczestników badania uzyskano model drzewa klasyfikacyjnego, który cechuje się błędem klasyfikacji równym 58,9 proc. W modelu tym zainteresowanie wyrażono w pełnej 5-stopniowej skali. – model osiągnął zatem ponad dwukrotnie lepsze parametry zdolności predykcyjnej w stosunku do klasyfikatora losowego.

Tabela 5. Macierz błędnych klasyfikacji modelu drzewa przy zainteresowaniu wyrażonym w skali binarnej

		Przewidywane zainteresowanie					Suma	Trafność klasyfikacji
		1	2	3	4	5		
Rzeczywiste zainteresowanie	1	8	19	8	9	4	48	16,7%
	2	8	35	13	12	4	72	48,6%
	3	6	30	50	24	10	120	41,7%
	4	1	17	28	55	17	118	46,6%
	5	5	9	34	33	55	136	40,4%
Suma		28	28	110	133	133	90	

Źródło: opracowanie własne.

W przypadku tego modelu najlepszą trafność klasyfikacji uzyskano kolejno dla poziomu zainteresowania: 2, 4, 3, 5, 1. Model ten klasyfikuje największą liczbę ocenionych produktów jako interesujące w stopniu 5 i 4. Model ten cechuje również stosunkowo niska nominalna średnia wartość błędu predykcji równa ±0,91.

Model zbudowany dla grupy najbardziej aktywnych uczestników i zainteresowania wyrażonego w skali binarnej charakteryzuje się niskim błędem klasyfikacji na poziomie 17,3 proc. Pole powierzchni pod krzywą AUC wynosi 0,796. Model cechuje wysoka zdolność do przewidywania rzeczywistego zainteresowania (patrz Tabela 6).

Tabela 6. Ocena klasyfikacji modelu drzewa dla próby najbardziej aktywnych uczestników przy zaintresowaniu wyrażonym w skali binarnej

Fałszywie negatywne	Prawdziwie negatywne	Fałszywie pozytywne	Prawdziwie pozytywne	Czułość	Specyficzność
24	59	61	350	0,9359	0,492

Źródło: opracowanie własne.

4. Podsumowanie

W artykule zaprezentowano szczegółowe wyniki badań identyfikacji preferencji użytkowników serwisów handlu elektronicznego z wykorzystaniem techniki pośredniej informacji zwrotnej i metody drzew klasyfikacyjnych. Przedstawiono parametry modeli dla wielo-, jak i dwuwartościowej logiki zmiennej decyzyjnej. Zwrócono uwagę na istotne wyniki powstałe z analiz zachowań najbardziej aktywnych uczestników badania.

Przedstawione rozważania stanowią punkt wyjścia do planowanych dalszych badań nad problematyką identyfikacji preferencji użytkowników na potrzeby systemów rekomendacyjnych w platformach e-handlu.

Bibliografia

[1] Allen, R. B.: *User models: Theory, method, and practice*. International Journal of Man-Machine Studies. 1990, Volume 32, Issue 32.
[2] Avery C., Zeckhauser R.: *Recommender systems for evaluating Computer Messages*. Communications of the ACM. Marzec 1997, Volume 40 Issue 3.
[3] Claypool M., et al.: *Implicit Interest Indicators*. Proceedings of the 6th international conference on Intelligent user interfaces. ACM, Nowy Jork 2001.
[4] Demski T.: *Drzewa klasyfikacyjne w przewidywaniu migracji klientów*. Systemy IT. 2005, Nr 3(57).
[5] Fischer G.: *User Modeling in Human-Computer Interaction*. User Modeling and User-Adapted.
[6] Goecks J., Shalvik J.: *Learning users' interests by unobtrusively observing their normal behavior*. Proceedings of the 5th international conference on Intelligent user interfaces. ACM, Nowy Jork 2000.
[7] Grundin J.: *Groupware and Social Dynamics: Eight Challenges for Developers*. Communications of the ACM, Styczeń1994. Volume 37, Issue 1.
[8] Hand D.J., Henley W.E.: *Statistical classification methods in consumer credit scoring: a review*. Journal of the Royal Statistical Society. 1997, Series A, No. 160(3).
[9] Interaction (UMUAI). 2001, Volume 11, Issue 1–2.
[10] Kellar M., et. al..: *Effect of task on time spent reading as an implicit measure of interest*. American Society for Information Science and Technology 2004, Volume 41, Issue 1.
[11] Montaner M.: *A Taxonomy of Personalized Agents on the Internet*. Technical Report TRUDG. Departament d'Electrònica, Informàtica i Automàtica. Universitat de Girona, 2001.

[12] Kelly D.: *Implicit Feedback: Using Behavior to Infer Relevance*. New Directions in Cognitive Information Retrieval. The Information Retrieval Series, 2005, Volume 19, Section IV.
[13] Nielsen J.: *Personalization is over-rated*. Alertbox. Źródło: *http://www.useit.com/alertbox/981004.html* [dostęp: 2015-06-15].
[14] Próchnicka M.: *Metody i techniki modelowania użytkownika w inteligentnych systemach informacyjnych*. Multimedialne i Sieciowe Systemy Informacyjne, Wrocław 2000.
[15] Velayathan G., Yamada S.: *Behavior Based Web Page Evaluation*. Proceedings of the 15th international conference on World Wide Web. ACM. Nowy Jork, 2006.
[16] Watson A., Sasse M. A.: *Measuring perceived quality of speech and video in multimedia conferencing applications*. ACM international conference on Multimedia. ACM, Nowy Jork 1998.

CLASSIFICATION TREES IN E-COMMERCE CUSTOMER PREFERENCES IDENTIFICATION

Summary

This paper aims to present full results of an e-comerce customer preferences identification study using classification trees. The study led to building classification models based on indicators obtained using implicit feedback. The paper constitutes a starting point for further research in the field of recommender systems.

Keywords: customer preferences, e-commerce, classification trees, recommender systems

Tomasz Zdziebko
Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytet Szczeciński

Piotr Sulikowski
Wydział Informatyki
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie