

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Inżynierii Metali i Informatyki Przemysłowej



PRACA DYPLOMOWA MAGISTERSKA

pt.

*„Implementacja algorytmu personalizacji w systemie
rekomendacji produktów na przykładzie wybranego sklepu
internetowego”*

Imię i nazwisko dyplomanta:	Grzegorz Zięba
Kierunek studiów:	Informatyka Stosowana
Specjalność:	Systemy Informatyki Przemysłowej
Nr albumu:	126137
Promotor:	dr inż. Łukasz Rauch
Recenzent:	dr hab. inż. Krzysztof Banaś, prof. AGH

Podpis dyplomanta:

Podpis promotora:

Kraków 2010



Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie i że nie korzystałem ze źródeł innych niż wymienione w pracy.

Kraków, dnia

Podpis dyplomanta

*Składam serdeczne podziękowania
Panu dr inż. **Łukaszowi Rauchowi**
za pomoc i cenne uwagi podczas
realizacji tej pracy*

Spis Treści

1.	WSTĘP	7
2.	WPROWADZENIE DO SYSTEMÓW PERSONALIZACJI I REKOMENDACJI.....	9
2.1.	INTERNETOWE TRENDY W WYKORZYSTANIU PERSONALIZACJI	9
2.1.1.	<i>Serwisy społecznościowe i informacyjne</i>	<i>10</i>
2.1.2.	<i>Reklama i marketing.....</i>	<i>12</i>
2.1.3.	<i>Handel elektroniczny.....</i>	<i>13</i>
2.1.4.	<i>Usługi mobilne</i>	<i>14</i>
2.2.	KORZYŚCI PŁYNĄCE Z WYKORZYSTANIA PERSONALIZACJI	15
2.2.1.	<i>Dla odbiorców.....</i>	<i>15</i>
2.2.2.	<i>Dla właścicieli witryn.....</i>	<i>15</i>
2.3.	CZYNNIKI WPLYWAJĄCE NA SKUTECZNOŚĆ PERSONALIZACJI I REKOMENDACJI.....	16
2.3.1.	<i>Dobór odpowiedniego systemu.....</i>	<i>16</i>
2.3.2.	<i>Sposób prezentacji.....</i>	<i>17</i>
3.	PERSONALIZACJA W SYSTEMACH REKOMENDACJI	18
3.1.	DANE WEJŚCIOWE DO SYSTEMU REKOMENDACJI	19
3.1.1.	<i>Dane wejściowe od użytkownika (odbiorcy rekomendacji).....</i>	<i>19</i>
3.1.2.	<i>Dane wejściowe od społeczności użytkowników i globalne.....</i>	<i>21</i>
3.2.	IDENTYFIKACJA UŻYTKOWNIKÓW	22
3.2.1.	<i>Cookies</i>	<i>22</i>
3.2.2.	<i>Logi serwera.....</i>	<i>23</i>
3.2.3.	<i>Uwierzytelnianie.....</i>	<i>23</i>
3.2.4.	<i>Wielu użytkowników, jeden komputer.....</i>	<i>24</i>
3.3.	METODY REKOMENDACJI	24
3.3.1.	<i>Na podstawie danych demograficznych.....</i>	<i>25</i>
3.3.2.	<i>Wyszukiwanie (ang. raw retrieval).....</i>	<i>25</i>
3.3.3.	<i>Selekcja ręczna (ang. manualny selected).....</i>	<i>25</i>
3.3.4.	<i>Podsumowanie statystyczne (ang. statistical summaries)</i>	<i>26</i>
3.3.5.	<i>Korelacja atrybutów (ang. attribute-based).....</i>	<i>26</i>
3.3.6.	<i>Korelacja obiektów (ang. item-to-item correlation).....</i>	<i>27</i>
3.3.7.	<i>Korelacja użytkowników (ang. user-to-user correlation).....</i>	<i>28</i>
3.3.8.	<i>Metody mieszane</i>	<i>29</i>
3.4.	POWSZECHNIE STOSOWANE TECHNIKI	30
3.5.	DANE WYJŚCIOWE SYSTEMU	31
3.6.	STOPIEŃ PERSONALIZACJI.....	34
3.6.1.	<i>Niespersonalizowana.....</i>	<i>34</i>

3.6.2.	<i>Personalizacja ulotna</i>	35
3.6.3.	<i>Personalizacja ciągła</i>	35
3.7.	SPOSÓB DOSTARCZANIA	35
3.7.1.	<i>Z naciskiem (ang. push)</i>	36
3.7.2.	<i>Na żądanie (ang. pull)</i>	36
3.7.3.	<i>Pasywny (ang. passive)</i>	36
3.8.	PROBLEMY SYSTEMÓW REKOMENDACJI	37
3.8.1.	<i>Problem nowego użytkownika (ang. cold start problem)</i>	39
3.8.2.	<i>Problem nowego elementu</i>	39
3.8.3.	<i>Problem rzadkości danych</i>	39
3.8.4.	<i>Inne problemy</i>	40
4.	PROPONOWANE ROZWIĄZANIE	42
4.1.	ZAŁOŻENIA OGÓLNE I FUNKCJONALNE	42
4.1.1.	<i>Dowolność przyznawania ocen</i>	43
4.1.2.	<i>Odrzucanie skrajnych wartości</i>	43
4.1.3.	<i>Personalizacja ciągła</i>	43
4.1.4.	<i>Rekomendacje dla użytkowników oraz elementów</i>	44
4.1.5.	<i>Dowolność prezentacji rekomendacji</i>	44
4.2.	ALGORYTM SLOPE ONE.....	44
4.2.1.	<i>Zalety algorytmu</i>	45
4.2.2.	<i>Działanie algorytmu</i>	46
4.2.3.	<i>Przykład obliczeń</i>	48
4.2.4.	<i>Złożoność algorytmu</i>	48
4.2.5.	<i>Testy empiryczne algorytmu</i>	49
4.3.	STRUKTURA BAZY DANYCH.....	51
4.4.	IDENTYFIKACJA UŻYTKOWNIKÓW	53
4.5.	DZIAŁANIE PROPONOWANEGO SYSTEMU REKOMENDACJI.....	54
4.5.1.	<i>Obliczenia i aktualizacje danych</i>	54
4.5.2.	<i>Usuwanie anomalii</i>	54
5.	WDROŻENIE ROZWIĄZANIA W SKLEPIE INTERNETOWYM.....	56
5.1.	DANE POCZĄTKOWE I DOBÓR OCEN	57
5.2.	STOSOWANE METODY REKOMENDACJI	58
5.3.	METODYKA TESTÓW.....	59
5.4.	UZYSKANE WYNIKI.....	60
5.4.1.	<i>Przygotowanie danych do analizy</i>	60
5.4.2.	<i>Skuteczność systemów</i>	60
5.4.3.	<i>Skuteczność w poszczególnych dniach tygodnia</i>	62

5.4.4.	<i>Liczba produktów w koszyku w poszczególnych systemach.....</i>	<i>63</i>
5.4.5.	<i>Podsumowanie wyników.....</i>	<i>64</i>
6.	PODSUMOWANIE.....	66
7.	BIBLIOGRAFIA	68

1. Wstęp

W dobie wszechobecnego Internetu, który ciągle dynamicznie się rozrasta. Kiedy to pojedynczy człowiek potrzebowałby wielu tysięcy lat na odwiedzenie wszystkich stron WWW. Niezbędne wydaje się stosowanie systemów informatycznych, które pomagają w dostosowaniu treści do zainteresowań odbiorcy. Takie systemy są powszechne i obecne praktycznie na każdym kroku. Pomagają wyodrębnić te informacje, których potencjalnie oczekuje użytkownik. Zmniejszając tym samym jego dezorientację. Wyszukiwarki internetowe codziennie dostosowują setki milionów wyników do kontekstu zapytania (chwilowego zainteresowania odbiorcy). Filtry wykorzystywane w poczcie elektronicznej, automatycznie decydują o tym jakie informacje są pożądane przez odbiorcę, a jakie należy usunąć. W przypadku poszukiwania produktu lub usługi wykorzystywane są porównywarki oraz serwisy opiniotwórcze. Sklepy internetowe proponują produkty, którymi użytkownik jest szczerze zainteresowany. Aplikacje internetowe dostosowują swój wygląd i funkcjonalność do preferencji użytkownika, a urządzenia mobilne pomagają znaleźć interesujące miejsce w pobliżu aktualnej lokalizacji. Wszystkie te mechanizmy wykorzystują systemy personalizacji, a szczególnym ich przypadkiem są systemy rekomendacji.

Najprostszym sposobem, aby otrzymać rekomendacje produktu, filmu lub interesującej strony WWW jest zapytanie o opinie znajomych. Z czasem, na podstawie obserwacji, można stwierdzić, który ze znajomych ma lepszy gust od innych. Jeśli pojawia się zbyt dużo dostępnych możliwości, mniej praktyczne staje się decydowanie o tym czego się chce na podstawie opinii stosunkowo małej grupy ludzi (znajomych), która prawdopodobnie nie jest nawet świadoma wszystkich dostępnych możliwości. Dlatego właśnie stosowane są systemy rekomendacji. Najczęściej jest to zbiór technik nazwany *filtrowaniem kolaboratywnym*. Techniki te „odpowiadają na pytanie” jak wykorzystać preferencje grupy użytkowników do zarekomendowania określonych elementów innym ludziom. Takie techniki wykorzystują dane o dużych grupach ludzi, w celu odnalezienia mniejszych, których preferencje bardziej pasują do użytkownika. Przypomina to łączenie opinii wielu ludzi na temat rzeczy, które lubią lub nie i budowanie na podstawie tych informacji rankingu sugestii. Istnieje wiele różnych dróg do zbadania podobieństw i preferencji danego użytkownika. W tej pracy przedstawiony został przegląd różnych technik oraz zademonstrowane zostało rozwiązanie, które wykorzystuje kilka z nich. [1]

Cel pracy

Celem niniejszej pracy jest zaimplementowanie skutecznego algorytmu personalizacji w systemie rekomendacji produktów, który przyczyni się do zwiększenia przychodów wybranego sklepu internetowego. Dobrze działający system rekomendacji produktów powinien przyczynić się do zwiększenia średniej wartości zamówienia, uzyskiwać dobre rekomendacje dostosowane do preferencji użytkownika i być wydajny technicznie.

W ramach realizacji tak postawionego celu wyznaczono następujące zadania:

- 1) Przeszukanie istniejących zasobów pod kątem uzasadnienia stosowania rozwiązań personalizacji i aktualnych oraz przyszłych trendów.
- 2) Analiza dostępnych rozwiązań personalizacji w systemach rekomendacji.
- 3) Wykonanie projektu systemu personalizacji w systemie rekomendacji.
- 4) Implementacja zaprojektowanego systemu.
- 5) Wykonanie testów na przykładzie wybranego sklepu internetowego w warunkach rzeczywistej sprzedaży.
- 6) Analiza uzyskanych wyników.

Rozdział drugi niniejszej pracy jest wprowadzeniem do systemów personalizacji i rekomendacji. Uzasadnia sensowność stosowania systemów personalizacji oraz pokazuje trendy występujące na rynku globalnym. W rozdziale opisane zostały korzyści ze stosowania tego typu systemów oraz dwa kluczowe czynniki mające wpływ na jego skuteczność.

W rozdziale trzecim przedstawione zostały dostępne rozwiązania stosowane w systemach rekomendacji. Oprócz stosowanych metod i technik scharakteryzowane zostały możliwe dane wejściowe i wyjściowe, sposoby identyfikacji użytkowników, stopień personalizacji, sposoby dostarczania rekomendacji oraz informacje na temat możliwych problemów z jakimi zmagają się twórcy tego typu systemów.

Rozdział czwarty to opis proponowanego rozwiązania. Zostały w nim przedstawione założenia, algorytm, struktura bazy danych oraz zaimplementowane podsystemy. Na prostym przykładzie zademonstrowany został sposób dokonywania obliczeń rekomendacji.

W kolejnym rozdziale przedstawione zostało wdrożenie rozwiązania w wybranym sklepie internetowym. Pokróctce scharakteryzowany został sklep, jego działalność oraz dotychczas stosowane systemy rekomendacji. W dalszej części znaleźć można opis metodyki przeprowadzonych testów oraz analizę uzyskanych wyników.

Ostatni rozdział podsumowuje otrzymane rezultaty, a także zawiera sugestie na temat możliwości dalszej optymalizacji sposobu wykorzystania proponowanego rozwiązania oraz dalszego rozwinięcia niniejszej pracy.

2. Wprowadzenie do systemów personalizacji i rekomendacji

W dobie dynamicznego rozwoju Internetu, użytkownicy zostają przytłoczeni nadmiarem informacji. Portale, serwisy, blogi emitują potężne ilości nowych wiadomości. Do odwiedzin stron przez jak najszerszą publiczność zachęcają coraz to nowe kampanie reklamowe. Należy jednak pamiętać, że potencjał odbiorcy poszukującego i weryfikującego przydatność rozpowszechnianych treści jest ograniczony. Koniecznością staje się takie kreowanie działań, aby odpowiadały indywidualnym preferencjom użytkowników. Z pomocą przychodzi personalizacja.

Systemy personalizacji mają na celu osiągnięcie najwyższego poziomu satysfakcji u odbiorcy poprzez dążenie do zaspokojenia jego indywidualnych potrzeb. Dają możliwość dostarczania odbiorcy właściwych treści, produktów i usług, którymi potencjalnie będzie bardziej zainteresowany niż innymi. Personalizacja sprawia, że odbiorca w krótkim czasie pozyskuje dostęp do właściwej informacji lub produktu.

Specyficznym typem personalizacji są systemy rekomendacji. Bazują one na idei sztucznej inteligencji, socjologii, technologii WWW. Ich działanie sprowadza się do zastąpienia automatycznymi (wyliczonymi) rekomendacjami, naturalnego sposobu rekomendowania, opartego na opiniach przyjaciół, znajomych, publikacjach i innych źródłach pomocnych w dokonywaniu wyborów. W tym celu wykorzystują zgromadzone dane na temat użytkowników i ich preferencji.¹

2.1. Internetowe trendy w wykorzystaniu personalizacji

Coraz więcej wydawców serwisów internetowych dostrzega korzyści płynące z zastosowania systemów personalizacji. Największe firmy starają się pozyskiwać jak najwięcej informacji o użytkownikach z różnych, również rozproszonych miejsc w Internecie. Tak zebrane informacje wykorzystywane są pośrednio lub bezpośrednio w celu zwiększenia skuteczności różnych przekazów marketingowych.

¹ <http://www.e-marketing.pl/artyk/artyk53.php>, Adamczyk J., Zbawienna personalizacja, 07.2010

2.1.1. Serwisy społecznościowe i informacyjne

Personalizacja pomaga w znalezieniu informacji zgodnych z predyspozycjami danego odbiorcy. Portale informacyjne na podstawie zebranych danych o użytkownikach mogą prezentować treści, które z większym prawdopodobieństwem zainteresują odbiorcę. Popularne są także możliwości personalizacji interfejsów użytkownika, dostosowując wygląd strony internetowej lub aplikacji do indywidualnych preferencji.

Serwis *Facebook.com* mający ponad 500 milionów użytkowników na całym świecie dysponuje jednym z najbardziej zaawansowanych systemów rekomendacji. Przyczynia się do tego popularność, charakter społecznościowy oraz udostępnienie wielu darmowych wtyczek ułatwiających integrację serwisu z innymi stronami internetowymi. Wychodząc z założenia, że każdego użytkownika da się zidentyfikować poprzez charakter jego działalności w sieci (kontakty, ulubione strony, filmy, obrazki, spotkania w których uczestniczy, produkty które poleca), firma wprowadziła w 2010 roku system *Open Graph* potrafiący wyszukiwać podobieństwa pomiędzy obiektami, prezentując je w formie grafu. Łącząc możliwości tej technologii rekomendacji z innymi oferowanymi przez *Facebook* otrzymujemy potężne narzędzie. Używając np. *Facebook Connect* użytkownik może zalogować się do innego serwisu, a wtedy serwis zewnętrzny powiąże jego działalność w *Facebooku* z własnymi funkcjonalnościami polecając mu np. określone restauracje z chińskim jedzeniem, które tak bardzo lubi. Pokaże również, co lubią i robią jego znajomi. Inną rewolucyjną technologią udostępnioną w postaci wtyczki jest przycisk „*I like it!*” (przedstawiony na rysunku 1), pozwalający polecać odnośniki do treści, produktu w serwisie społecznościowym. Oprócz pokazania informacji co lubią znajomi, system posiada możliwość rekomendacji rzeczy pasujących do zainteresowań np. na podstawie podobieństw między użytkownikami. Większość możliwości można ze sobą integrować tworząc zaawansowane narzędzia rekomendacji ².

² <http://blog.mitx.org/Blog/bid/45106/Facebook-Open-Graph-Personalizing-the-Social-Web>, Gena Folts, Facebook Open Graph - Personalizing the Social Web, 07.2010



ASORTYMENT
Woda koloniska
Woda perfumowana
Woda po goleniu
Woda toaletowa
Pobierz liste producentów
BESTSELLERY
Produkt1
Produkt2
Produkt3

INFORMACJE O PRODUKCIE
EDIMAX (EW-7209APG) Access Point 54Mbps 802.11g, 5xLAN, WDS, APC
producent: 2S Media kod produktu: STEEDIBAK0008

Krótko o: Kod Producenta: EW-7209APG
Moc wyjściowa radii: 18 dBm
Ilość gniazd antenowych: 1 szt.
Typ gniazd antenowych: RSM
Antena: Odłączana
Ilość gniazd kablowych RJ45: 5 szt.
Standard radiowy: 54 Mbps - 802.11g
Pasmo: 2,4 GHz
Zarządzanie: Przejrzysta www

Cena: 1146.55Zł / szt.

Dostępność:

Wysokość: **w ciągu 24h**
Koszt dostawy: **10.00Zł**

Kod Producenta: EW-7209APG

LOGOWANIE
nazwa użytkownika
twoje hasło
Zaloguj się

POMOC I DANE KONTAKTOWE
Nazwa
ul. Wolska Północna
05-500 Warszawa
e-mail: **napisz do nas**
tel. **987 987 877 - test**
fax **test**
00 test
00 test
skype **test**
test

INFORMACJE
Przesyłki
Koszty przesyłki
Jak jest dostawca czasu dostawy?

Rys 1. Zrzut ekranu prezentujący wykorzystanie przycisku „I like it!”. Źródło: <http://www.iai-shop.com>, dostawca oprogramowania dla sklepów internetowych, 07.2010

Wyszukiwarka *Google* od wielu lat zbiera informacje o każdym swoim użytkowniku. Największe znaczenie dla firmy ma wyszukiwanie spersonalizowane dostarczające olbrzymich ilości danych, również dla celów poprawy algorytmów wyszukiwania niespersonalizowanego. Dane dostarczane *Google* przez użytkownika mogą posłużyć do modyfikacji wyników wyszukiwania tak, aby najbardziej odpowiadały zainteresowaniom tego użytkownika, m.in.:

- strony często wybierane przez użytkownika mogą otrzymywać wyższe pozycje w jego spersonalizowanych wynikach wyszukiwania,
- częstotliwość wybierania tych często odwiedzanych stron może wpływać na ranking preferowanych witryn danego użytkownika,
- witryny pomijane przez użytkownika w wynikach wyszukiwania mogą być usuwane z jego indywidualnych wyników wyszukiwania,
- podstrony z danej witryny mogą pojawiać się wyżej w personalnych wynikach wyszukiwania, jeżeli użytkownik poprzednio odwiedzał inne podstrony z tych witryn lub z witryn powiązanych z tą witryną,

- do wyników wyszukiwania użytkownika mogą być dodane określone wyniki, jeżeli były one wybierane w powiązanych wyszukiwaniach użytkownika.³

2.1.2. Reklama i marketing

Obok przedstawionych w pracy możliwości wykorzystania systemów personalizacji nie sposób nie wspomnieć o targetowaniu behawioralnym oraz e-mail marketingu.

Targetowanie behawioralne wykorzystywane jest obecnie przede wszystkim w systemach adserwerowych służących do emisji i zarządzania internetowymi kampaniami reklamowymi. O jego popularności decyduje skuteczność, która jest wyższa nawet o kilkadziesiąt procent względem tradycyjnego sposobu serwowania reklam banerowych. Targetowanie behawioralne w reklamie polega na emisji reklam dobranych zależnie od profilu behawioralnego odbiorcy. Taki profil tworzony jest w oparciu o odwiedzane strony, dane społeczno-demograficzne, zainteresowania produktowe i styl życia internauty. Informacje te pozwalają na dopasowanie reklam do zainteresowań poszczególnych użytkowników, co sprawia, że stają się one bardziej przyjazne. Istota działania takiego systemu polega na śledzeniu ruchów użytkowników na różnych stronach internetowych. Następnie na podstawie zebranych danych użytkownicy przypisywani są, zazwyczaj, do kilkunastu modeli zachowań. Próbie użytkowników z danego modelu wyświetlana jest dodatkowa ankieta, która pomaga w określeniu danego profilu dla całej grupy modelowej. W ten sposób powstają profile o charakterystyce związanej z zainteresowaniami np.: automaniacy, podróżnicy, biznesmeni. Drugą część systemu stanowi sposób emisji reklamy. Po ustaleniu profilu danego odbiorcy, otrzymuje on tematycznie dobrane do profilu reklamy niezależnie od tematyki strony, na której w danej chwili przebywa. Zwiększa to czas ekspozycji odbiorcy na reklamę, która jest dostosowana do jego zainteresowań, co w konsekwencji pozytywnie wpływa na skuteczność⁴.

W marketingu wykorzystuje się również pocztę elektroniczną do kontaktu z klientami. Niektóre firmy wykorzystują tą formę do posprzedażowej rekomendacji produktów. Klient po danej akcji pozwalającej na identyfikację jego adresu e-mail (np. zakup produktu), może otrzymać od sklepu e-mail z rekomendacjami produktów, których jeszcze nie kupił, najczęściej w ograniczonej czasowo promocji. W wielu przypadkach prowadzi to do zwiększenia wartości koszyka zamówień. Należy również pamiętać o personalizacji

³ <http://www.ittechnology.us/przyszosc-google-personalizacja/>, Gosiewski M., Przyszłość Google – personalizacja, 07.2010

⁴ <http://behavia.pl>, dostawca systemów targetowania behawioralnego, 07.2010

stosowanej w tytule lub treści wiadomości e-mail. Zwroty bezpośrednie lub użycie imienia odbiorcy często, a w szczególności w określonych branżach i przy budowaniu dłuższych relacji, poprawia skuteczność działań e-mail marketingowych ⁵.

2.1.3. Handel elektroniczny

Personalizacja w handlu elektronicznym coraz częściej jest traktowana jako klucz do zwiększenia sprzedaży. Rekomendacje produktów mają wpływ na wielkość zamówienia. Sklepy internetowe decydują się na wykorzystywanie mechanizmów zbierających i analizujących dane na temat użytkowników. Rekomendacje udostępniane są na podstawie zachowań w sieci (targetowanie behawioralne), dotychczasowych zakupów w sklepie, ocen wystawianych przez użytkowników, jak i, w przypadku przede wszystkim dużych ilości danych, informacji na temat zachowań poszczególnych grup. W Stanach Zjednoczonych ponad połowa sprzedawców online oferuje rekomendacje produktów odpowiadające profilowi klienta. Co czwarty deklaruje wprowadzenie takich mechanizmów w najbliższym czasie. W Polsce co szósty sklep internetowy stosuje rekomendacje w oparciu o profil klienta, a co czwarty stosuje indywidualne rekomendacje produktów w oparciu o dotychczasowe zakupy klientów. Co szósty zamierza wprowadzić taką możliwość w najbliższym czasie ⁶.

Istotę tego typu systemów podkreśla fakt, że coraz więcej firm decyduje się nie tylko na wprowadzenie systemów personalizacji, ale również na optymalizację już istniejących. Przykładem takiego działania może być firma *Netflix* prowadząca w Stanach Zjednoczonych usługi wypożyczania filmów na nośnikach fizycznych. Klient zamawia wybrany film poprzez Internet. *Netflix* w systemach rekomendacji wykorzystuje oceny filmów wystawione przez klienta. W 2006 roku firma wyznaczyła nagrodę w wysokości 1 miliona dolarów za poprawę wyniku działania ich algorytmu rekomendacji filmów. Udostępnione zostało ponad 100 milionów ocen wystawionych przez ponad 480 tysięcy użytkowników dla prawie 18 tysięcy filmów. Zadanie polegało na empirycznej poprawie o 10% działania algorytmu. Wyzwanie to zaangażowało do pracy zespół z całego świata, a mimo to cel został osiągnięty dopiero w roku 2009 [2].

⁵ <http://interaktywnie.com/biznes/newsy/e-mail-marketing/e-mail-przechodzi-z-odbiorca-na-ty-11582>, Krawiec A., E-mail przechodzi z odbiorcą na „Ty”, 07.2010

⁶ http://www.praktycy.com/infoteka/Raport_E-commerce_w_Polsce_2009_Praktycy_com.pdf, Raport E-commerce w Polsce 2009, 07.2010

2.1.4. Usługi mobilne

Dostęp do Internetu na urządzeniach mobilnych (głównie telefony komórkowe, PDA, netbooki) nabiera coraz większego znaczenia. Przyczynia się do tego rozwój technologii, szybsze łącza bezprzewodowe i bardziej użyteczne urządzenia mobilne (większe ekrany o dużych rozdzielczościach, szybkie i użyteczne interfejsy użytkownika, nanotechnologia). Na świecie ponad 650 milionów ludzi (około 13,4% abonentów sieci komórkowych) korzysta przynajmniej raz w miesiącu z Internetu, za pomocą urządzeń mobilnych. Przewiduje się dalszy, dynamiczny wzrost liczby użytkowników (około 1 miliard użytkowników w 2013 roku). Natomiast w 2020 roku urządzenia mobilne prawdopodobnie będą podstawowym narzędziem dostępu do Internetu dla większości ludzi na świecie⁷. Użytkownicy Internetu na urządzeniach mobilnych poszukują dostępu do serwisów informacyjnych, wyszukiwarek internetowych, multimedii oraz poczty elektronicznej i komunikatorów. Internet mobilny ma kilka przewag nad klasycznym. Połowa ludzi na świecie ma możliwość korzystania z Internetu za pomocą urządzeń mobilnych. Zasięg jest około 2 razy większy niż Internetu klasycznego. Urządzenia takie są mniejsze i można je mieć zawsze przy sobie, gdziekolwiek znajduje się użytkownik. Obecne technologie i oprogramowanie zainstalowane w takim urządzeniu nie wymaga specjalnie przygotowanych serwisów internetowych dla tych urządzeń. Jednak ze względu na optymalizację transferu danych często przygotowuje się „lżejsze” wersje z przeznaczeniem na urządzenia mobilne. Sieci komórkowe i wbudowane urządzenia nawigacyjne GPS umożliwiają precyzyjną lokalizację użytkownika, co prowadzi do powstawania nowych usług z wykorzystaniem geolokalizacji, np. rekomendacje najbliższych restauracji z kuchnią polską. 2/3 obecnych użytkowników Internetu mobilnego wyraża zainteresowanie usługami spersonalizowanymi, oczekując informacji zgodnych z ich preferencjami, miejscem pobytu, porą dnia oraz kontekstem społecznym. Odnotowuje się również wzrost znaczenia usług m-commerce. Coraz więcej sklepów internetowych dostrzega możliwości urządzeń mobilnych z dostępem do Internetu^{8 9 10}.

⁷ http://www.pewinternet.org/~media/Files/Reports/2008/PIP_FutureInternet3.pdf.pdf, Quitney Anderson J., The Future of the Internet III, 07.2010

⁸ <http://interaktywnie.com/biznes/newsy/raporty-i-badania/uzytkownicy-telefonow-komorkowych-chca-spersonalizowanych-uslug-11068>, Matuszewska B., Użytkownicy telefonów komórkowych chcą spersonalizowanych usług, 07.2010

⁹ http://pl.wikipedia.org/wiki/Mobilny_Internet, informacje o mobilnym Internecie, 07.2010

¹⁰ <http://www.idg.pl/news/353705/IDC.w.roku.2013.miliard.urzadzen.mobilnych.bedzie.podlaczonych.do.internetu.html>, Muszyński J., podsumowanie raportu IDC, 07.2010

2.2. Korzyści płynące z wykorzystania personalizacji

Stosowanie systemów personalizacji niesie ze sobą wiele korzyści zarówno dla odbiorców, jak również dla właścicieli stron internetowych, na których są stosowane. Niektóre z nich zostały wymienione niżej.¹¹

2.2.1. Dla odbiorców

Wśród korzyści, jakie czerpią odbiorcy witryn internetowych wykorzystujących personalizację, można wymienić:

- możliwość kompozycji indywidualnej strony internetowej z dostępnych treści,
- szybki dostęp do interesujących informacji,
- sprawny dostęp do interesującej oferty sklepu internetowego,
- zabezpieczenie przed znużeniem, poprzez dynamizm zawartości,
- możliwość prenumeraty biuletynu zgodnego z preferencjami i oczekiwaniami odbiorcy (w przypadku portalu np. informacje ze świata biznesu),
- możliwość składania zamówień na nietypowe produkty (np. własna kompozycja butów lub zamawianie indywidualnego druku książki),
- atrakcyjność i zadowolenie poprzez prezentacje treści dostosowanej do preferencji.

2.2.2. Dla właścicieli witryn

Wśród korzyści, jakie czerpią wydawcy witryn internetowych wykorzystujących systemy personalizacji, można wymienić:

- pozyskanie uwagi i zdobycie uznania wśród społeczności internetowej za unikalność świadczonych usług,
- zdobycie zadowolenia i uznania w oczach klienta, które może być źródłem najefektywniejszej reklamy rozpowszechnianej drogą rekomendacji,
- możliwość zdefiniowania grupy docelowej do której można kierować indywidualne propozycje pod kątem zainteresowań grupy,
- zainicjowanie i podtrzymanie sprawnego procesu komunikacyjnego z określonymi użytkownikami (w przypadku użytkowników zarejestrowanych możemy skorzystać w procesie komunikacji z danych personalno - demograficznych),
- możliwość udostępniania płaszczyzny reklamowej zdefiniowanej poprzez profil klienta lub grupy,

¹¹ <http://www.e-marketing.pl/artyk/artyk53.php>, Adamczyk J., Zbawienna personalizacja, 07.2010

- szansa na pozyskanie lojalności e-klienta,
- generowanie ponownych wizyt odbiorców,
- dłuższy czas wizyt na stronie,
- zdobycie pozytywnych opinii na temat strony lub firmy,
- konkurencyjność wobec firm nie stosujących personalizacji.

2.3. Czynniki wpływające na skuteczność personalizacji i rekomendacji

Strony internetowe wykorzystujące systemy personalizacji mają na celu zwiększenie czasu wizyty użytkownika poprzez prezentację interesujących go materiałów, a w dłuższej perspektywie zdobycie jego lojalności. Głównym celem sklepów internetowych jest zwiększenie sprzedaży, poprzez zwiększenie zainteresowania ofertą i zwiększenie wartości zamówień. Należy przy tym pamiętać o dwóch bardzo istotnych czynnikach. Wybór nawet najlepszego systemu rekomendacji nie przyniesie bowiem pożądanego efektu, jeśli zapomni się o użytecznej prezentacji wyników systemu.

2.3.1. Dobór odpowiedniego systemu

Na dobór odpowiedniej metody wykorzystywanej w procesie personalizacji lub rekomendacji wpływ mają:

- cel działania systemu,
- branża,
- sezonowość rynku,
- grupa docelowa,
- ilość rekomendowanych elementów,
- ilość użytkowników,
- dane wykorzystywane przez system oraz sposób ich gromadzenia,
- przewidywane miejsce i sposób prezentacji.

Ustalenie powyższych czynników jest kluczowe dla wyboru odpowiedniej metody i technik, które zostaną wykorzystane w procesie personalizacji. Właściwy dobór algorytmu powinna cechować wysoka jakość rekomendowanych stron lub produktów i ogólne zadowolenie użytkowników. System powinien wykonywać obliczenia najszybciej jak to możliwe, najlepiej w czasie rzeczywistym. Z uwagi na bardzo duże ilości danych, na niektórych stronach, nie zawsze jest to możliwe. Praktycznie każda strona internetowa wymaga indywidualnego podejścia, chociażby poprzez dostosowanie odpowiednich współczynników w systemie

personalizacji. Często, aby zwiększyć prawdopodobieństwo sukcesu wykorzystuje się kilka metod, które składają się na systemy nazywane hybrydowymi.

2.3.2. Sposób prezentacji

Oprócz samego wyboru miejsca prezentacji spersonalizowanej treści czy rekomendacji, należy pamiętać o całości strony internetowej, jej użyteczności i dostępności informacji. Istotny jest projekt strony, jak również szybkość jej wczytywania. Wielu klientów rezygnuje z przeglądania zawartości z powodu niejasnego przekazu marketingowego (np. brak sprecyzowania pożądanej akcji), powolnego wczytywania zawartości czy konfuzji spowodowanej źle zaprojektowanym systemem nawigacyjnym. Należy przy tym zaznaczyć, że możliwych błędów w użyteczności strony internetowej, przez które wiele serwisów codziennie traci klientów, jest znacznie więcej. Każda strona internetowa powinna być dobrze zaprojektowana, przemyślana pod kątem grupy docelowej tak, żeby ułatwić użytkownikom przejście przez pożądany przez nich proces (przeglądanie, kupowanie).

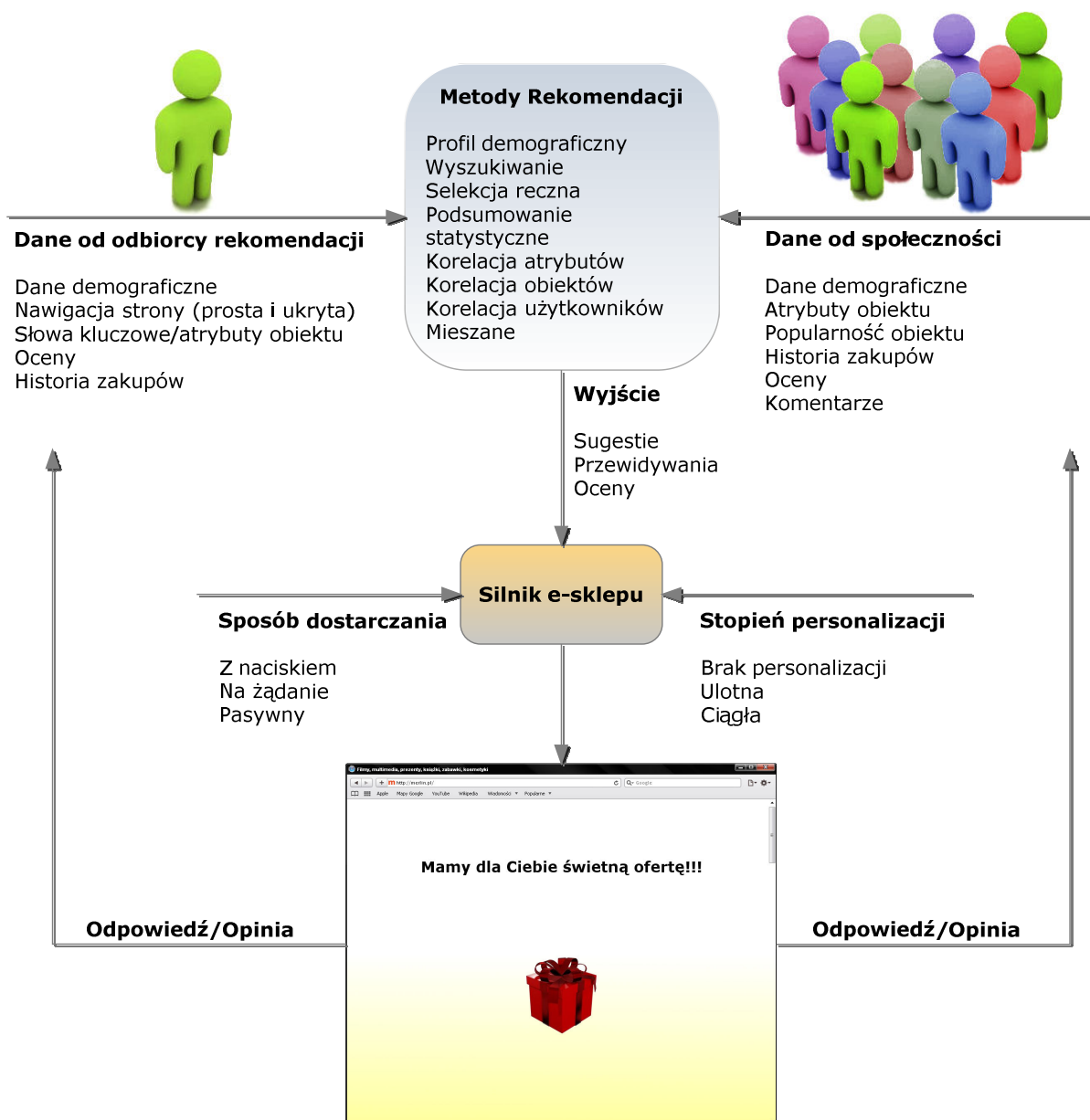
Miejsce prezentacji rekomendacji w serwisach również powinno być przemyślane. Naturalnym położeniem linków rekomendujących artykuły czy produkty wydaje się miejsce pod treścią artykułu bądź opisu produktu. Wyświetlone w ten sposób rekomendacje nie powinny zmniejszać zainteresowania przeglądającą stroną. Odnośniki źle oznaczone, słabo widoczne będą prawdopodobnie rzadziej wybierane niż te które są dobrze widoczne i właściwie zatytułowane (np. „klienci kupujący ten produkt wybrali również”). Mimo ogólnych zaleceń stosowania dobrych praktyk użyteczności, należy pamiętać, że każda strona internetowa jest inna i ma inny charakter. Coś, co świetnie sprawdza się w jednym serwisie, niekoniecznie będzie dobre w innym. Dlatego należy zachować umiar i zdrowy rozsądek w prezentacji spersonalizowanych treści. Tu z pomocą przychodzą wyspecjalizowane firmy oraz przeróżne testy funkcjonalności i użyteczności stron internetowych.

Przykładem może być firma *Amazon.com* (jeden z największych sklepów internetowych na świecie), która w fazie akceptacji koszyka, odłącza wszystkie elementy niezwiązane bezpośrednio z zakupem. Rekomendacje, a nawet menu nawigacyjne. Takie podejście maksymalizuje liczbę zrealizowanych transakcji, ograniczając liczbę możliwych wyborów. Zbyt wiele możliwości prowadzi bowiem do paraliżu decyzyjnego, co w konsekwencji może spowodować rezygnację klienta. Tak restrykcyjne podejście nie zawsze jest jednak konieczne.¹²

¹² http://interaktywnie.com/index/index/?file=uzytecznosc_w_internecie.pdf, Kuśmierk W. Skuteczność w e-commerce, Użyteczność w Internecie, 07.2010

3. Personalizacja w systemach rekomendacji

System rekomendacji wykorzystuje w swoim działaniu szereg różnych elementów i danych, pochodzących z najróżniejszych źródeł. W tym rozdziale zostaną omówione dane wejściowe do systemu, metody ich analizy oraz dane wyjściowe, sposoby ich prezentacji i problemy jakie należy mieć na uwadze przy projektowaniu takiego systemu. Ogólny schemat działania systemu rekomendacji został zaprezentowany na rysunku 2.



Rys. 2 Działanie systemu rekomendacji. Źródło: opracowanie własne na podstawie [3]

3.1. Dane wejściowe do systemu rekomendacji

Każdy system rekomendacji gromadzi dane z różnych miejsc, takich jak preferencje użytkownika, atrybuty, oceny, komentarze, korelacje. Ponieważ często dane te zajmują dużą przestrzeń można wyróżnić w nich kilka źródeł. Podstawowy podział dzieli dane na informacje o użytkowniku (demografia, wiedza, zainteresowania) oraz informacje o użytkowaniu (zachowania, akcje).

Na dane o użytkowniku składają się:

- dane demograficzne,
- wiedza użytkownika,
- zdolności i umiejętności,
- zainteresowania i preferencje,
- zamierzenia i cele użytkownika.

Na dane o użytkowaniu składają się:

- akcje wyborów,
- czasowe zachowanie podczas przeglądania,
- ocenianie i komentowanie,
- zakupy i akcje związane z zakupami,
- inne akcje potwierdzające zainteresowania.

Do ww. danych można dodać regularność użycia i częstotliwość akcji (specyficzne w odniesieniu do branży), które wiele systemów uwzględnia w obliczeniach rekomendacji. Pomocne mogą okazać się również informacje o środowisku sprzętowym (środowisku działania np. czy jest to komputer czy telefon komórkowy) i oprogramowaniu użytkownika. Wśród danych wejściowych możemy wyróżnić informacje od odbiorcy rekomendacji, na które składają się informacje o użytkowniku jak i o jego akcjach oraz zgromadzone dane od społeczności użytkowników i dane zewnętrzne. Dane te pomagają określić podobieństwo zainteresowań odbiorcy względem danej grupy użytkowników, w ramach całej populacji użytkowników. [5]

3.1.1. Dane wejściowe od użytkownika (odbiorcy rekomendacji)

Wszystkie systemy uwzględniające dane od użytkownika, któremu prezentują rekomendacje są systemami spersonalizowanymi. Systemy, które nie uwzględniają takich danych mogą prezentować jedynie niepersonalne rekomendacje (korelacje elementów, podsumowanie statystyczne). Uwzględnienie jednego lub więcej wejść danych pozwala systemowi bazować na aktualnej aktywności użytkownika, długoterminowych preferencjach lub na obydwu

źródłach. Wiele systemów rekomendacji zaczyna reagować na obecny stan użytkownika i nie korzysta wyłącznie z danych globalnych czy demograficznych. Istotną informacją jest np. aktualnie używana ścieżka nawigacji klienta, pokazująca jego aktualne zachowania. Wykorzystanie nawigacji może odbywać się w dwojaki sposób. Użytkownik może nie mieć świadomości wykorzystania ścieżki, aktualnie przeglądanej strony produktu czy koszyka zakupu, w procesie rekomendacji. Do rekomendacji może być wykorzystana np. treść zawarta na danej stronie, polecając klientowi zestaw podobnych przedmiotów do obecnie przeglądanych. Innym sposobem jest polecanie skończonego zbioru możliwości, z których klient świadomie dokonuje wyboru poprzez określenie atrybutów jakie powinny zawierać linki lub strony docelowe. W ten sposób można zaproponować klientowi linki do produktów spełniających podane przez niego kryteria, co ma charakter świadomego wyboru w procesie nawigacji. Przykładem takiego zachowania mogą być systemy ekspertowe, które na podstawie świadomego wyboru użytkownika, ostatecznie rekomendują produkty lub usługi. W niektórych przypadkach dane wejściowe od klienta nie mogą ograniczać się do jednej kategorii. Systemy wykorzystują słowa kluczowe wpisywane przez użytkownika w wyszukiwarce lub zaczerpnięte z samego opisu produktu (np. jako atrybuty produktu), który aktualnie jest przeglądany, traktując je jako pomoc w ustaleniu aktualnych potrzeb i zaspokojeniu ich poprzez odpowiednio dobrane rekomendacje.

Odbiorca rekomendacji może zapewnić najbardziej precyzyjne i pomocne dane poprzez samodzielne ocenianie i komentowanie przedmiotów, których jest użytkownikiem lub które kupił. Może to polegać na zaprezentowaniu klientowi kilku wybranych przedmiotów z prośbą o ich ocenę lub prośbą o ocenę zakupionego przedmiotu. W celu oceny najczęściej wykorzystuje się skale od 1 do 5 lub proste oceny binarne („Czy lubisz ten produkt?”). Tego typu informacje wykorzystywane są w systemach kolaboratywnych (korelacja elementów, użytkowników). Im więcej przedmiotów oceni, w sposób rzetelny, dany użytkownik, tym lepsze rekomendacje otrzyma w przyszłości.

Inne systemy zamiast pytać użytkownika o oceny, wykorzystują ukryte postacie ocen na podstawie historii zakupów. Zawartość zamówienia przedstawia bowiem listy przedmiotów, którymi klient wyraził bardzo konkretne zainteresowanie. Zakupiony przedmiot może być automatycznie oceniony przez system jako przedmiot lubiany przez użytkownika (ocena binarna). Nie ma to jednak przełożenia na zadowolenie klienta z zakupionego przedmiotu, np. po zakupie klient może być zawiedziony jakością produktu. Dlatego lepszym sposobem jest dodatkowe wykorzystanie ręcznie podanej przez użytkownika oceny. Niektóre systemy pozwalają nawet na wpisanie oceny negatywnej. Tym sposobem można uwzględnić, czy

zakupiony produkt spodobał się klientowi, czy klient jest nim bardzo zawiedziony, czy ma nastawienie neutralne. [3]

3.1.2. Dane wejściowe od społeczności użytkowników i globalne

Dane te obejmują szeroki zakres informacji dotyczących sposobu postrzegania przedmiotów przez grupy użytkowników lub społeczność jako całości. Może to obejmować przypisanie atrybutów do danej pozycji na podstawie ogólnych opinii. Przykładem jest grupowanie filmów co do gatunku lub książek według kategorii, które to jest konkluzją opinii całego społeczeństwa. Podobnie na popularność może wpływać element zewnętrzny, taki jak światowa lista bestsellerów czy globalna sprzedaż biletów na koncerty. Takie dane mogą być zaczerpnięte z zewnętrznych rankingów i zestawień statystycznych lub od partnerów biznesowych. Podsumowanie statystyczne najczęściej kupowanych produktów lub ręczne metody selekcji, również oparte są o dane globalne pochodzące od społeczności klientów. Można również stosować sposoby opisane we wcześniejszym podrozdziale. Stosując np. metody ocen automatycznych, podobnie jak ma to miejsce w przypadku pojedynczego użytkownika, możemy przeprowadzić ocenianie produktów dla całej populacji użytkowników, korzystając z ich historii zakupów. Analiza zachowań poszczególnych użytkowników może sprowadzać się do poszukiwania grup użytkowników o podobnych zainteresowaniach. Może też pomóc w odnalezieniu trendów występujących w sprzedaży.

O ile dane wejściowe od społeczności postrzegane są jako całość, to pierwotnie pochodzą one od pojedynczych użytkowników.

Interesującą możliwością jest stosowanie komentarzy do produktów. Jednak analiza tego typu danych jest dość złożona i czasochłonna, zarówno dla systemu rekomendacji jak i dla użytkownika. Użytkownik musi bowiem przeczytać każdy komentarz i zinterpretować na ile jest on pozytywny albo negatywny. Aby uprościć ten proces większość stron oferuje oprócz komentarzy możliwość oceny w postaci liczbowej. Są to dane o wiele szybciej przyswajane oraz łatwiejsze do przetworzenia w systemie rekomendacji. Oceny odbiorcy rekomendacji jak i wszystkie oceny od społeczności wykorzystywane są w systemach kolaboracyjnych do wyliczeń optymalnych rekomendacji.

Istotnym czynnikiem są również atrybuty elementów. Mogą one pochodzić z opisu danego przedmiotu. Często dostarczane są przez producentów w katalogach produktów. Na ich podstawie właściciele mogą kategoryzować produkty. Niestety w wielu przypadkach pożądane atrybuty bywają pomijane, przez co proces dopasowania do użytkownika jest utrudniony. Dlatego ogromne znaczenie mają opinie w postaci ocen i/lub komentarzy użytkowników, które przetworzone w systemie rekomendacji mogą dostarczać znakomitych

danych dla sprzedawców jak i kupujących. Cały proces może być zautomatyzowany do takiego stopnia, że większość użytkowników nie będzie nawet wiedzieć o jego istnieniu (automatyczne oceny zachowań). [3]

3.2. Identyfikacja użytkowników

W systemach rekomendacji, przede wszystkim w tych o wysokim stopniu personalizacji, zachodzi potrzeba identyfikowania poszczególnych użytkowników. Niestety proces ten również stanowi swoiste wyzwanie. W przypadku sklepów internetowych najczęściej użytkownicy logują się (uwierzytelniają swoją osobę) do sklepu już po procesie wyboru produktów, najczęściej przy składaniu zamówienia. Uniemożliwienie przeglądania produktów lub dodawania ich do koszyka zakupowego przed uwierzytelnieniem najczęściej skutkuje poirytowaniem i rezygnacją z usług sklepu stosującego tego typu praktyki. Natomiast pokazywanie rekomendacji produktów już po procesie uwierzytelniania to zdecydowanie za późno. Istnieje kilka metod identyfikacji użytkownika, które są wykorzystywane do określenia preferencji zaraz po wejściu do sklepu, ale niestety pewność, że dany użytkownik jest tym, za którego go uważa system można mieć dopiero po procesie uwierzytelniania, gdy użytkownik zostanie zidentyfikowany i osobiście potwierdzi swoją tożsamość poprzez podanie loginu i hasła lub rejestrację w sklepie. Najczęściej wykorzystywaną metodą identyfikacji są pliki *cookies*, obsługiwane przez przeglądarki około 99% Internautów¹³.

3.2.1. Cookies

Cookie to plik przechowujący informacje na dysku użytkownika, przypisany do danej witryny, wykorzystywany w obrębie konkretnej przeglądarki, która dany plik utworzyła. Plik taki tworzony jest na polecenie aplikacji internetowej lub serwera (przekazywany w nagłówku HTTP wysyłanego do przeglądarki dokumentu). Przeglądarka tak przekazaną informację zachowuje i udostępnia serwerowi na jego polecenie. W przypadku kolejnej wizyty na stronie WWW, serwer otrzyma zapisane w pliku *cookie* informacje, które może wykorzystać np. do identyfikacji użytkownika. Plik *cookie* jest formatu tekstowego, zaś maksymalna jego pojemność wynosi 4 kilobajty. Czas przechowywania jest zmienny i może być zdefiniowany przez system rekomendacji. Najczęściej systemy rekomendacji przechowują w nim jedynie identyfikator użytkownika (do kilkunastu bajtów) pozwalający przypisać do oglądającego szereg informacji zapisanych w bazie danych na temat jego zainteresowań i preferencji.

¹³ <http://ranking.pl/pl/rankings/cookies-acceptance-rate.html>, badania statystyczne gemiusRanking PL, 07.2010

Technologia *cookie* jest uważana za bezpieczną, a użytkownik ma pełną kontrolę nad zapisywanymi i udostępnianymi informacjami. Często jednak informacje takie przechowywane są w plikach bez świadomego udziału użytkownika. Atutem metody jest identyfikowanie użytkownika niezależnie od jego adresu IP. Jest to szczególnie istotne przy dynamicznych adresach przydzielanych przy wykorzystaniu popularnych łączy ADSL. Mankamentem natomiast jest brak identyfikacji (przekazania *cookie*) w przypadku zmiany sprzętu, oprogramowania lub skorzystania z innej przeglądarki internetowej.[8]

3.2.2. Logi serwera

Logi serwera to pliki tekstowe zapisywane przez serwer, które dostarczają wielu konkretnych informacji o użytkownikach wchodzących na daną stronę. Każda próba odwołania do strony skutkuje zapisem informacji w logach. Analiza logów pozwala na określenie wielu cennych informacji:

- z jakiego adresu IP łączył się użytkownik,
- jakiej przeglądarki używał,
- datę i czas nawiązania połączenia (przesłanie żądania do serwisu),
- liczbę unikalnych wizyt,
- liczbę odwiedzin (globalnie jak i poszczególnych użytkowników),
- ilość pobranych informacji i ściągniętych plików,
- jakie słowo kluczowe wpisał użytkownik w wyszukiwarce aby odnaleźć daną stronę,
- stronę, z której użytkownik przeszedł do serwisu,
- informację jak często użytkownik oglądał dany produkt zanim dokonał jego zakupu.

Dane te dotyczą każdej podstrony serwisu. Atutem metody jest możliwość identyfikacji niezależnie od użytej przeglądarki oraz brak konieczności zapisywania plików *cookie*, które nie zawsze są obsługiwane. Minusem natomiast jest potrzeba analizy dużych ilości danych oraz brak prawidłowej identyfikacji w przypadku zmiany adresu IP (np. dynamiczny adres IP – łącza modemowe i ADSL). Może nawet dojść do sytuacji, w której użytkownikowi z innej lokalizacji geograficznej zostanie przypisany taki sam adres IP. Po wejściu do danego sklepu może on zostać źle zidentyfikowany, jako inna osoba. Ta technika polecana jest najczęściej w ulotnych systemach rekomendacji, gdzie największe znaczenie dla systemu mają stosunkowo „świeże” dane. [8]

3.2.3. Uwierzytelnianie

Identyfikacja użytkownika poprzez rejestrację lub jego uwierzytelnienie (w przypadku wcześniejszej rejestracji) jest najskuteczniejszą metodą identyfikacji. Jednak wymusza na

użytkownika wykonanie dodatkowej akcji – wypełnienie formularza, podając swoje dane. Dlatego powinno być to stosowane jedynie w ostateczności, kiedy nie ma już innego wyjścia. Mankamentem jest potrzeba zalogowania w celu identyfikacji użytkownika, co ze względu na charakter sklepów internetowych może być kłopotliwe dla użytkownika. Plusem tej metody jest duża pewność, że mamy do czynienia z daną osobą. Metoda może działać bez wykorzystania plików *cookies* oraz bez potrzeby analizy logów serwera. Jest niezależna od wykorzystywanego przez użytkownika sprzętu, oprogramowania oraz jego lokalizacji i adresu IP. [8]

3.2.4. Wielu użytkowników, jeden komputer

Dodatkowym problemem występującym zarówno przy identyfikacji na podstawie logów jak i *cookie*, może być wykorzystywanie jednego komputera i przeglądarki (w przypadku *cookie*) przez więcej niż jednego użytkownika, który przegląda ten sam serwis internetowy. Częściowym rozwiązaniem są profile użytkowników w systemach operacyjnych. Jednak nie zawsze są wykorzystywane. Najczęściej jednak do tego problemu podchodzi się sceptycznie ze względu na domniemanie osobistej znajomości między tymi osobami, co może sugerować podobieństwo zainteresowań. Potwierdzeniem tego podobieństwa jest natomiast chęć przeglądania tych samych serwisów internetowych, które może wynikać również z osobistych rekomendacji między osobami. Ostatecznie te same rekomendacje produktów w sklepie internetowym mogą mieć podobną skuteczność dla obu tych osób. Do takiego wniosku należy jednak podchodzić sceptycznie, gdyż znane są wyjątki, w postaci publicznych miejsc dostępu do komputerów jak biblioteki czy kafejki internetowe. Sprawia to, że największą pewność co do identyfikacji danego użytkownika uzyskujemy dopiero po jego uwierzytelnieniu.

3.3. Metody rekomendacji

Bardzo istotnym elementem systemów rekomendacji produktów jest to, aby rekomendacje były „świeże”, czyli obliczane bezpośrednio przed wysłaniem ich wraz ze stroną do przeglądarki internetowej użytkownika. Daje to największą efektywność tego typu działań. Ze względów wydajnościowych nie wszystkie metody na to pozwalają. W przypadku niektórych metod wykorzystuje się obliczenia w trybie offline. Preferowane są natomiast rekomendacje online, które w sposób optymalny wykorzystują preferencje odbiorcy, poprzez natychmiastową odpowiedź na jego potrzeby. Większość procesów rekomendacji powinna być przeprowadzana w czasie rzeczywistym. Metody takie jak wyszukiwanie, selekcja

ręczna, podsumowanie statystyczne, korelacja atrybutów są wydajne i w miarę proste do zastosowania, co sprawia że mogą być wyliczane w czasie rzeczywistym. Natomiast systemy korelacji obiektów bądź użytkowników są bardziej złożone i czasochłonne. Dlatego najczęściej metody te wykorzystują obliczenia w trybie offline, w celu przygotowania modelu danych, który można skutecznie wykorzystać w czasie rzeczywistym. Swoistym wyzwaniem dla programistów jest takie zaprojektowanie algorytmu, aby prezentowane rekomendacje w czasie rzeczywistym były jak najlepiej dostosowane do różnych akcji przeprowadzanych na stronie internetowej przez użytkownika [3, 4].

3.3.1. Na podstawie danych demograficznych

Metoda ta oparta jest na cechach demograficznych przechowywanych w profilu odbiorcy. Wykorzystanie takiej metody sprowadza się do stereotypowego rozumowania, np. kobiety wolą kosmetyki niż samochody. Dla wielu użytkowników rekomendacje mogą być zbyt ogólne. Metoda nie przewiduje też adaptacji do zainteresowań odbiorcy, które mogą się zmieniać wraz z upływem czasu. Nie jest też odporna na błędy powstałe w procesie uzupełniania profilu demograficznego odbiorcy [5].

3.3.2. Wyszukiwanie (ang. raw retrieval)

Strona internetowa dostarcza odbiorcy mechanizmu wyszukiującego. Odbiorca może w ten sposób odpytać bazę danych o pożądany przedmiot poszukiwań. Chociaż technicznie nie stosuje się w tym przypadku systemów rekomendacji, odbiorca może otrzymać interesujące go rezultaty co przyspieszy proces dotarcia do informacji bądź przedmiotu. W konsekwencji osoba przeglądająca rezultaty wyszukiwania może dojść do wniosku, że są one rekomendowanymi odpowiedziami na jego potrzebę (zapytanie). Przykładem może być poszukiwanie albumów „The Rolling Stones” w sklepie z muzyką. System wyszukiwania zwróci listę albumów, która może być pomocna i może doprowadzić klienta do albumu, o którym wcześniej nie wiedział [3].

3.3.3. Selekcja ręczna (ang. manualy selected)

Informacje lub przedmioty są rekomendowane poprzez ręczny wybór przez pracowników sklepów, krytyków i innych ekspertów. Lista rekomendacji tworzona jest na podstawie własnych upodobań, zainteresowań i celów tych osób. Ich rekomendacjom często towarzyszą komentarze pomagające zrozumieć ocenę i zalecenia. Przykładem takiej selekcji mogą być rekomendacje np. ulubionych filmów lub powiązanych stron internetowych.

W serwisach tworzących wokół siebie społeczności można spotkać się z możliwością tworzenia przez osoby list zaleceń lub rekomendacji. Takie listy mogą posłużyć do budowy list TOP10 (np. dziesięć najpopularniejszych) przy minimalnym obciążeniu serwerów. Proces ten najbardziej naśladuje możliwości tradycyjnych krytyków i redaktorów, w tym potencjalne błędy [3].

3.3.4. Podsumowanie statystyczne (ang. statistical summaries)

W przypadku kiedy personalizacja nie jest praktyczna lub jest niepotrzebna, aplikacje internetowe mogą w bardzo skuteczny sposób wykorzystywać mechanizmy statystyczne w celu podsumowania opinii społecznych. Na takie podsumowania składają się m.in. popularność w społeczności (odsetek osób, które lubią daną rzecz lub ją zakupiły), zbiorcze podsumowania (liczba osób, które rekomendują produkt, średnia ocena produktu). Takie dane często wykorzystywane są do tworzenia list rankingowych ułatwiających rekomendacje produktu (najczęściej kupowane, najlepiej oceniane). Są to jednak nie indywidualne rekomendacje, lecz zbiorcze wyliczenia. Przewagą tej metody jest łatwość obliczeń oraz możliwość wykorzystania danych w marketingu tradycyjnym (np. promocja produktów najbardziej popularnych) [3].

3.3.5. Korelacja atrybutów (ang. attribute-based)

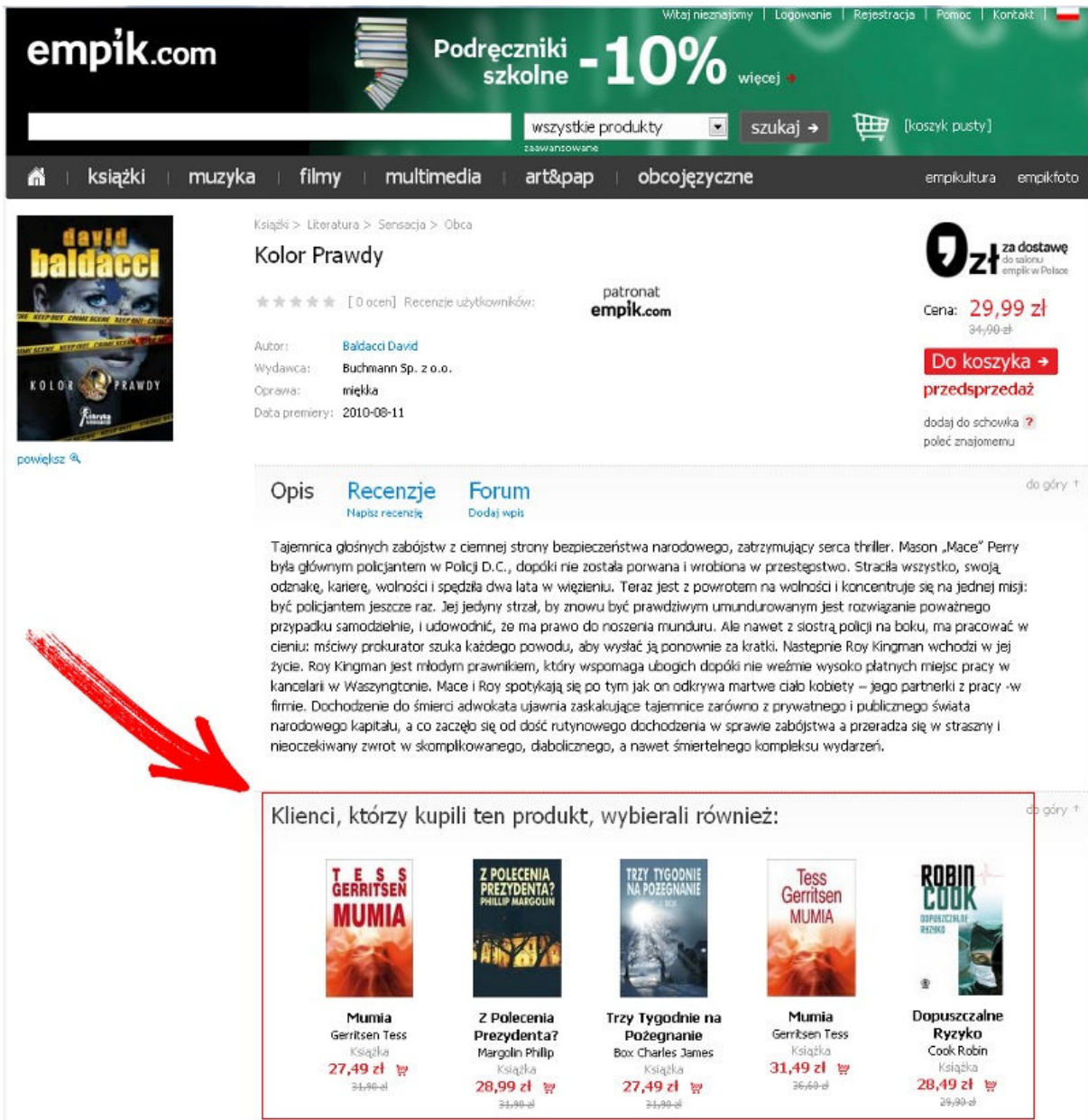
Rekomendacje oparte na zainteresowaniu odbiorcy atrybutami poszukiwanego przedmiotu. Najprostsze tego typu rekomendacje mogą być zwykłymi wynikami wyszukiwania (raw retrieval). W bardziej zaawansowanych systemach wykorzystuje się atrybuty modelu zainteresowań klienta. Użytkownik poszukujący słowników w języku angielskim może powiedzieć, np. poprzez system nawigacji księgarni, że poszukuje „słowników”, których atrybutem będzie „angielski”. Osobie, która przegląda kategorię muzyka rockowa w sklepie z muzyką i dodaje do koszyka tylko przecenione płyty, można wyświetlić inne płyty sugerując rabat. Innym przykładem może być analiza koszykowa lub inna metoda pozyskania wiedzy o profilu użytkownika w celu wyświetlania tylko produktów będących w zakresie zainteresowań. Jeśli dana osoba kupuje tylko przecenione rzeczy można skoncentrować się na rekomendowaniu kolejnych przecenionych rzeczy. Jeśli ktoś, o dużej historii zakupów, nigdy nie kupił płyt z muzyką z lat 80-tych można przypuszczać, że nie będzie taką muzyką zainteresowany i nie wyświetlać rekomendacji takiej muzyki [3].

Szczególnym przypadkiem tej metody rekomendacji są rekomendacje oparte o zawartość stron (ang. content based) wykorzystują one opisy treści lub elementu (atrybuty) wcześniej ocenionych (np. poprzez zapisanie zdarzenia wizyty) przez odbiorcę elementów,

by znaleźć związek pomiędzy odbiorcą, a opisem innego elementu. Takie podejście niestety jest zbyt uzależnione od obiektywnych opisów rekomendowanych elementów, które z kolei często mogą być zbyt wyspecjalizowane [5].

3.3.6. Korelacja obiektów (ang. item-to-item correlation)

Metoda oparta na identyfikowaniu obiektów, które często występują w asocjacji z innymi obiektami, którymi zainteresowany jest odbiorca. Powiązania między obiektami mogą być oparte o analizę koszykową, wspólne preferencje klientów i inne. W najprostszych zastosowaniach można spotkać się ze znalezieniem przedmiotów pasujących do danego przedmiotu, np. inne akcesoria komputerowe kupowane wraz z płytą główną (przykład został przedstawiony na rysunku 3). Bardziej zaawansowane systemy wykorzystują informacje o wszystkich przedmiotach zainteresowań danej osoby (np. wszystkie przedmioty z koszyka) do rekomendacji odpowiednich, innych przedmiotów. Systemy oparte o korelacje obiektów zwykle wykorzystują bieżące dane o zawartości koszyka lub innych zainteresowaniach odbiorcy, a nie długoterminową historię danego klienta. Dzięki temu świetnie nadają się do rekomendacji prezentów. Klient szukający prezentu musi jedynie określić elementy zainteresowania (w tym przypadku elementy zainteresowania odbiorcy prezentu) np. poprzez przeglądanie konkretnych ofert produktowych, w celu dostosowania prezentu do odbiorcy, a nie kupującego [3].



empik.com Podręczniki szkolne **-10%** więcej

Witaj nieznanomy | Logowanie | Rejestracja | Pomoc | Kontakt

wszystkie produkty [szukaj] [koszyk pusty]

książki | muzyka | filmy | multimedia | art&pap | obcojęzyczne

empikultura empikfoto

Książki > Literatura > Sensacja > Obca

Kolor Prawdy

★★★★★ [0 ocen] Recenzje użytkowników:

patronat empik.com

Autor: **Baldacci David**

Wydawca: **Buchmann Sp. z o.o.**

Oprawa: **miękka**

Data premiery: **2010-08-11**

9 zł za dostawę do salonu empik w Polsce

Cena: **29,99 zł** 34,90 zł






Do koszyka → **przedsprzedaż**

dodaj do schowka poleć znajomemu

Opis **Recenzje** Forum

Tajemnica głośnych zabójstw z ciemnej strony bezpieczeństwa narodowego, zabrzmiący serca thriller. Mason „Mace” Perry była głównym policjantem w Policji D.C., dopóki nie została porwana i wrobiona w przestępstwo. Straciła wszystkie, swoją odznakę, karierę, wolności i spędziła dwa lata w więzieniu. Teraz jest z powrotem na wolności i koncentruje się na jednej misji: być policjantem jeszcze raz. Jej jedynym strzał, by znowu być prawdziwym umundurowanym jest rozwiązanie poważnego przypadku samobójstwa, i udowodnić, że ma prawo do noszenia munduru. Ale nawet z siostrą policji na boku, ma pracować w cieniu: mściwy prokurator szuka każdego powodu, aby wystać ją ponownie za kratki. Następnie Roy Kingman wchodzi w jej życie. Roy Kingman jest młodym prawnikiem, który wspomaga ubogich dopóki nie weźmie wysoko płatnych miejsc pracy w kancelarii w Waszyngtonie. Mace i Roy spotykają się po tym jak on odkrywa martwe ciało kobiety – jego partnerki z pracy -w firmie. Dochodzenie do śmierci adwokata ujawnia zaskakujące tajemnice zarówno z prywatnego i publicznego świata narodowego kapitału, a co zaczęło się od dość rutynowego dochodzenia w sprawie zabójstwa a przeradza się w straszny i nieoczekiwany zwrot w skomplikowanego, diabolicznego, a nawet śmiertelnego kompleksu wydarzeń.

Klienci, którzy kupili ten produkt, wybierali również:

 <p>Mumia Gerritsen Tess Książka 27,49 zł</p>	 <p>Z Polecenia Prezydenta? Margolin Philip Książka 28,99 zł</p>	 <p>Trzy tygodnie na Pożegnanie Box Charles James Książka 27,49 zł</p>	 <p>Mumia Gerritsen Tess Książka 31,49 zł</p>	 <p>Dopuszczalne Ryzyko Cook Robin Książka 28,49 zł</p>
--	---	--	--	--

Rys 3. Zrzut ekranu prezentujący wykorzystanie rekomendacji. Źródło: <http://empik.com>, sklep internetowy, 08.2010

3.3.7. Korelacja użytkowników (ang. user-to-user correlation)

Metoda oparta na rekomendowaniu na podstawie asocjacji pomiędzy odbiorcą, a innymi osobami, które np. kupiły produkty w sklepie internetowym. Technologia ta często nazywana jest filtrowaniem kolaboratywnym (collaborative filtering), ponieważ powstała jako technika filtrowania używająca opinii grup użytkowników w celu wygenerowania rekomendacji dla indywidualnego odbiorcy. Mechanizm tego typu wykorzystuje korelację użytkowników w celu identyfikacji społeczności klientów, którzy mają tendencje do posiadania bądź lubienia tego

samego zestawu przedmiotów. Zasadą tego mechanizmu jest to, że jeśli kilku członków społeczności danego odbiorcy jest posiadaczem bądź lubi np. książkę pt. Harry Potter, to jest wysoce prawdopodobne, że odbiorca rekomendacji również będzie tą książką zainteresowany. Metoda najczęściej wykorzystuje algorytmy korelacji liniowej w celu identyfikacji najbliższego sąsiedztwa danego odbiorcy (ang. nearest neighbour), które jest osobą bądź grupą osób najbardziej pasujących do odbiorcy pod względem preferencji i zainteresowań. Może być implementowana przy wykorzystaniu również wielu innych algorytmów [3].

3.3.8. Metody mieszane

Najskuteczniejsze systemy rekomendacji wykorzystują różne techniki i metody w celu optymalizacji efektu. Często łącząc opisane wyżej metody. Powstałe w ten sposób systemy są dostosowywane do stron internetowych na których działają. Uzależnione od bardzo wielu czynników np. charakteru branży, celu działania, profilu odbiorców, liczby elementów czy użytkowników i wielu innych. Można stwierdzić, że każda strona internetowa niesie ze sobą potrzebę, co najmniej, dostosowania takiego systemu do charakteru działalności, jaką prowadzi. Celem wykorzystania metod mieszanych, oprócz samego zwiększenia efektywności i jakości rekomendacji, może być, zwłaszcza przy bardzo dużych ilościach danych, poprawa wydajności systemu, która pośrednio przyczynia się do polepszenia rekomendacji poprzez bardziej aktualne ich generowanie (szybciej lub nawet w czasie rzeczywistym). Wykorzystanie mieszanych metod rekomendacji prowadzi do powstawania coraz większej liczby systemów, optymalizowanych dla danej działalności. Przykładem może być złożony system rekomendacji filmów stworzony przez grupę *BellKor's Pragmatic Chaos* dla firmy *Netflix*, wykorzystujący kilkadziesiąt różnych algorytmów w celu optymalizacji rezultatów^{14 15 16}.

¹⁴ http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf, Koren Y., The BellKor Solution to the Netflix Grand Prize, 07.2010

¹⁵ http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf, Toscher A., Jahrer M., The BigChaos Solution to the Netflix Grand Prize, 07.2010

¹⁶ http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf, Piotte M., Chabbert M., The Pragmatic Theory solution to the Netflix Grand Prize, 07.2010

3.4. Powszechnie stosowane techniki

Systemy rekomendacji wykorzystują bardzo dużo różnych technik i algorytmów w swoim działaniu. W tabeli 1 przedstawiono klasyfikację najpowszechniej stosowanych technik. Oprócz podziału na metody wyróżnić można dwa typy stosowanych technik do predykcji rekomendacji.

Tabela 1. Klasyfikacja powszechnie stosowanych technik rekomendacji. Źródło: [6]

Metoda rekomendacji	Powszechnie stosowane techniki rekomendacji	
	Oparte o heurystykę	Oparte o model
Oparta o zawartość (content-based)	<ul style="list-style-type: none"> • TF – IDF (wyszukiwanie informacji) • klasteryzacja 	<ul style="list-style-type: none"> • klasyfikator Bayes'a • klasteryzacja • drzewa decyzyjne • sztuczne sieci neuronowe
Oparta o filtrowanie kolaboratywne (korelacje elementów, użytkowników)	<ul style="list-style-type: none"> • najbliższy sąsiad (cosinus, korelacja) • klasteryzacja • teoria grafu 	<ul style="list-style-type: none"> • sieci Bayes'a • klasteryzacja • sztuczne sieci neuronowe • regresja liniowa • modele probabilistyczne
Hybrydowa (mieszana – połączenie kilku metod)	<ul style="list-style-type: none"> • liniowe połączenie przewidzianych wyników • połączenie różnych systemów oceniania • zawierające jeden element jako część heurystyki dla innej metody 	<ul style="list-style-type: none"> • zawierające jeden komponent jako część modelu dla innej metody • budowanie jednego ujednoczonego modelu

Metody oparte o heurystykę, zwane też metodami opartymi o obliczenia w pamięci, pozwalają na wyliczenie przybliżonych wartości na podstawie próby lub całości zgromadzonych danych. W przypadku najczęściej stosowanych algorytmów opartych o filtrowanie kolaboratywne, każdy użytkownik jest częścią grupy o podobnych zainteresowaniach. Poprzez określenie tzw. sąsiadów danego użytkownika, wykonywane są dla niego obliczenia predykcji preferencji dla elementów. Najczęściej wyliczane są

podobieństwa, odzwierciedlone w postaci odległości, korelacji lub średniej ważonej pomiędzy dwoma użytkownikami lub dwoma elementami. Algorytmy tego typu wymagają pewnej minimalnej ilości danych do generowania dobrych rekomendacji. Nie nadają się do generowania rekomendacji dla nowych użytkowników lub elementów.

Metody oparte o model, najpierw określają modele, a następnie wykorzystują je do końcowych wyliczeń. Najczęściej są to algorytmy wykorzystujące maszyny uczące oraz techniki eksploracji danych. Systemy tego typu uczą się rozpoznawać pewne wzorce zachowań na danych treningowych, po to aby w sposób inteligentny generowały predykcje na danych realnych, wykorzystując przy tym wyuczone modele. System taki może wcześniej np. określić 200 profili, będących modelem, do których przypisze, podczas generowania rekomendacji, poszczególnych użytkowników.

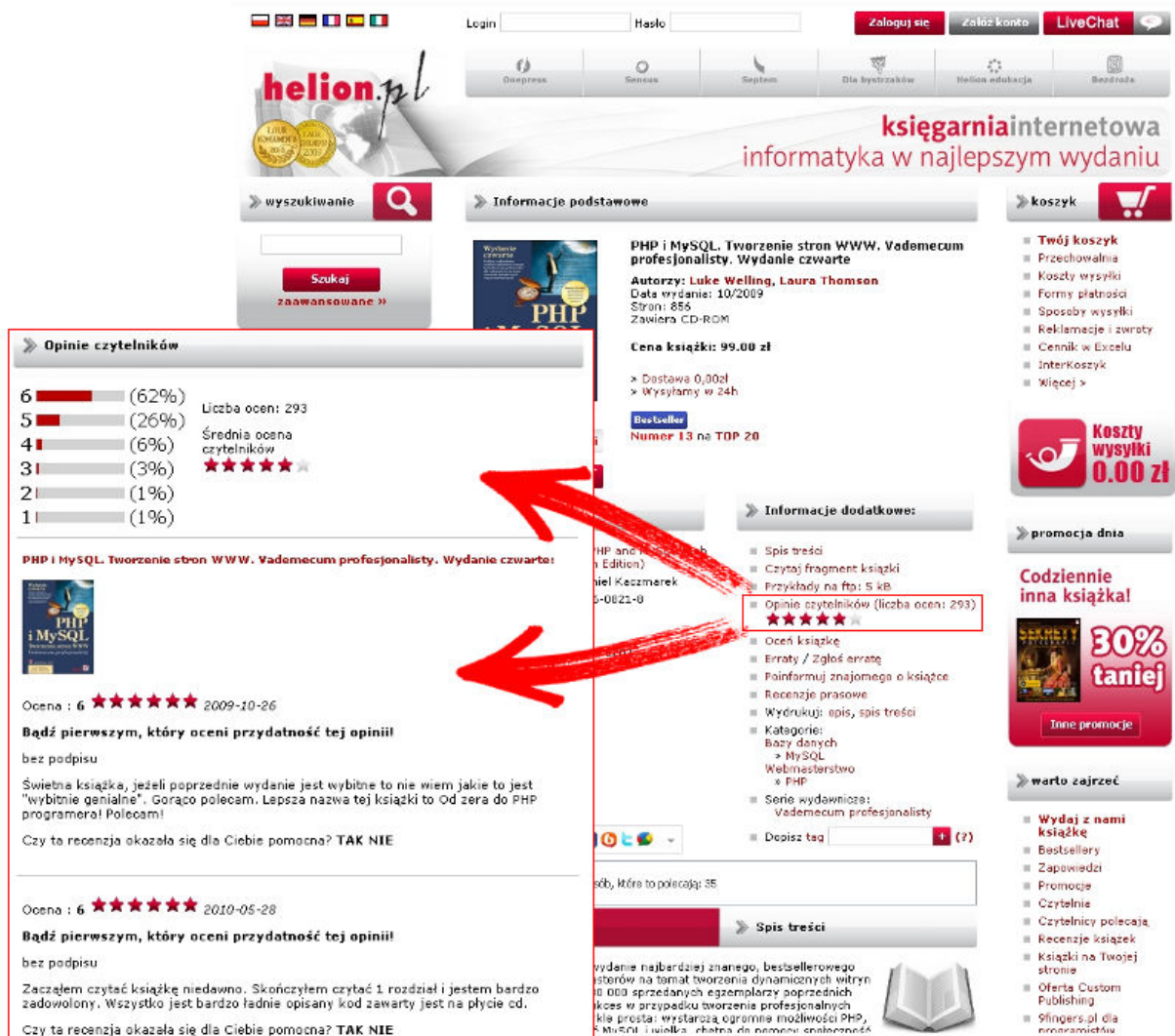
Metody oparte o model cechują się bardzo dużą szybkością generowania rekomendacji, często szybszą niż w przypadku metod opartych o heurystykę. Wymagają jednak budowy modelu, który jest dużym obciążeniem dla systemów informatycznych. Jednak nie zmienia się on często, dlatego można go budować rzadziej, np. w nocy, w czasie mniejszego obciążenia serwerów. Oprócz tego systemy budujące model mogą tracić wiele cennych informacji podczas czynności związanych z redukcją ilości danych (filtrowanie anomalii, algorytmy *SVD* i inne). Jednak w przeciwieństwie do systemów opartych o heurystykę mogą generować dobre rekomendacje dla rzadkich danych, a nawet dla nowych użytkowników i elementów. Lepiej radzą sobie również z problemami skalowalności. Systemy oparte o model zalecane są w przypadku dużych ilości danych. Odnośniki do badań opartych na przedstawionych w tabeli 1 technik można znaleźć w źródle [2]. [2, 6]

3.5. Dane wyjściowe systemu

Dane wyjściowe, prezentowane użytkownikowi w postaci rekomendacji różnią się pod względem typu, ilości czy wspomnianego już wyglądu. Najczęstszym typem rekomendacji są sugestie postaci „sprawdź to” lub po prostu wprowadzenie „tego” na stronę przeglądaną przez użytkownika. Jednak takie podejście ogranicza do rekomendacji jednej pozycji. Może to być pozytywne w procesie wyboru – ogranicza to również proces decyzyjny użytkownika, co zwiększa szansę na zainteresowanie. Jest to jednak ryzykowne ze względu na to że klient, w przypadku sklepów internetowych, może już posiadać dany produkt lub wiedzieć o nim coś czego nie wie sprzedający, co przyczyni się do rezygnacji z zakupu. Reklamy kierowane na użytkownika typu różne promocje np. „kupując razem oszczędzasz”, „dodając do koszyka jeszcze jeden produkt pokryjemy koszty wysyłki” itp., również wykorzystują

rekomendacje produktów w celu zwiększenia sprzedaży. Systemy rekomendacji stanowią zestaw wskazówek dla użytkownika w danym kontekście. Niektóre prezentują nieuporządkowane listy rekomendacji, aby uniknąć wrażenia, że konkretna rekomendacja jest najlepsza. W ten sposób można uniknąć zniechęcenia do rekomendacji tylko na podstawie odrzucenia pierwszej pozycji (teoretycznie najlepszej). Każda lista jednak powinna prezentować pewien ład. Nawet nieuporządkowane listy, pod względem prawdopodobieństwa zainteresowania, są celowo przedstawiane np. w kolejności alfabetycznej lub na podstawie danych o towarach polecanych. Każdy sposób uporządkowania listy zapewnia dodatkowe informacje, które mogą być pomocne dla użytkownika.

Istnieją systemy, które prezentują przewidywane oceny użytkownika dla danego elementu. Szacunki te mogą być przedstawiane jako indywidualne lub niespersonalizowane dla typowych członków społeczności. Prezentacja takich ocen może pomóc użytkownikowi zrozumieć siłę rekomendacji. Przewidywana ocena może być wyświetlana w ramach indywidualnych zaleceń, list zaleceń lub może być wyświetlana w kontekście jako element informacji.



The screenshot shows the Helion.pl website interface. At the top, there are language selection options and login fields. The main header includes the Helion.pl logo and navigation links like 'Dziennik', 'Senas', 'System', etc. Below the header, there's a search bar and a 'wyszukiwanie' button. The main content area displays a book titled 'PHP i MySQL. Tworzenie stron WWW. Vademecum profesjonalisty. Wydanie czwarte'. The book's price is listed as 99.00 zł. A 'Opinie czytelników' (User Reviews) section is highlighted with a red box and arrows. This section shows a star rating of 6 stars (62% of 293 reviews) and two user comments. The first comment is dated 2009-10-26 and the second is dated 2010-05-28. Both reviewers gave the book a 6-star rating. The review section also includes a 'Bądź pierwszym, który oceni przydatność tej opinii!' prompt and a 'Czy ta recenzja okazała się dla Ciebie pomocna?' question.

Rys 4. Zrzut ekranu prezentujący oceny i opinie użytkowników. Źródło: <http://helion.pl>, księgarnia internetowa, 08.2010

W przypadku małych społeczności użytkowników lub społeczności, które dobrze się znają przydatnym sposobem prezentacji jest wyświetlanie indywidualnych ocen członków społeczności, aby użytkownik docelowy mógł wykorzystać tę listę do własnej oceny elementu. Technika ta jest szczególnie przydatna gdy użytkownik jest powiązany w społeczności ze swoimi znajomymi (przykład *Facebook.com*) lub gdy ocenie towarzyszą opinie (rysunek 4). Opinie są przykładem rekomendacji zawierającej ocenę, ale nie są w pełni zrozumiałe dla systemów informatycznych. Trudno jest odróżnić systemowi rekomendacji opinię pozytywną od negatywnej. Dlatego często oprócz możliwości komentowania, recenzenci dostają możliwość oceniania. Oceny analizowane są przez system rekomendacji, a tekst opinii daje użytkownikom dodatkowe uzasadnienie danej oceny. Takie firmy jak *Helion.pl*, *Allegro.pl* prezentują zarówno oceny pozycji (książki,

partnerzy handlowi) jak i indywidualne komentarze i oceny. Oznacza to, że każdy użytkownik widzi to samo, kompletny zestaw komentarzy wraz z ocenami. Na podstawie tych danych może podjąć decyzję o zakupie. Należy jednak pamiętać o zachowaniu wysokiej wiarygodności takich opinii, co nie zawsze jest procesem prostym. Należałoby zabezpieczyć się przed fałszywymi opiniami i nadać wartość wagową każdej opinii z osobna. Taki mechanizm stosowany jest w *Amazon.com*, który udostępnia „meta-rating”, pozwalający na komentowanie opinii. Takie komentarze następnie mogą posłużyć do nadawania, w sposób automatyczny, wartości wagowej opiniom. W podobny sposób postępuje firma *Helion.pl* udostępniająca binarny mechanizm oceny opinii: „Czy ta recenzja okazała się dla Ciebie pomocna?” widoczny na rysunku 4. [3]

3.6. Stopień personalizacji

Rekomendacje mogą być prezentowane w różnym stopniu personalizacji, obejmującym zarówno ich dokładność jak i przydatność oraz inne czynniki. Szczególne znaczenie dla poprawy systemu rekomendacji jest jego przydatność obejmująca trafność – wtedy, gdy system potrafi uzyskiwać cenne, ale niespodziewane rezultaty oraz indywidualizm – wtedy, gdy system zapewnia różne rezultaty dla różnych osób. Dobry system rekomendacji zalecający jedynie książki najlepiej sprzedające się jest mniej wartościowy od systemu, który prezentuje indywidualne rekomendacje użytkownikom pod kątem ich zainteresowań. Podobnie system rekomendujący mało trafne elementy, nie będzie dobrym pomysłem w dłuższym terminie. Dopóki na jakość personalizacji składa się tak wiele czynników można określić trzy poziomy personalizacji. [3, 4]

3.6.1. Niespersonalizowana

Kiedy system rekomendacji prezentuje takie same rekomendacje dla każdego użytkownika, system jest nazywany niespersonalizowanym. Tego typu rekomendacje oparte są o wyszukiwanie ręczne, podsumowanie statystyczne lub inne metody. Wiele sklepów internetowych wykorzystuje niespersonalizowane systemy w swoim działaniu, jako systemy podstawowe. Przykładem mogą być rankingi – najlepiej sprzedające się, najlepsi sprzedawcy, nie filtrowane komentarze, średnie oceny itd.. Wszystkie tego typu rekomendacje prezentują taką samą treść każdemu użytkownikowi. [3]

3.6.2. Personalizacja ulotna

Personalizacja ulotna wykorzystuje aktualne dane uzyskane od użytkownika w celu dostosowania rekomendacji do jego zainteresowań. Daje to rekomendacje wyświetlane w ulotnym kontekście, dostosowane do klienta, jego ścieżki nawigacyjnej lub wyboru jakiego dokonuje. Poszczególne implementacje mogą być mniej lub bardziej osobiste. Systemy o wysokim stopniu personalizacji ulotnej wykorzystują bieżącą sesję przeglądania lub bieżącą zawartość koszyka zakupów. Przeciwnie systemy o niskim stopniu personalizacji ulotnej wykorzystują powiązania z aktualnie przeglądaną stroną, co powoduje że rekomendacje są często zbliżone do niespersonalizowanych. Personalizacja ulotna wykorzystuje najczęściej korelacje elementów lub atrybutów, wyszukiwanie lub metody mieszane oparte o powyższe. [3]

3.6.3. Personalizacja ciągła

Najbardziej spersonalizowane systemy rekomendacji używają personalizacji ciągłej do prezentacji różnych rekomendacji dla różnych użytkowników, nawet jeśli użytkownicy przebywają na tej samej stronie. W tego typu rekomendacjach wykorzystuje się metody oparte o korelacje użytkowników oraz o korelacje atrybutów lub elementów połączoną z ciągłym monitorowaniem preferencji użytkownika. Systemy tego typu wymagają obecności użytkowników oraz zaimplementowanych algorytmów identyfikacji, w celu poznania ich preferencji w dłuższym okresie czasu. W zamian system jest w stanie dostarczać rekomendacje o najwyższym, osobistym stopniu personalizacji. [3, 4]

3.7. Sposób dostarczania

Sposób dostarczenia rekomendacji pasujący do aktywności użytkownika jest kwestią kluczową w projektowaniu rozwiązania. W rzeczywistości sklepy internetowe stanowią analogię do modeli tradycyjnej akwizycji i sprzedaży detalicznej. Analogię stanowi również tradycyjnie stosowany marketing. Od dawna wykorzystuje się mailing tradycyjny, telemarketing próbując pozyskać nowych klientów lub rozpocząć nową sesję zakupów. Podobnie wygląda to w przypadku sklepów internetowych.

Preferencje w metodach dostarczania rekomendacji zmieniają się z upływem czasu. Wczesne aplikacje wykorzystywały rekomendacje z naciskiem lub na żądanie. Również w celu optymalizacji wydajności i pokazaniu klientom, że się o nich dba. Aktualnie aplikacje wykorzystują częściej sposób pasywny do rekomendacji interakcji z użytkownikiem, w chwili przebywania na stronie, oraz sposób nacisku do sprowadzenia użytkowników z powrotem na

stronę internetową. Trzy sposoby dostarczania rekomendacji zostały przedstawione w kolejnych podrozdziałach. [3]

3.7.1. Z naciskiem (ang. push)

Rekomendacje z naciskiem mają za zadanie dotrzeć do klienta, nawet jeśli nie pozostaje on w interakcji ze sklepem internetowym lub inną witryną. E-mail sprzedażowy jest obecnie najczęściej stosowaną techniką dostarczenia rekomendacji dla takich osób. Wysyłane rekomendacje często są powiązane z ofertami promocyjnymi i mają na celu zachęcić odbiorcę do powrotu do witryny (sklepu internetowego). Obecnie stosowane technologie pozwalają na przygotowanie spersonalizowanej wiadomości e-mail, zawierającej m.in. bezpośredni odnośnik do rekomendowanego produktu. [3] Szacuje się, że przy dobrze przygotowanym e-mailu, liczba otwieranych wiadomości tego typu może sięgać nawet 50%, a liczba kliknięć w linki rekomendacyjne w mailach nawet 10% wszystkich wysyłanych do klientów.¹⁷

3.7.2. Na żądanie (ang. pull)

Systemy rekomendacji wykorzystujące ten sposób dostarczania umożliwiają użytkownikowi kontrolę procesu wyświetlania rekomendacji. To użytkownik decyduje, kiedy chce zobaczyć rekomendacje. Jest świadomy dostępności rekomendacji, ale nie są one wyświetlane w sposób aktywny. Prośba o pokazanie rekomendacji jest realizowana w reakcji na akcję użytkownika (np. kliknięcie w odnośnik). Może pojawić się w różnym kontekście np. „pokaż oceny produktu”, „znajdź prezent”, „pokaż zalecaną kategorię” itd. Niektóre, wczesne aplikacje wykorzystywały ten sposób dostarczania rekomendacji jako podstawowy, ze względu na to, że obliczenia dla rekomendacji były drogie i mogły powodować spowolnienie interaktywności witryny internetowej. Obecnie wykorzystanie tej metody leży w kwestii projektantów. Stosowana jest w takich opcjach jak rekomendacje prezentów czy listy 10 (TOP 10). W innych przypadkach rekomendacje są najczęściej integralną częścią strony internetowej. [3]

3.7.3. Pasywny (ang. passive)

Pasywny sposób dostarczania rekomendacji, przedstawia je w naturalnym kontekście działania witryny internetowej. Czasami można spotkać się z nazwą naturalny lub organiczny sposób dostarczania rekomendacji. Systemy wyświetlające rekomendacje stron lub

¹⁷ Chaffey D., Total E-mail Marketing, Elsevier, 2007

produktów powiązane z aktualnie przeglądany, rekomendacje związane z tematem aktualnie czytanej artykułu lub dostosowane do kontekstu poszukiwań użytkownika zaliczane są do tej grupy. Zaletą tego typu prezentacji rekomendacji jest dotarcie do użytkownika w momencie, gdy jest on już otwarty na nowe pomysły. Sklepy internetowe wykorzystują rekomendacje pasywne w procesie składania zamówienia, np. w celu modyfikacji opcji dostawy już po procesie akceptacji zamówienia. Rekomendacje pasywne nie są aktywnie odbierane przez użytkowników. Mogą oni nie mieć świadomości ich wykorzystania na stronie. Natomiast mają one charakter wtapiania się w ogólne doświadczenia użytkownika związane ze stroną internetową. [3]

3.8. Problemy systemów rekomendacji

W zależności od wyboru metody rekomendacji można spotkać się z różnymi problemami. Każda z metod rekomendacji ma swoje mocne i słabe strony. Niektóre z możliwych problemów zostały przedstawione w tabeli 2. Algorytmy kolaboratywne oraz niektóre oparte na analizie zawartości strony nie radzą sobie z nowymi użytkownikami i produktami, dla których nie ma historii zakupów lub ocen. Kolejną słabością jest rzadkość danych. Problematyczne może być oszacowanie rzetelnych korelacji pomiędzy użytkownikiem a produktem w środowisku z dużą ilością pustych danych. Może to również powodować generowanie zbyt krótkich list rekomendacji.

Metody zależne od użytkownika, takie jak filtrowanie kolaboratywne i demograficzne wykazują konieczność identyfikacji użytkownika, przez co nie mogą być stosowane w anonimowych systemach. Wymagają również wprowadzania przez użytkowników danych osobowych, co często jest dla nich bardzo uciążliwe.

Podsumowanie statystyczne, filtrowanie kolaboratywne, jak również wzorce sekwencyjne z uwagi na analizowanie danych już zgromadzonych nie są w stanie polecić nowych pozycji dla użytkownika ze względu na brak informacji o sprzedaży, oglądalności lub ocenach / opiniach dla nowego elementu. Inaczej jest w przypadku metod opartych o atrybuty lub analizę zawartości (opisu produktu). Te metody wykorzystują wprowadzone informacje od razu po dodaniu nowego elementu.

W przypadku sklepów internetowych można wyróżnić dwa rodzaje stron. Strony produktów prezentujące odpowiednie oferty i opisy produktów oraz inne strony nie produktowe zawierające informacje o firmie, warunkach gwarancji, wysyłki, nowości z branży, recenzje produktów, itp. Strony nie produktowe rzadko są powiązane z konkretnymi produktami. Metody oparte wyłącznie na korelacjach obiektów, nie radzą

sobie z prezentacją rekomendacji na stronach nie produktowych, ponieważ nie są w stanie wybrać produktów pokrewnych. [4]

Dla większości z przedstawionych tu problemów istnieją propozycje rozwiązań. Często stosuje się systemy hybrydowe oparte o kilka metod, które dzięki temu są w stanie prezentować rekomendację niezależnie od problemów występujących w metodach stosowanych indywidualnie. Twórcy systemów rekomendacji biorą pod uwagę słabości stosowanych metod i w odpowiedni sposób na nie reagują. Przykładem może być system oparty o rozdzielne metody rekomendacji (korelacje obiektów i podsumowanie statystyczne), który wyświetla produkty np. najczęściej kupowane, zamiast produktów powiązanych z nowym produktem, dopóki nie uzyska odpowiedniej ilości danych pozwalających na prezentację rzetelnych rekomendacji dla nowego produktu poprzez metodę korelacji obiektów. [6]

Tabela 2. Cechy podstawowych metod rekomendacji. „+” oznacza rozwiązany problem, „-” nierozwiązany. Źródło: [4]

Metoda	Źródło danych	Adaptacja w zależności od użytkownika	Adaptacja w zależności od przeglądanej pozycji	Rozwiązanie problemu nowych pozycji	Rozwiązanie problemu nowych użytkowników	Rozwiązanie problemu rzadkości danych
Podsumowanie statystyczne	zakupy, oceny, wizyty	-	-	--	+	+
Filtrowanie kolaboratywne	oceny	+	-	--	-	-
Powiązania, wzorce sekwencyjne	zakupy, koszyk zakupów, ścieżki nawigacyjne	-	+	--	+	-/+
Na podstawie zawartości (ang. content based)	atrybuty pozycji, zawartość tekstowa	-	+	+	+	+

Na podstawie danych demograficznych	atrybuty pozycji i użytkowników	+	-	+	+	+
-------------------------------------	---------------------------------	---	---	---	---	---

3.8.1. Problem nowego użytkownika (ang. cold start problem)

Problem polega na braku połączenia nowego użytkownika z obiektami (produktami w sklepie). System rekomendacji musi najpierw „nauczyć” się preferencji danego użytkownika poprzez np. poznanie (minimalnej ilości) ocen, jakie ten wystawia produktom, albo wprowadzone dane personalne w przypadku systemów opartych o dane demograficzne. Problem ten występuje również w przypadku użytkowników, o których posiadamy informacje potrzebne do generowania rekomendacji, ale których nie można zidentyfikować przed procesem uwierzytelniania (logowania) w systemach opartych o korelacje użytkowników.

Istnieje kilka technik adresowanych rozwiązaniu tego problemu, większość z nich oparta jest na rozwiązaniach hybrydowych np. połączeniu odseparowanych metod kolaboratywnych i opartych na zawartości. [6]

3.8.2. Problem nowego elementu

Problem występuje w przypadku dodania nowego produktu lub strony. Nowe elementy są ciągle dodawane do systemów rekomendacji. Systemy kolaboratywne oparte są wyłącznie na preferencjach użytkowników względem produktów. Dlatego dopóki nowy produkt nie zostanie oceniony przez znaczną liczbę użytkowników, system nie będzie go uwzględniał przy generowaniu rekomendacji.

Ten problem również rozwiązywany jest poprzez zastosowanie metod hybrydowych. Można również obliczyć rekomendacje dla nowego produktu w oparciu o analizę zawartości (tytuł, opis, kategorię). [6]

3.8.3. Problem rzadkości danych

W systemach rekomendacji liczba ocen już uzyskanych od użytkowników jest bardzo mała w stosunku do liczby ocen jakie należy przewidzieć. Skuteczna predykcja ocen z bardzo małej ilości danych jest bardzo ważna, zwłaszcza w metodach kolaboratywnych. Do sukcesu przyczynia się również obecność krytycznej (minimalnej) ilości użytkowników. Przykładem może być system rekomendacji książek w księgarni internetowej, gdzie bardzo wiele książek jest ocenianych przez niewielką grupę użytkowników. Będą one rekomendowane bardzo

rzadko nawet w przypadku kiedy użytkownicy oceniają je bardzo wysoko. Ponadto w przypadku specyficznych zainteresowań danego użytkownika, które nie mają odniesień w całej populacji użytkowników (słabe podobieństwa pomiędzy użytkownikami), rekomendacje prezentowane temu użytkownikowi będą słabe.

Jednym z rozwiązań tego problemu jest użycie danych demograficznych do kalkulacji podobieństw między użytkownikami. Wzięcie pod uwagę nie tylko jego ocen, ale również wieku, miejscowości, płci, wykształcenia itd. Jednak i tym razem będzie to miało charakter hybrydowej metody rekomendacji.

Innym sposobem na rzadkość danych w systemach kolaboratywnych jest redukcja wymiarowości macierzy rzadkich. Popularnie stosowaną techniką, ze względu na najlepsze przybliżenia macierzy oryginalnej (w stopniu niższym), jest dekompozycja głównych składowych, inaczej rozkład *SVD* (ang. *Singular Value Decomposition*). Metoda pozwala na zastąpienie jednej macierzy o wymiarach $m \times n$ (np. oceny produktów wystawiane przez użytkowników), dwoma o wymiarach $m \times k$ oraz $k \times n$, gdzie k jest wartością stałą. Przy metodach kolaboratywnych, gdzie złożoność obliczeń jest wykładnicza, zmniejsza to wymaganą przestrzeń zużywaną na przechowywanie danych oraz przyspiesza obliczenia w kolejnym etapie działania systemu. Metody dekompozycji są zalecane przy dużych ilościach danych. Najczęściej wykorzystywane w systemach wyliczających korelację w trybie offline. [6, 7]

3.8.4. Inne problemy

Oprócz wymienionych wyżej problemów, można spotkać się również z innymi, które często wynikają ze specyfiki prowadzonej działalności, branży czy grupy docelowej użytkowników.

3.8.4.1. Zmienność obiektów i zainteresowań użytkowników

Jednym z tego typu problemów jest zmiana zainteresowań użytkowników wraz z upływem czasu. Przyczyną zmiany zainteresowań może być sezonowy charakter sprzedaży lub zmiana preferencji użytkownika spowodowana czynnikami zewnętrznymi, występująca zazwyczaj w długich przedziałach czasowych. Przykładem może być użytkownik robiący zakupy w danej kategorii produktów, powracający np. po okresie roku do sklepu z zainteresowaniem produktami z innej kategorii. W tym przypadku dane zgromadzone we wcześniejszym etapie mogą być mało przydatne w rekomendacjach produktów dla tego użytkownika (dotyczy to przede wszystkim metod opartych na korelacji użytkowników). Rozwiązaniem takiego problemu może być zmniejszanie wagi ocen użytkowników dla danych produktów wraz z upływem czasu.

Innym problemem może być aktualizowanie lub usuwanie produktów ze sklepu internetowego. Przykładem może być nowe wydanie bardzo popularnej i często rekomendowanej książki w księgarni internetowej. Rozwiązaniem takiego problemu jest stworzenie dodatkowych asocjacji między produktami i w przypadku wygenerowania rekomendacji starego wydania popularnej książki, zarekomendowanie nowej.

Systemy rekomendacji często sprawdzają dostępność produktu już po wygenerowaniu rekomendacji, z których następnie usuwane są produkty niedostępne w sprzedaży lub proponowane zamienniki (np. nowe wydania produktów). Należy przy tym pamiętać, że niedostępność produktu w przypadku sklepu internetowego może być ograniczona czasowo.

3.8.4.2. Pozyskiwanie połączeń między elementami

Systemy kolaboratywne oparte są o oceny produktów lub stron internetowych, wystawianych przez użytkowników. Niestety w tym przypadku występuje awersja użytkowników do akcji, które w ich świadomości nie przynoszą im rzeczywistych korzyści. Podobna sytuacja ma miejsce w systemach opartych o profil demograficzny. Użytkownicy niechętnie podają swoje dane personalne, mając na uwadze chęć zachowania jak największej anonimowości. Mogą nawet podawać fałszywe dane w celu ochrony ich prywatności. W przypadku tworzenia treści w Internecie przez użytkowników można mówić o *Piramidzie cyklu tworzenia wartości*¹⁸, w której spośród 100 użytkowników danej społeczności (np. sklepu internetowego) tylko 10 aktywnie uczestniczy w tworzeniu treści (np. komentowaniu i ocenianiu produktów). Problem w pozyskiwaniu danych od użytkowników może prowadzić do problemu rzadkości danych oraz mało rzetelnych rekomendacji. Rozwiązaniem tego problemu jest automatyzacja pozyskiwania danych od użytkowników. Można np. pozyskiwać dane na podstawie sesji HTTP. W systemach kolaboratywnych z automatycznym generowaniem ocen, poszczególnym akcjom użytkownika (kliknięcie, dodanie do koszyka, zamówienie) przypisywane są różne wagi ocen. W systemach opartych o powiązania, wzorce sekwencyjne (np. targetowanie behawioralne) śledzone są ścieżki nawigacyjne. Taka automatyzacja nie wymaga od użytkownika żadnych dodatkowych akcji, poza standardowym przeglądaniem stron i korzystaniem z podstawowej funkcjonalności sklepu internetowego.

¹⁸ Kaznowski D., Nowy Marketing, Warszawa, 2008

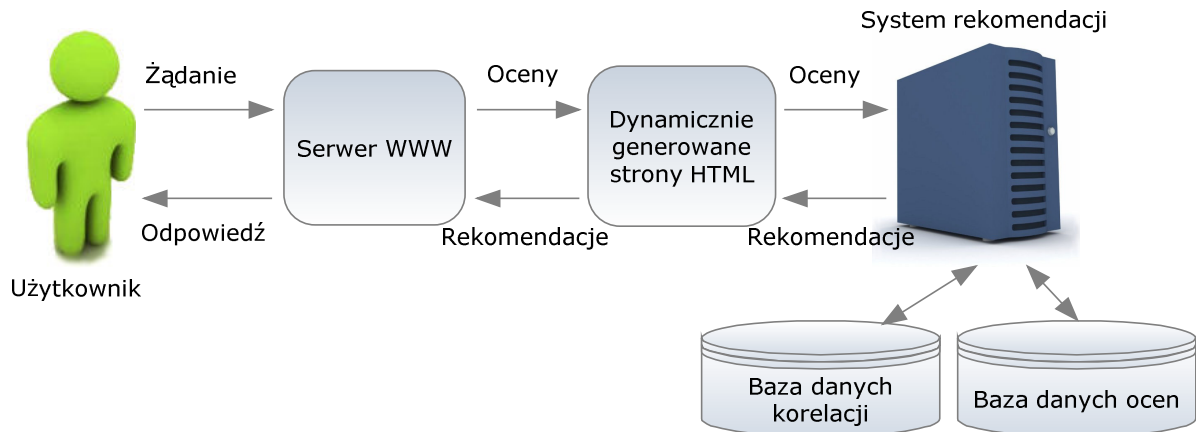
4. Proponowane rozwiązanie

Stworzenie dobrze działającego systemu rekomendacji jest czynnością złożoną i często niewystarczającą do osiągnięcia zamierzonych celów biznesowych. Aby spełnić oczekiwania biznesowe wobec takiego systemu trzeba również zadbać o wdrożenie i sposób prezentacji wyników w sklepie internetowym. Niesie to za sobą konieczność dostosowania stworzonego systemu rekomendacji do charakteru branży i działalności sklepu. W tym rozdziale przedstawiono sposób działania i funkcjonalność proponowanego systemu. Natomiast wdrożenie i dostosowanie proponowanego rozwiązania przedstawiono w kolejnym rozdziale na przykładzie wybranego sklepu internetowego.

4.1. Założenia ogólne i funkcjonalne

Ze względu na powszechność oraz bezpłatny dostęp do technologii, system rekomendacji został stworzony w języku *PHP* w wersji 5 przy wykorzystaniu bazy danych *MySQL* w wersji 5. Do wykonania obliczeń został wykorzystany, opisany w kolejnych podrozdziałach, algorytm *Slope One* wykorzystujący filtrowanie kolaboratywne.

Wysoka elastyczność systemu została osiągnięta poprzez rozdzielenie od siebie poszczególnych elementów. Na sam system rekomendacji składają się bazy danych ocen i korelacji, algorytm wyliczeń korelacji oraz interfejs wejście/wyjścia systemu rekomendacji (przyjmujący oceny, wydający rekomendacje). Dodatkowo zaimplementowana została metoda identyfikacji i asocjacji użytkowników. Elementy oceniane, skala ocen, wagi poszczególnych ocen oraz sposób ich przyznawania są ustalane na poziomie wdrożenia i są niezależne od samego systemu rekomendacji. Podobnie sam sposób przetwarzania oraz prezentacji rekomendacji jest zależny od interpretacji wdrożonej w sklepie internetowym. Ogólny schemat działania systemu przedstawia rysunek 5.



Rys. 5 Architektura systemu rekomendacji. Źródło: opracowanie własne na podstawie [7]

4.1.1. Dowolność przyznawania ocen

Oceny w systemie mogą być przyznawane w sposób automatyczny, jak również manualny (poprzez ręczne przyznanie oceny przez użytkownika). Skala ocen jest ustalana na poziomie wdrożenia. System działa prawidłowo dla dowolnej skali np. 1-5, 1-10, 0-1. System posiada zaimplementowaną metodę do przyznawania i normalizacji ocen w skali 0-1 z precyzją do czterech miejsc po przecinku. Umożliwia to dostosowanie systemu do dowolnych zadań, ocenianie dowolnych akcji użytkownika. Oceny manualne mogą być normalizowane i zapisywane w dowolnej innej skali.

4.1.2. Odrzucanie skrajnych wartości

Zbyt duża liczba ocen przyznanych przez jednego użytkownika może być nienaturalna i prowadzić do zmniejszenia wydajności algorytmu. Dodatkowo metoda predykcji dla takiego użytkownika może nie mieć praktycznego zastosowania ze względu na zbyt szerokie zainteresowania. W oparciu o metody statystyczne system zapobiega takim sytuacjom, ignorując bądź usuwając oceny użytkowników zaliczane do skrajnych wartości.

4.1.3. Personalizacja ciągła

System oparty został o personalizację ciągłą co umożliwia dostarczenie najbardziej spersonalizowanych rekomendacji. W takim podejściu konieczne jest zastosowanie systemu identyfikacji użytkowników. Proponowany system realizuje to w oparciu o pliki *cookies*. Umożliwia również identyfikację użytkowników, którzy nie są zalogowani. Po logowaniu lub rejestracji następuje procedura asocjacji przyznanych ocen z rzeczywistym identyfikatorem użytkownika. Takie podejście umożliwia również przyznawanie ocen przez użytkownika bez

potrzeby wcześniejszego uwierzytelnienia jego osoby. Daje to dodatkowe możliwości właścicielom witryn internetowych oraz nie zniechęca potencjalnych klientów.

4.1.4. Rekomendacje dla użytkowników oraz elementów

System wykorzystuje metody filtrowania kolaboratywnego oparte o automatycznie wygenerowane oceny użytkowników. Oprócz rekomendacji stron lub produktów podobnych do danego, system umożliwi rekomendacje dla danego użytkownika, które mogą być prezentowane nie tylko na karcie danego produktu, ale również w miejscach nie powiązanych z żadnym produktem (strony tekstowe, informacje o firmie itd.). Rekomendacje dla użytkowników mogą być również wykorzystane w procesach posprzedażowych np. w kampaniach mailingowych, gdzie oczekuje się dobrych rezultatów osiąganych poprzez spersonalizowaną ofertę sprzedażową.

4.1.5. Dowolność prezentacji rekomendacji

Wyjście systemu zawiera identyfikatory proponowanych produktów lub innych elementów. Można określić liczbę rekomendacji, jaką oczekujemy, że system zwróci. Umożliwia to prezentację rekomendacji w dowolnej formie, w dowolnym miejscu na stronie. Wyniki rekomendacji mogą być sortowane w dowolny sposób. Można również łączyć rekomendacje oparte o korelacje elementów z rekomendacjami opartymi o korelacje użytkowników. Pozwala to na ciągłe testowanie możliwości i udoskonalanie sposobu prezentacji rekomendacji. System może zwrócić większą listę rekomendacji, z której następnie usunięte zostaną produkty czasowo niedostępne w sprzedaży, w celu prezentacji tylko tych aktualnie dostępnych.

4.2. Algorytm Slope One

Slope One jest algorytmem filtrowania kolaboratywnego opartym o analizę ocen elementów przyznawanych przez użytkowników. Jest zaliczany do rodziny algorytmów opartych o model. Algorytm oblicza zależności między ocenami różnych użytkowników, przewidując w ten sposób ocenę danego użytkownika dla danego elementu. Jest w stanie konkurować z wolniejszymi algorytmami opartymi o heurystykę wyliczaną w pamięci komputera, prezentując przy tym lepsze rezultaty działania.

Algorytmy tego typu wykorzystują w działaniu tablicę ocen użytkownika w postaci pary: oceniony element, ocena. W odpowiedzi prezentując tablicę przewidywań w postaci:

element, przewidywana ocena dla elementów, których użytkownik jeszcze nie ocenił. Przykładowa tablica ocen została przedstawiona w tabeli 3. [9]

Tabela 3. Przykładowa macierz ocen użytkowników dla elementów, gdzie O_{ij} to ocena użytkownika i dla elementu j , a skala ocen wynosi od 0 do 5. Źródło: opracowanie własne

	Element ₁	Element ₂	...	Element _{n-1}	Element _n
Użytkownik ₁	4	3		4	2
Użytkownik ₂	5	1		0	4
...			O_{ij}		
Użytkownik _{m-1}	1	2		4	3
Użytkownik _m	0	3		5	4

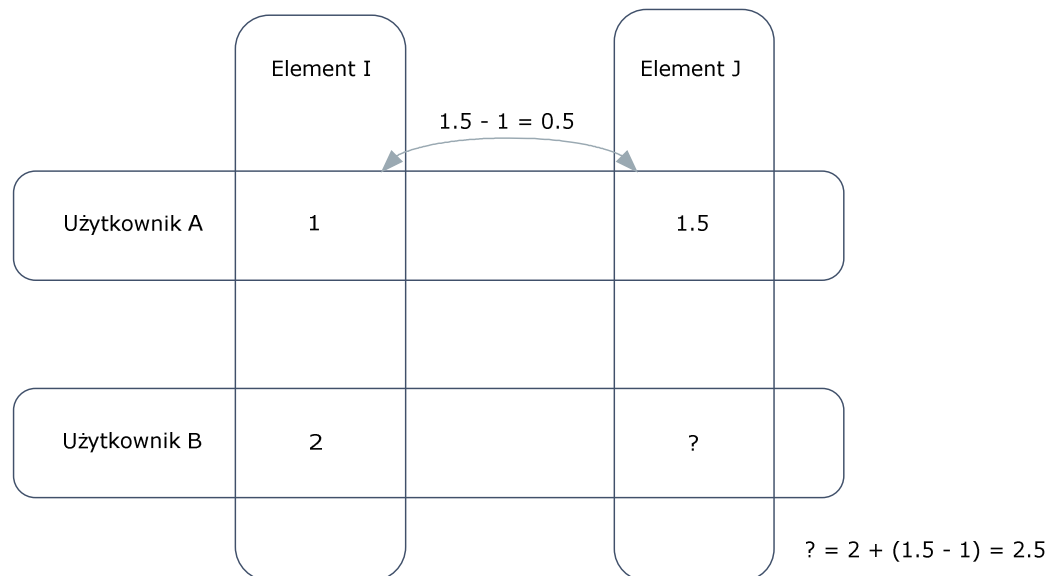
4.2.1. Zalety algorytmu

Algorytm *Slope One* jest stosunkowo łatwy do wdrożenia, efektywny w działaniu, w miarę dokładny, wspiera rekomendacje generowane w czasie rzeczywistym oraz dynamiczne aktualizowanie danych. Czyni go to dobrym kandydatem do zastosowania w rzeczywistych systemach. Do głównych zalet algorytmu można zaliczyć:

- łatwość wdrożenia i utrzymania – wszystkie zagregowane dane są łatwe w interpretacji,
- aktualizacje w locie (na bieżąco) – dodanie nowej oceny zmienia natychmiast przewidywania (rekomendacje),
- szybkie odpowiedzi na zapytania – algorytm jest efektywny m.in. dzięki zoptymalizowanej bazie danych,
- dobre działanie dla nowych użytkowników – osoby z kilkoma ocenionymi elementami mogą spodziewać się kilku konkretnych rekomendacji,
- dokładność – algorytm przy zachowaniu prostoty i skalowalności jest w stanie konkurować na tym polu z innymi. [9]

4.2.2. Działanie algorytmu

Slope One jest algorytmem filtrowania kolaboratywnego predykującym oceny użytkowników dla elementów w oparciu o prostą regresję liniową, w postaci funkcji $f(x)=x+b$. Wykorzystuje przy tym obliczone wcześniej średnie odchylenia pomiędzy ocenami użytkowników dla elementów. Występująca w funkcji $f(x)$ stała b jest średnią różnicą między ocenami innych użytkowników, natomiast x to zmienna reprezentująca wartość oceny danego użytkownika. Dla każdej pary elementów poszukiwana jest najlepsza funkcja f , która przewiduje ocenę elementu, na podstawie ocen innych elementów. Funkcja ta może być różna dla każdej pary elementów. Funkcja, dla danych najprostszej postaci, została przedstawiona na rysunku 6. W przedstawionym przypadku użytkownik A ocenił element I na 1, a użytkownik B na 2. Użytkownik A ocenił również element J na 1.5, natomiast użytkownik B nie ocenił tego elementu. Można postarać się o przewidzenie oceny użytkownika B dla elementu J . J jest oceniany wyżej od pozycji I o $1.5-1=0.5$, stąd można wywnioskować, że użytkownik B oceni pozycję J na $2+0.5=2.5$.



Rys. 6 Podstawa algorytmu *Slope One*. Źródło: opracowanie własne na podstawie [9]

Algorytm *Slope One* uwzględnia zarówno informacje od innych użytkowników oceniających ten sam element, jak również informacje od innych elementów ocenianych przez tego użytkownika. Dlatego wyróżnia się dwa kroki w obliczeniach rekomendacji.

W pierwszym kroku obliczane jest średnie odchylenie dwóch elementów. Dla każdego dwóch elementów E_i i E_j , algorytm oblicza średnie odchylenie elementu E_i w odniesieniu do E_j , według wzoru:

$$D_{ij} = \sum_{u_c \in U(E_i \cap E_j)} \frac{O_{ci} - O_{cj}}{\text{count}(E_i \cap E_j)} \quad (1)$$

gdzie:

D_{ij} – średnie odchylenie dla elementu i oraz elementu j ,

O_{ci} – ocena użytkownika c dla elementu i ,

O_{cj} – ocena użytkownika c dla elementu j ,

E_i – element i ,

E_j – element j ,

$\text{count}()$ – liczba wystąpień.

W drugim kroku algorytm oblicza przewidywania. Uwzględnia przy tym wagi będące liczbą wystąpień użytkowników oceniających obydwie elementy, dla których policzone zostało średnie odchylenie. Dzięki temu ocena elementu, który jest popularniejszy (częściej oceniany) ma większe znaczenie od elementu mniej popularnego. Na podstawie dostępnej macierzy średnich odchyleń, dla użytkownika t oraz docelowego elementu i przewidywanie określone jest wzorem:

$$P_{ti} = \frac{\sum_{c \in S(u_t) - \{i\}} (O_{tc} + D_{ic}) \cdot w_{ic}}{\sum_{c \in S(u_t) - \{i\}} w_{ic}} \quad (2)$$

gdzie:

P_{ti} – przewidywanie dla użytkownika t oraz elementu i ,

O_{tc} – ocena użytkownika t dla elementu c ,

D_{ic} – średnie odchylenie dla elementu i oraz elementu c ,

$S(u_t)$ – podzbiór wszystkich elementów ocenionych przez użytkownika t ,

w_{ic} – liczba użytkowników, którzy ocenili element i oraz element c .

Wykorzystując powyższe macierze można obliczyć wektor będący listą rekomendowanych produktów. Rekomendacje elementów dla elementu E_i to wszystkie elementy, dla których wyliczono średnie odchylenie. Im średnie odchylenie jest mniejsze, tym element lepiej pasuje (lepsza rekomendacja) do elementu E_i . Rekomendacje elementów dla użytkownika (rekomendacje spersonalizowane) to wszystkie elementy, dla których obliczono przewidywaną ocenę P_{ti} elementu E_i dla użytkownika t . Im przewidywana ocena jest wyższa,

tym bardziej prawdopodobne jest pozytywne zainteresowanie użytkownika t danym elementem (przy założeniu, że wyższe oceny nadawane przez użytkowników odpowiadają większemu zainteresowaniu). [1, 9, 10]

4.2.3. Przykład obliczeń

W tabeli 4 zaprezentowano przykład macierzy ocen.

Tabela 4. Przykładowa baza danych ocen. Źródło: opracowanie własne

Użytkownik	Element 1	Element 2	Element 3
Tomek	5	3	4
Marek	2	3	Nie oceniono
Ania	Nie oceniono	4	1

Na podstawie przedstawionej tabeli ocen (tabela 4), można policzyć średnie odchylenie pomiędzy elementami $D_{12}=[(5-3)+(2-3)]/2=(2-1)/2=0.5$, co za tym idzie średnio *Element 1* jest wyżej oceniany od *Elementu 2* o 0.5. Analogicznie średnie odchylenie pomiędzy elementami 1 i 3 wynosi $D_{13}=(5-4)/1=1$. Stąd przewidywana ocena użytkownika Ania dla *Elementu 1*, używając jej oceny dla elementu 2, wynosi $4+0.5=4.5$. Analogicznie wykorzystano ocenę *Elementu 3*, $1+1=2$. Jeśli użytkownik oceni wiele elementów, przewidywania są średnią ważoną, gdzie brana jest pod uwagę liczba użytkowników oceniająca obydwie elementy (ten, dla którego przewiduje się ocenę oraz pozostałe). Z powyższych obliczeń można przewidzieć ocenę użytkownika Ania dla *Elementu 1*, w następujący sposób:

$$P_{AI} = \frac{4.5 \cdot 2 + 2 \cdot 1}{2 + 1} = \frac{11}{3} = 3.67$$

4.2.4. Złożoność algorytmu

Algorytm oblicza i przechowuje średnie odchylenie dla maksymalnie n^2 par elementów, gdzie n jest liczbą elementów przy założeniu, że wszyscy użytkownicy ocenili wszystkie elementy. Dla przypadku n elementów, m użytkowników i N ocen, obliczenie średniego odchylenia ocen dla każdej pary elementów może dochodzić do $n(n-1)/2$ wierszy w bazie danych oraz mn^2 kroków obliczeń. Jeśli założy się, że użytkownicy ocenili maksymalnie do y elementów, wtedy możliwe jest obliczenie średnich odchyleń w nie więcej niż $n^2 + my^2$ kroków. Jeśli użytkownik oceni x elementów, predykcja pojedynczej oceny wymaga x kroków, a predykcja

wszystkich brakujących ocen użytkownika to maksymalnie $(n-x)x$ kroków. Aktualizacja bazy danych w chwili, gdy użytkownik ocenił x elementów i ocenia kolejny, wymaga x kroków. Aby zredukować wielkość bazy danych można wykorzystać sztucznie stworzoną rzadkość danych: pary elementów, które nie mają ocen lub mają ich niewiele mogą być pominięte bez znaczącego wpływu na jakość predykcji.¹⁹ [9, 10]

4.2.5. Testy empiryczne algorytmu

Wykorzystany algorytm *Slope One* (ważony) został porównany do innych, znanych algorytmów filtrowania kolaboratywnego, wykorzystywanych w systemach rekomendacji produktów. Do testów wykorzystane zostały bazy danych ocen filmów *EachMovie* oraz *MovieLens* udostępnione przez *GroupLens Research Group* z Uniwersytetu Minnesota²⁰. Bazy danych posiadały 50 000 ocen filmów do procesu nauczania (budowania modelu) oraz 100 000 ocen filmów do testów. W celu zachowania kompatybilności oceny oraz predykcje zostały znormalizowane do skali 0 – 1. Zbadano średni błąd *MAE* (ang. *Mean Average Error*), będący średnim odchyleniem pomiędzy przewidywaną oceną użytkownika dla danego elementu, a oceną rzeczywistą użytkownika dla tego elementu. Predykcja odbywała się z pominięciem rzeczywistej oceny danego użytkownika dla danego elementu.

Średni błąd wyrażony jest następującym wzorem:

$$MAE = \frac{1}{count(X')} \sum_{u \in X'} \frac{1}{count(S(u))} \sum_{i \in S(u)} |P(u^{(i)}) - u_i| \quad (3)$$

gdzie:

X' – zbiór testowy,

$count()$ – liczba wystąpień,

$S(u)$ – podzbiór wszystkich elementów ocenionych przez użytkownika,

$P(u^{(i)})$ – predykcja oceny elementu i dla użytkownika,

u_i – ocena elementu i użytkownika.

Na potrzeby testów wybrano algorytmy wykorzystujące różne metody predykcji. Wyniki testów prezentuje tabela 5.

¹⁹ http://en.wikipedia.org/wiki/Slope_One, opis algorytmu, 08.2010

²⁰ <http://www.grouplens.org>, GroupLens Research Group, 08.2010

Średnia ocena użytkownika jest najprostszym algorytmem predykcji polegającym na obliczeniu średniej arytmetycznej ze wszystkich ocen użytkownika. Ze względu na małą przydatność, rzadko jest stosowany, samodzielnie, w rzeczywistych systemach. Często występuje jako element innych algorytmów.

Bias From Mean to algorytm dodający do średniej arytmetycznej użytkownika, średnie odchylenie pomiędzy oceną innych użytkowników dla danego elementu, a ich średnią arytmetyczną. Jest to już lepsze rozwiązanie od samej średniej arytmetycznej, biorące pod uwagę oceny innych użytkowników.

Pearson jest znanym algorytmem opartym o obliczenia w pamięci. Sumuje średnią arytmetyczną ocen użytkownika z ważonymi sumami wszystkich użytkowników, gdzie wagą jest stopień korelacji pomiędzy danymi użytkownikami obliczony na podstawie metryki *Pearson'a*. Dzięki temu otrzymujemy dosyć dokładne przewidywania. Algorytm jest jednak mało wydajny w przypadku dużych ilości danych.

Na potrzeby testów wykorzystano również algorytm korelacji elementów, oparty o podobieństwa elementów liczone metryką cosinusa. Algorytm został dostosowany do predykcji ocen dla użytkowników poprzez wykorzystanie regresji liniowej.

Tabela 5. Porównanie algorytmów, mniejsza wartość *MAE* to lepszy wynik. Źródło: [9]

Algorytm	EachMovie	MovieLens
Slope One (ważony)	0.198	0.188
Bias From Mean	0.203	0.191
Korelacje elementów (cosinus)	0.209	0.198
Średnia ocena użytkownika	0.231	0.208
Pearson	0.194	0.190

Wyniki przedstawione w tabeli 5 są zadawalające dla algorytmu *Slope One*. Pomimo prostoty, nazywanej przez autorów predykcją naiwną, osiąga lepsze rezultaty od kilku przetestowanych algorytmów oraz porównywalne z algorytmem *Pearsona*. Cechuje się również większą wydajnością niż wiele znanych algorytmów, zwłaszcza tych opartych o heurystyki, np. algorytm *Pearsona*. Na tej podstawie można wywnioskować, że jest to dobry kandydat do zastosowania w rzeczywistych systemach rekomendacji. [9]

4.3. Struktura bazy danych

Proponowane rozwiązanie wykorzystuje darmowy silnik *Open Slope One*²¹, oparty o algorytm *Slope One*. Rozwiązanie to jest bardzo wydajne ze względu na przeniesienie prawie wszystkich obliczeń do bazy danych *MySQL*, łącznie z wykorzystaniem procedur składowych do wyliczenia tabeli pośredniej. Na samo rozwiązanie *Open Slope One* składają się dwie tabele: *user_ratings* (tabela 6) oraz *slope_one* (tabela 7).

Tabela 6. Struktura tabeli *user_ratings*. Źródło: opracowanie własne

user_ratings	
Nazwa	Typ
user_id	Int(11)
item_id	Int(11)
rating	Decimal(14,4)

Tabela *user_ratings* (tabela 6) przechowuje oceny użytkowników dla danego elementu. Przy wykorzystaniu dodatkowej metody zaimplementowanej na potrzeby proponowanego rozwiązania, oceny mogą być znormalizowane do skali 0-1 z czterema miejscami po przecinku. Dodatkowo zaimplementowana metoda nie tworzy nowych rekordów dla nowych lub dodatkowych ocen użytkownika dla tego samego elementu. Takie oceny są sumowane ze starymi do maksymalnej wartości wynoszącej 1. W wielu przypadkach (zależnych od wdrożenia) pozwala to zaoszczędzić przestrzeń bazy danych i przyspieszyć obliczenia predykcji.

²¹ <http://code.google.com/p/openslopeone/>, silnik rekomendacji Open Slope One, 08.2010

Tabela 7. Struktura tabeli *slope_one*. Źródło: opracowanie własne

slope_one	
Nazwa	Typ
item_id1	Int(11)
item_id2	Int(11)
times	Int(11)
rating	Decimal(14,4)

Tabela *slope_one* (tabela 7) przechowuje dane pozwalające na wyliczenie średniego odchylenia pomiędzy ocenami dwóch elementów. Kolumna *times* to liczba wystąpień tych samych par elementów w ocenach jednego użytkownika, natomiast *rating* to suma różnic pomiędzy tymi ocenami. Średnie odchylenie pomiędzy dwoma elementami jest więc ilorazem *rating* przez *times*.

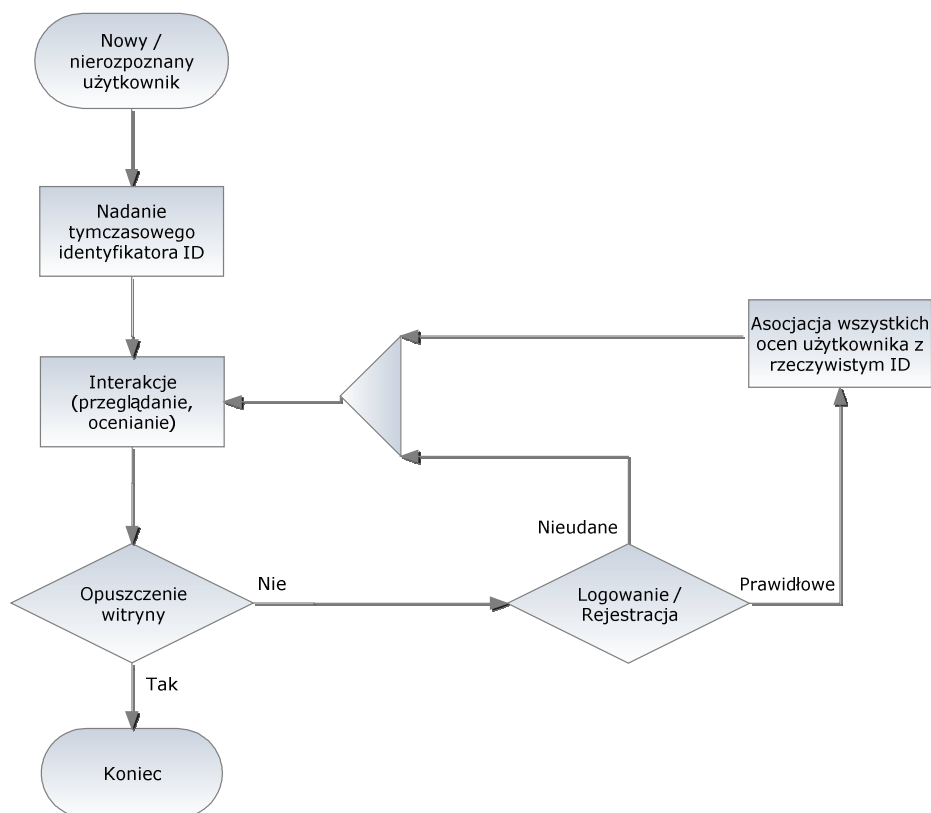
Tabela 8. Struktura tabeli *user_tmp*. Źródło: opracowanie własne

user_tmp	
Nazwa	Typ
user_id_tmp	Int(11)
datetime	Datetime

Tabela *user_tmp* (tabela 8) jest dodatkową tabelą stworzoną na potrzeby proponowanego rozwiązania. Przechowuje ona tymczasowe identyfikatory użytkowników, pozwalając na ocenianie elementów bez potrzeby uwierzytelniania. Kolumna *datetime*, to czas ostatniego dostępu danego użytkownika do systemu. Dzięki temu, stare niewykorzystywane identyfikatory mogą być sukcesywnie usuwane. Również pozwala to na usuwanie ocen użytkowników tymczasowych, którzy nie logując się do serwisu mogą zostać uznani za przypadkowych. Oceny takich osób mogą mieć niską wartość dla prawidłowej predykcji w systemie rekomendacji.

4.4. Identyfikacja użytkowników

Personalizacja ciągła zastosowana w proponowanym systemie rekomendacji działa dla wszystkich użytkowników, również bez uwierzytelnienia. Takie podejście umożliwia automatyczne ocenianie np. odsłon stron (wymaga to jednak dodatkowych metod filtracji danych). Sam system identyfikacji użytkowników wykorzystuje tabelę *user_tmp* (tabela 8) do przechowywania tymczasowego identyfikatora użytkownika nadawanego podczas pierwszej wizyty na stronie. Tymczasowy identyfikator jest zawsze większy od 1 000 000 000 (do ustawienia podczas wdrożenia), dzięki czemu można go łatwo rozpoznać, gdyż jest on wykorzystywany tak samo jak rzeczywisty identyfikator w procesie przyznawania ocen (ta sama tabela danych). W przypadku logowania lub rejestracji następuje asocjacja tymczasowego identyfikatora do rzeczywistego identyfikatora nadawanego przez system strony internetowej, łącznie z asocjacją nadanych ocen. Identyfikator przechowywany jest po stronie klienta w pliku *cookie*, przez co może być wykorzystany przy kolejnej wizycie. W bazie danych przechowywana jest data ostatniego dostępu do strony przez „tymczasowego użytkownika”, jeśli w określonym czasie po ostatniej wizycie nie nastąpi asocjacja z realnym identyfikatorem, tymczasowy identyfikator zostanie usunięty wraz z przypisanymi do niego ocenami. Schemat działania podsystemu identyfikacji użytkowników przedstawia rysunek 7.



Rys. 7 Schemat identyfikacji użytkowników. Źródło: opracowanie własne

4.5. Działanie proponowanego systemu rekomendacji

Proponowany system, w przeciwieństwie do systemów opartych o obliczenia w pamięci, daje prawidłowe rekomendacje już przy kilku ocenach, kilku użytkowników. Może być wykorzystywany do rekomendowania produktów podobnych do danego jak również rekomendowania produktów użytkownikom. Jest elastyczny we wdrożeniu, co pozwala dostosować go do realiów biznesowych wielu stron internetowych.

4.5.1. Obliczenia i aktualizacje danych

Zastosowanie bazy danych korelacji pomiędzy elementami (tabela pośrednia przedstawiona w tabeli 6) pozwala na wcześniejsze wyliczenie średnich różnic między wszystkimi ocenionymi elementami. Dla dużych ilości danych, użytkowników oraz ocen, zgromadzonych w odpowiednio długim czasie, podobieństwa pomiędzy produktami zmieniają się bardzo powoli. Oznacza to, że wraz ze wzrostem ilości ocen można rzadziej wyliczać tabelę korelacji między produktami (np. docelowo raz na dobę w czasie mniejszego obciążenia systemu). Proponowane rozwiązanie oprócz bazy danych korelacji pomiędzy produktami, wykorzystuje bazę danych ocen użytkowników, w przypadku generowania rekomendacji dla konkretnego użytkownika. Pozwala to na bieżącą aktualizację ocen użytkownika, która docelowo ma wpływ na generowane dla niego rekomendacje, pomimo ewentualnej niezmienności bazy danych korelacji pomiędzy produktami. Takie podejście jest efektywne, ponieważ korelacje pomiędzy użytkownikami zmieniają się znacznie częściej niż korelacje pomiędzy elementami. Jednocześnie zapewnia to rekomendacje w czasie rzeczywistym przy minimalnym obciążeniu obliczeniowym.

Zarówno rekomendacje dla użytkownika, jak również dla elementu są wyliczane w czasie rzeczywistym, w odpowiedzi na żądanie strony internetowej. Rekomendacje dla elementu w proponowanym rozwiązaniu opierają się jednak wyłącznie na wcześniej wyliczonej tabeli korelacji między elementami (pozwalającej na szybkie obliczenie średnich odchyłeń). Natomiast rekomendacje dla użytkownika opierają się zarówno na wcześniej wyliczonej tabeli korelacji pomiędzy elementami, jak i na tabeli ocen użytkowników aktualizowanej na bieżąco. [1, 9]

4.5.2. Usuwanie anomalii

Ze względu na możliwość generowania bardzo dużej ilości ocen oraz możliwość przyznawania ocen automatycznych nawet dla użytkowników nie uwierzytelnionych, powstaje potrzeba zastosowania filtru, który w sposób skuteczny ograniczy występowanie

anomalii, jednocześnie optymalizując cały system generowania rekomendacji. Zaimplementowany podsystem redukcji anomalii ogranicza liczbę ocen nadawaną przez pojedynczego użytkownika, usuwając oceny użytkowników, których liczba wzbudza podejrzenia. Zbyt duża liczba ocen nadana przez jednego użytkownika może nie być przydatna w procesie generowania rekomendacji, ponieważ wskazuje na bardzo szerokie zainteresowania danej osoby lub anomalie nieznanego pochodzenia. System podsumowuje liczbę przyznanych ocen przez jednego użytkownika, a następnie wykorzystując *nierówność Czebyszewa* dla dowolnych rozkładów usuwa maksymalnie do 0,1% ocen użytkowników, których liczba jest zbyt duża w stosunku do innych. Wykorzystano przy tym następujący wzór (*nierówność Czebyszewa*):

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2} \quad (4)$$

gdzie:

P – prawdopodobieństwo zajścia zdarzenia,

X – zmienna losowa (liczba ocen użytkownika),

μ – średnia arytmetyczna (średnia liczba ocen na użytkownika),

c – liczba większa od 1,

σ – odchylenie standardowe (liczby ocen).

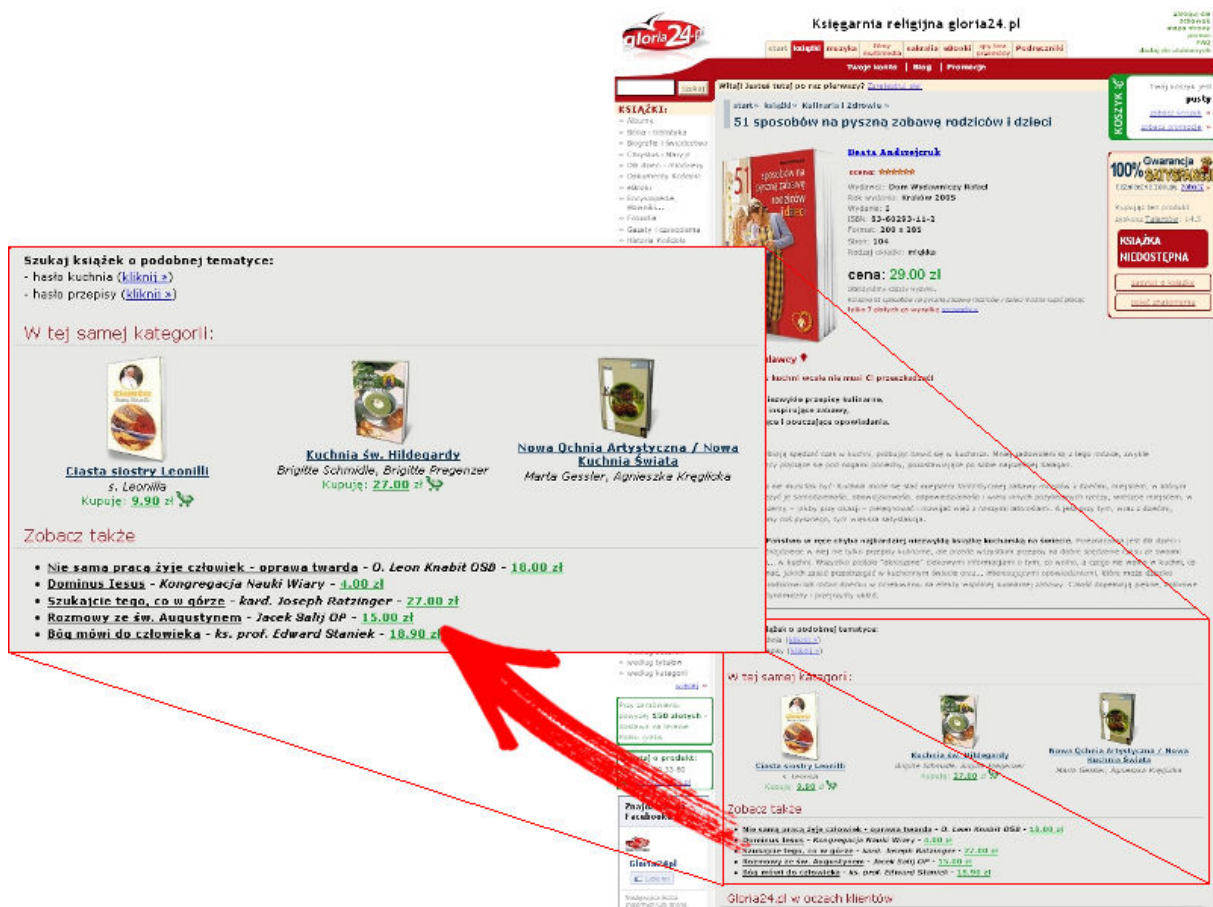
W przypadku proponowanego rozwiązania $c=32$, usuwane są wszystkie oceny użytkowników, których liczba jest większa od $\mu+32\sigma$. Prawdopodobieństwo takiego zajścia (maksymalny odsetek usuniętych użytkowników wraz z ich ocenami) wynosi:

$$P(|X - \mu| \geq 32\sigma) \leq \frac{1}{32^2} = 0.000977$$

Opisany podsystem jest uruchamiany bezpośrednio przed każdym generowaniem bazy danych korelacji pomiędzy ocenianymi elementami. [11]

5. Wdrożenie rozwiązania w sklepie internetowym

Zaproponowane rozwiązanie zostało wdrożone i przetestowane w warunkach rzeczywistych. Sklep internetowy, w którym rozwiązanie zostało wprowadzone jest księgarnią internetową, wyspecjalizowaną w sprzedaży przedmiotów o charakterze religijnym, w tym książek, filmów, muzyki, gier, przedmiotów sakralnych. Sklep posiada ponad 8 tysięcy artykułów, w tym ponad 7 tysięcy w aktywnej sprzedaży. Kilkadziesiąt tysięcy zarejestrowanych użytkowników. Jest odwiedzany każdego dnia przez kilka tysięcy unikalnych użytkowników. Ze względu na skalę działalności testy zostały ograniczone do działu książek. Rekomendacje były prezentowane tylko na karcie produktu, w postaci pięciu linków prowadzących do rekomendowanych produktów. Rysunek 8 prezentuje wygląd karty produktu w sklepie, wraz z rekomendacjami w sekcji *Zobacz także*.



Rys 8. Zrzut ekranu prezentujący wdrożone rozwiązanie. Źródło: <http://gloria24.pl>, księgarnia internetowa, 08.2010

Celem dla sklepu był wzrost przychodów poprzez poprawę istniejących lub wdrożenie nowych rozwiązań rekomendacji. Analiza historycznych danych wykazała również związek pomiędzy średnią wartością zamówienia a przychodami. Dlatego bardzo ważne było zwiększenie średniej wartości zamówienia. Okazało się bowiem, że 70% zamówień, które przyniosło sklepowi 32% zysków jest poniżej średniej, natomiast 30% zamówień, które były powyżej średniej przyniosło 68% przychodów. Interesujące okazało się również, że ponad 16% zamawiających ponawia zakupy, a 1/3 z nich robi to w ciągu miesiąca od pierwszego zakupu. Takie informacje mogą mieć wpływ na politykę posprzedażową sklepu, a odpowiedni system rekomendacji produktów może być pomocny w polepszeniu tych współczynników.

5.1. Dane początkowe i dobór ocen

W celu dobrania odpowiednich ocen przeprowadzona została analiza koszykowa historycznych danych (ostatnie 3 lata działalności). Na podstawie historii zamówień, zawartości zamówień oraz dostępnych statystyk oglądalności ustalono oceniane automatycznie akcje oraz ich wartość liczbową w postaci oceny dla danej akcji. Oceny akcji odbywały się przy wykorzystaniu zaimplementowanej metody normalizacji ocen, w skali 0 – 1, z precyzją czterech miejsc po przecinku. W tabeli 9 przedstawiono oceniane akcje oraz przyznawane im wartości.

Tabela 9. Wartości ocen przyznawane poszczególnym akcjom. Źródło: opracowanie własne

Akcja	Wartość
Odsłona strony produktu	0.0001
Dodanie produktu do schowka	0.0050
Dodanie produktu do koszyka	0.0100
Zamówienie produktu	0.1000

Przyjęte oceny oznaczają, że średnio na 1 dodany produkt do koszyka może przypadać 100 odsłon tego produktu. Dodanie do koszyka w większości przypadków, nie kończy się złożeniem zamówienia. Szanse na zakup produktu dodanego do schowka (odpowiednik przechowalni interesujących pozycji) jest około dwukrotnie mniejszy niż w przypadku dodania do koszyka. Maksymalna ocena wynosząca 1 dla danego produktu może oznaczać, że został on zakupiony przez użytkownika 10 razy, w różnych zamówieniach.

Dane z historii zamówień zostały wykorzystane jako dane początkowe do wdrożonego systemu rekomendacji. Każdy zamówiony produkt przez danego użytkownika dostał ocenę 0.1. Maksymalna ocena dla produktu wyniosła 1. Jednak ze względu na dużą rzadkość danych (prawie 49% zamówień to zamówienia pojedynczego produktu) postanowiono przyspieszyć proces gromadzenia danych, poprzez ocenianie odsłon stron poszczególnych produktów. Niesie to za sobą zwiększone ryzyko gromadzenia fałszywych danych i występowania większych anomalii. Dlatego istotnym elementem systemu rekomendacji okazał się mechanizm usuwania anomalii.

Podsystem obliczający tabelę pośrednią podobieństw między ocenianymi produktami, był uruchamiany co 20 minut. Z czasem częstotliwość uruchomień mogła zostać zmniejszona, ze względu na minimalne zmiany w generowanej tabeli w krótkim przedziale czasu. Jednak czas ten nie został zmieniony, ze względu na dużą wydajność i brak znaczącego obciążenia dla serwera. Tabela pośrednia, z jednym milionem wierszy generowana była w ciągu około 8 sekund.

5.2. Stosowane metody rekomendacji

Opisywany sklep internetowy wykorzystuje wiele, różnych mechanizmów rekomendacji produktów. Na stronie głównej można zobaczyć sekcje: *Nowości*, *Wyprzedaże*, *Bestsellery*. Sklep wykorzystuje mechanizmy analizy koszykowej historycznych danych prezentując rekomendacje w postaci produktów często kupowanych z aktualnie przeglądanych. Na rysunku 8 pokazane zostały jeszcze dwie inne metody stosowane przez sklep. Sekcja *W tej samej kategorii*, prezentuje losowe produkty z kategorii, do której należy przeglądany produkt. Oprócz tego można zauważyć sekcję *Szukaj książek o podobnej tematyce*, gdzie prezentowane są linki z określoną frazą dla dostępnej na stronie wyszukiwarki. Każdy produkt może być skomentowany przez uwierzytelnionego użytkownika w postaci opinii oraz przyznanej oceny w skali 1-6.

Proponowane rozwiązanie zostało przetestowane jako niezależnie działający system, który nie został zintegrowany z pozostałymi mechanizmami rekomendacji. Jednym z powodów była duża rzadkość danych w stosowanych metodach, co uniemożliwiło skuteczne wykorzystanie tych informacji w proponowanym rozwiązaniu.

5.3. Metodyka testów

Ze względu na skalę działalności testy zostały ograniczone do działu książek. Rekomendacje były wyświetlane na karcie produktów. W celu uzyskania relatywnie porównywalnych wyników z istniejącymi systemami testy były wyświetlane w sposób równoległy (test A/B) z istniejącym rozwiązaniem. W ramach testów sekcja *Kupujący ten produkt wybrali również*, na karcie produktu została zastąpiona sekcją *Zobacz także*. Rekomendacje z proponowanego systemu były wyświetlane w tej sekcji. Równoległy system wyświetlał rekomendacje na podstawie prostej analizy koszykowej (dotychczasowy *Kupujący ten produkt wybrali również*) w tej samej sekcji.

Główną różnicą w obydwu systemach była wewnętrzna metoda rekomendacji. Dotychczasowy system oparty o prostą analizę koszykową wyświetlał produkty kupowane z danym. Oznacza to, że rekomendacje oparte były na aktualnie wyświetlanym produkcie. Tak więc wszystkie produkty posiadały inne rekomendacje. W przypadku proponowanego rozwiązania rekomendacje wyświetlane były personalnie, na podstawie podobieństw zainteresowań użytkowników. Oznacza to, że na wszystkich kartach produktów rekomendacje były podobne, a często nawet identyczne. Przyczyniło się to do zwiększenia ilości odsłon danej rekomendacji. Zaimplementowana w proponowanym rozwiązaniu metoda obliczania rekomendacji dla danego produktu wyświetlała rekomendacje w przypadku braku rekomendacji dla użytkownika (brak jakichkolwiek ocen). Nie została wykorzystana w sposób bezpośredni, jako samodzielny system, w czasie testów, choć potencjalnie mogła dawać lepsze rekomendacje od dotychczasowego rozwiązania. Powodem niewykorzystania metody był ograniczony czas testów oraz chęć porównania zupełnie różnych systemów rekomendacji (rekomendacje dla danego produktu ze spersonalizowanymi rekomendacjami dla użytkownika).

Zbieranie danych do analizy wyników wymagało zaimplementowania dodatkowych mechanizmów. Użytkownik wchodząc na stronę po raz pierwszy był w sposób losowy przypisywany do grupy, której wyświetlają się rekomendacje z dotychczasowego systemu lub do grupy, której wyświetlają się rekomendacje z nowego, proponowanego systemu. Takie przypisanie miało charakter permanentny w czasie trwania testów. Użytkownik przy kolejnej wizycie był rozpoznawany jako należący do konkretnej grupy. Informacje te były przechowywane w pliku *cookie* użytkownika oraz w bazie danych po stronie serwera wraz z innymi danymi. Pozostałe dane dotyczyły odsłon oraz zamówień, występujących w poszczególnych systemach rekomendacji. Przy czym złożenie zamówienia przez osoby

przypisane do danego systemu rekomendacji, niekoniecznie oznaczało zainteresowanie tych osób samymi rekomendacjami.

Zastosowana metodyka w postaci testów równoległych pozwoliła na uzyskanie precyzyjnych informacji, służących porównaniu obu systemów. Jednocześnie, praktycznie zniwelowane zostały inne czynniki wpływające na wielkość przychodów, takie jak: promocje, czy przede wszystkim sezonowość sprzedaży.

5.4. Uzyskane wyniki

Uzyskane wyniki prezentują różnice pomiędzy zastosowanymi systemami rekomendacji. Testy porównawcze trwały 35 dni, w tym 4 pełne tygodnie z rzędu, od 16 lipca 2010 roku do 19 sierpnia 2010 roku. Proponowany system był wykorzystywany do prezentacji rekomendacji, w stosunku liczby użytkowników 49,89% do 50,11% z dotychczasowym systemem. Przy czym odnotowano niewielki wzrost liczby odsłon rekomendacji na użytkownika w nowym systemie.

5.4.1. Przygotowanie danych do analizy

Zebrane dane statystyczne zostały porównane i przypisane do danych pochodzących z systemu zamówień sklepu. W zestawieniach uwzględnione są wyłącznie dane pochodzące ze zrealizowanych zamówień. Dodatkowo dane te zostały przefiltrowane pod kątem występowania przypadkowych zamówień, które mogłyby mieć duży wpływ na uzyskane wyniki. Do filtracji posłużyła opisana w rozdziale 4.5.2 *nierówność Czebyszewa* dla dowolnych rozkładów. Usunięte zostały zamówienia, których wartość przekraczała $\mu+3\sigma$ (średnia + 3 razy odchylenie standardowe). Co według teorii nie powinno stanowić więcej niż 11,1% wszystkich zamówień. W praktyce z zestawienia usuniętych zostało 0,93% wszystkich zamówień. Wśród usuniętych zamówień 62,5% przychodów przypadło na proponowany system rekomendacji. Co może oznaczać, że proponowane rozwiązanie w rzeczywistości mogło przynieść lepsze wyniki niż te, które zostały przedstawione w tym rozdziale.

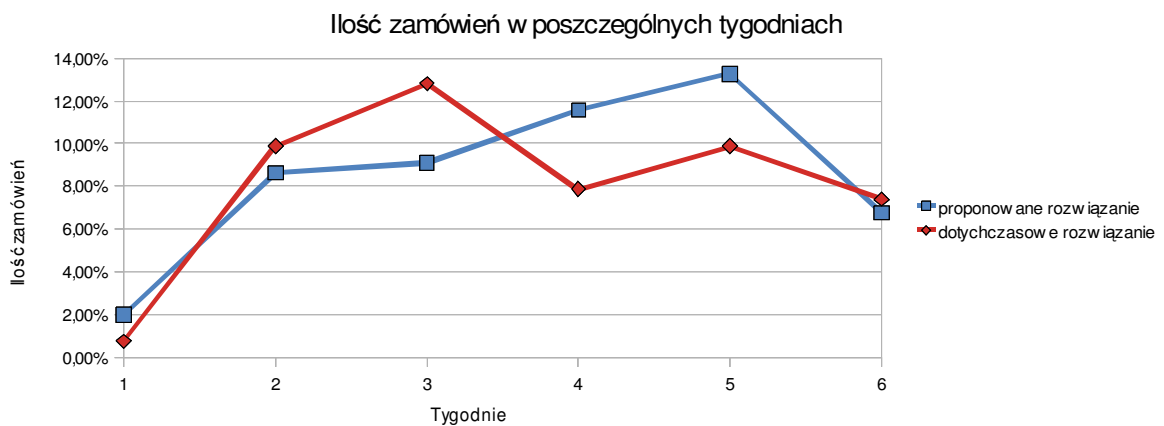
W tabeli 10 przedstawiono podsumowanie testu. Prezentowane w tym rozdziale wyniki zostały znormalizowane.

5.4.2. Skuteczność systemów

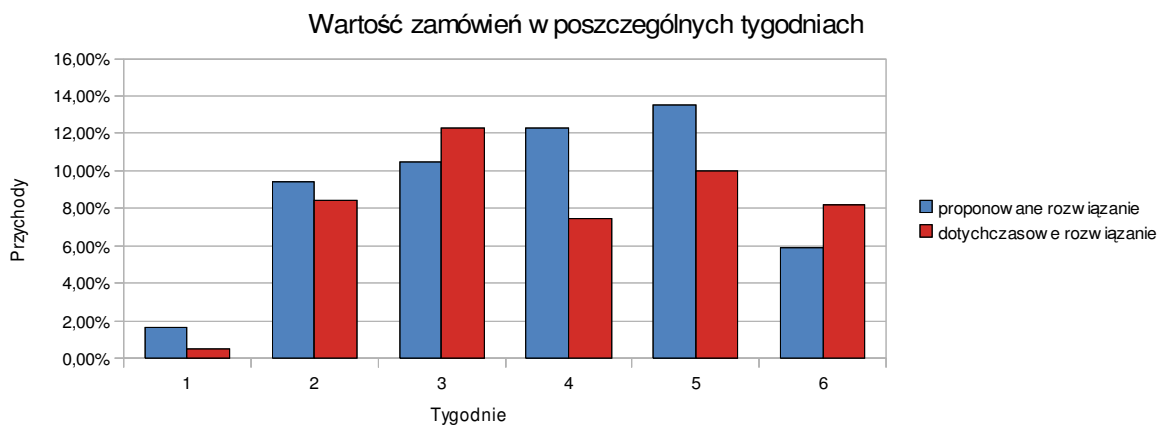
Proponowane rozwiązanie przyniosło wyższe przychody dzienne w 25 dniach na 35 dni testów. W przypadku 23 dni z 35 dni testów wyższa była średnia wartość pojedynczego zamówienia. Ilość zamówień była wyższa w 18 dniach, równa w 5 dniach i niższa w 12

dniach z 35 dni testów. W 20-stu z 35 dni wyższa była liczba różnych produktów w pojedynczym zamówieniu.

Na rysunku 9 została pokazana ilość zamówień w poszczególnych tygodniach. Natomiast na rysunku 10 przychody z tych zamówień. Na przedstawionych wykresach widać, że proponowany system zyskiwał skuteczność wraz z upływem czasu. Punkt krytyczny nastąpił pomiędzy 3 i 4 tygodniem testów. Po tym okresie system zyskał zauważalną przewagę zarówno pod kątem ilości zamówień jak i przychodów. Pierwszy tydzień przedstawia dane z końcowych dni tygodnia, natomiast ostatni z początkowych dni tygodnia. W podrozdziale 5.4.3 pokazano, że wdrożony system jest skuteczniejszy w drugiej połowie tygodnia. Dlatego skrajne tygodnie w tym przypadku nie prezentują dokładnych danych i można je pominąć. Pełne tygodnie na przedstawionych wykresach to tygodnie od 2 do 5.



Rys 9. Ilość zamówień w poszczególnych tygodniach. Źródło: opracowanie własne



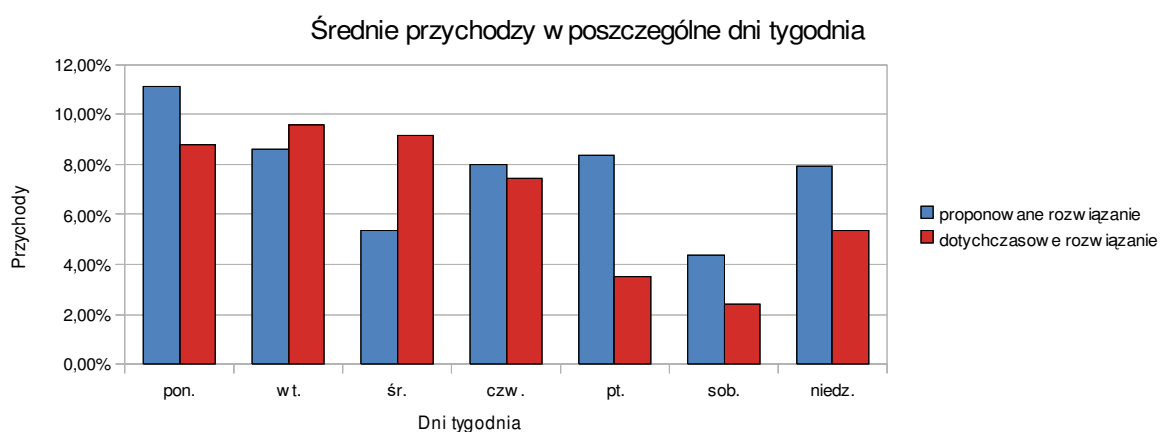
Rys 10. Wartość zamówień w poszczególnych tygodniach. Źródło: opracowanie własne

5.4.3. Skuteczność w poszczególnych dniach tygodnia

Analiza wyników pokazała różnice w skuteczności systemów w zależności od dni tygodnia. W zestawieniu zostały uwzględnione wszystkie dni testów. Wykresy pokazane na rysunku 11 oraz rysunku 12 przedstawiają średnią liczbę zamówień oraz średnie przychody w poszczególne dni tygodnia. Proponowane rozwiązanie najgorsze rezultaty uzyskiwało w środy oraz wtorki, a najlepsze od piątku do niedzieli. Przy czym w sobotę odnotowano duży ogólny spadek sprzedaży związany prawdopodobnie z okresem wypoczynku, jakim jest weekend.



Rys 11. Średnia ilość zamówień w poszczególnych dniach tygodnia. Źródło: opracowanie własne



Rys 12. Średnie przychody w poszczególnych dniach tygodnia. Źródło: opracowanie własne

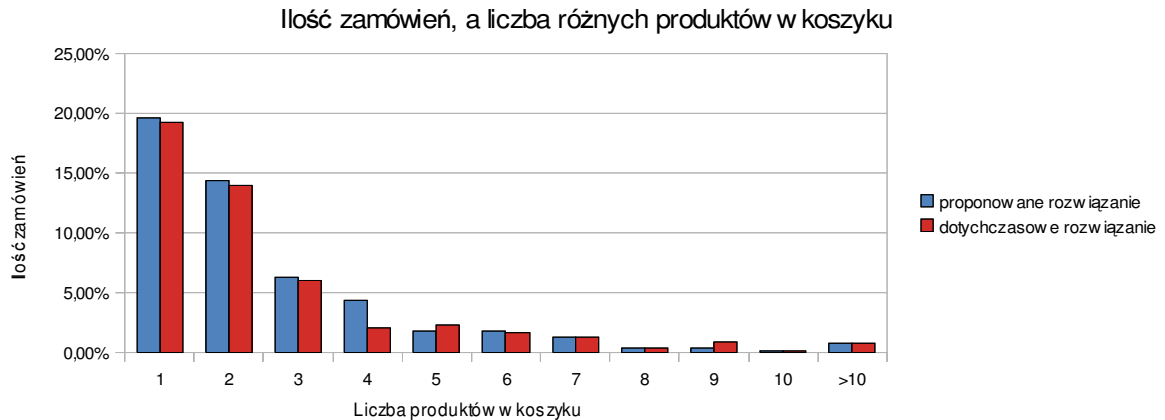
Przy zbliżonej liczbie użytkowników oraz odsłon rekomendacji w obu systemach zaobserwowano, że największa liczba wyświetleń rekomendacji przypada na środę. Wzrost ten wynosi 21,5% w stosunku do najmniejszej liczby wyświetleń przypadającej na sobotę.

Jednocześnie w sobotę, przy tak niewielkim spadku obserwuje się znaczny, ponad dwukrotny spadek liczby zamówień i przychodów. Oznacza to, że najtrudniej o zamówienie w sobotę oraz piątek i niedzielę. Średnia liczba odsłon w stosunku do liczby zamówień jest wtedy największa. Najłatwiej natomiast na początku tygodnia, w dni powszednie. Przyczyną takiego zachowania może być charakter grupy docelowej sklepu. Należy również zaznaczyć, że biuro obsługi klienta oraz infolinia funkcjonuje tylko w dni powszednie. Oznacza to, że proponowany system osiąga lepsze rezultaty wtedy gdy klienci, statystycznie, więcej uwagi poświęcają na przeglądanie oferty. Przy czym użytkownicy proponowanego rozwiązania szybciej podejmują decyzję o zakupie (na podstawie średniej liczby odsłon na zamówienie). Wdrożone rozwiązanie proponuje produkty pod kątem preferencji użytkownika. Nie zawsze rekomendacje te są zgodne z chwilowym zainteresowaniem odbiorcy. Rozwiązanie jest w stanie zaproponować rekomendacje nie związane tematycznie z aktualnie przeglądany produkt. Być może bliżej weekendu osoby odwiedzające sklep mają więcej czasu na przeglądanie oferty i chętniej klikają w rekomendacje produktów, które dostosowane są do ich preferencji. Dodatkowego potwierdzenia należałoby szukać w statystykach średniego czasu spędzonego na stronach w poszczególne dni tygodnia.

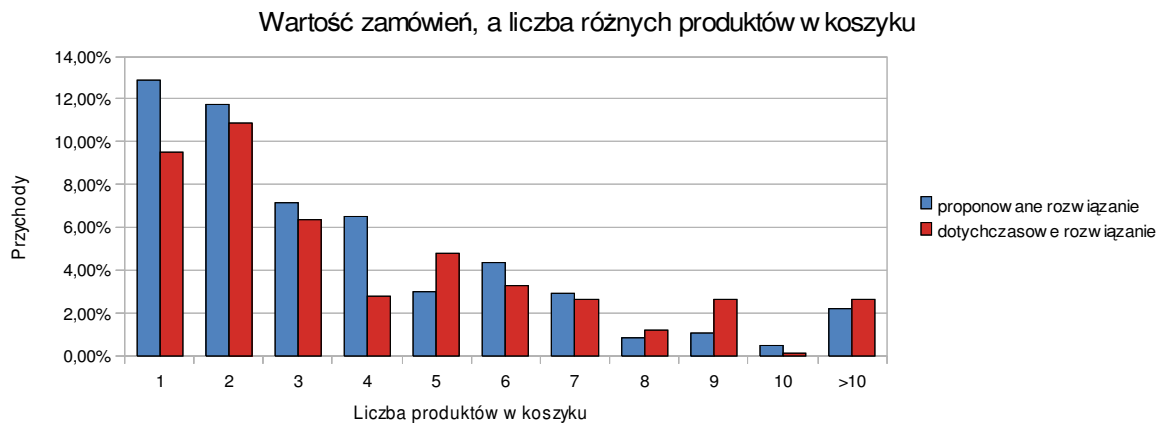
5.4.4. Liczba produktów w koszyku w poszczególnych systemach

Liczba zamawianych pojedynczych sztuk produktów jest równa dla obu systemów. Natomiast zauważalne są różnice w ilości różnych produktów w poszczególnych zamówieniach. Wykresy przedstawione na rysunku 13 oraz rysunku 14 pokazują ilość oraz wartość zamówień pod kątem liczby różnych produktów w danym zamówieniu. Uwzględniają również liczbę sztuk danego produktu, dlatego widoczna jest różnica przy jednym produkcie w koszyku.

Proponowane rozwiązanie wypada lepiej, jeśli w koszyku jest od 1 do 4 różnych produktów. Najlepsze rezultaty uzyskiwane są w przypadku 4 różnych produktów w koszyku. W przypadku większej liczby produktów lepszy jest dotychczasowy system rekomendacji. Należy jednak zauważyć że niecałe 80% zakupów, to zamówienia od 1 do 3 różnych przedmiotów. Na te zamówienia przypada 58,5% przychodów. Uwzględniając jeszcze zamówienia z 4 różnymi produktami, jest to odpowiednio 85,6% zamówień, na które przypada już 67,9% przychodów.



Rys 13. Ilość zamówień, a liczba różnych produktów w koszyku. Źródło: opracowanie własne



Rys 14. Wartość zamówień, a liczba różnych produktów w koszyku. Źródło: opracowanie własne

5.4.5. Podsumowanie wyników

W tabeli 10 przedstawiono dane podsumowujące przeprowadzone testy. Poza *ilością odsłon rekomendacji*, dane przedstawiają różnicę w systemach wynoszącą więcej niż 1%. Takie dane zostały uznane za istotne. Wiele nie zaprezentowanych wskaźników nie przekracza tej różnicy, co wskazuje na duże podobieństwo warunków pracy obu testowanych systemów.

Tabela 10. Podsumowanie testu. Źródło: opracowanie własne

Wskaźnik	Zmiana wartości w stosunku do dotychczasowego systemu
Ilość odsłon rekomendacji	+0,44%
Ilość zamówień	+5,71%
Ilość zamawianych różnych produktów	+5,35%
Średnia liczba odsłon rekomendacji na zamówienie, przez osoby zamawiające	-5,10%
Średnia wartość zamówienia	+7,62%
Przychody	+13,77%

Praktycznie wszystkie prezentowane wskaźniki uległy poprawie w stosunku do dotychczasowego systemu rekomendacji. Niewielkiej poprawie uległa liczba odsłon rekomendacji, przy czym należy pamiętać, że proponowany system posiadał podczas testów niewiele mniej użytkowników. Oznacza to, że średnia liczba odsłon rekomendacji na użytkownika była większa w przypadku proponowanego systemu, jednak różnica ta nie przekroczyła 1%. Sugeruje to również, że system dał lepsze rezultaty nawet przy mniejszej liczbie użytkowników, którzy chętniej przeglądali ofertę sklepu. O 5.71% zwiększyła się liczba zamówień. Ilość zamawianych różnych produktów również uległa poprawie, chociaż dynamika poprawy nie jest już tak duża jak w przypadku ilości zamówień. Niewielkiemu pogorszeniu uległa więc średnia liczba różnych produktów na zamówienie. Jednak różnica ta również nie przekroczyła 1%. Spadek *średniej liczby odsłon rekomendacji na zamówienie przez osoby zamawiające* może być odebrany jako pozytywny. Osoby te prawdopodobnie szybciej podejmowały decyzję o zakupie. W obu systemach średnia liczba odsłon rekomendacji na jedno zamówienie przekraczała 10. O 7.62% zwiększyła się średnia wartość zamówienia. Przy uwzględnieniu opisanych wskaźników *ilości zamówień* oraz *ilości różnych produktów w zamówieniach* wskazuje to na kupowanie droższych produktów w stosunku do zakupów dokonywanych przy dotychczasowym systemie rekomendacji. Rezultatem działania proponowanego rozwiązania był wzrost przychodów sklepu internetowego, w czasie przeprowadzania testów.

6. Podsumowanie

Celem niniejszej pracy było zaimplementowanie algorytmu personalizacji w systemie rekomendacji produktów wybranego sklepu internetowego. System rekomendacji miał być skuteczny w działaniu i wydajny technicznie. Te założenia zostały zrealizowane. System przyczynił się do zwiększenia przychodów sklepu internetowego oraz zwiększenia średniej wartości zamówień. Wydajność techniczna została osiągnięta poprzez implementację stosunkowo prostego algorytmu wykorzystującego w swoich obliczeniach szybkie mechanizmy bazy danych. Prostota oraz rozłożenie obliczeń na przygotowywany co określony czas (w tle) model danych połączony z bieżącymi ocenami użytkowników, pozwoliło na obliczanie rekomendacji w czasie rzeczywistym, bez potrzeby ich buforowania.

Pierwszym elementem niniejszej pracy było uzasadnienie stosowania systemów personalizacji oraz zbadanie trendów występujących w tej dziedzinie. Przedstawione rezultaty poszukiwań wskazują na wykorzystywanie systemów personalizacji w wielu dziedzinach życia oraz tendencję wzrostu ich popularności. Przy tak ogromnych ilościach informacji, do których obecnie ma dostęp praktycznie każdy człowiek, nie sposób obejść się bez systemów, które pomagają w dopasowaniu danych do preferencji odbiorcy. Zebrane informacje zostały zaprezentowane w rozdziale drugim. Kolejnym elementem wymaganym do realizacji celów pracy, było znalezienie oraz zaznajomienie się z technologiami, które umożliwią jak najprostszą realizację wszystkich założeń systemu rekomendacji. Sposób funkcjonowania systemów rekomendacji, wykorzystywane technologie oraz napotymane problemy zostały omówione w rozdziale trzecim. W rozdziale czwartym przedstawione zostały założenia oraz implementacja proponowanego rozwiązania. Do realizacji zadania posłużył, jako „serce” systemu, ważony algorytm *Slope One*. Rozwiązanie wymagało zaimplementowania dodatkowych mechanizmów normalizacji automatycznych ocen, identyfikacji użytkowników oraz usuwania anomalii. Zastosowanie ogólnie dostępnych technologii oraz prostych algorytmów przyczyniło się do dużej wydajności, skuteczności oraz elastyczności rozwiązania. Po dobraniu odpowiednich parametrów w procesie wdrożenia, system nadaje się do wykorzystania praktycznego, na każdej stronie internetowej wymagającej mechanizmów rekomendacji. Automatyzacja procesu oceniania, spowodowała że system nie jest obciążeniem psychicznym dla użytkownika. Nie wymaga podawania żadnych danych personalnych, ani ręcznego oceniania czy opiniowania danego produktu. Zmniejsza to też rzadkość danych, ponieważ liczba ocen przypadająca na jednego

użytkownika jest większa niż w przypadku systemów wykorzystujących manualne mechanizmy. W rozdziale czwartym przedstawiono również, korzystne dla algorytmu *Slope One*, empiryczne porównanie systemu z innymi rozwiązaniami. Kolejny rozdział przedstawia wdrożenie zaproponowanego rozwiązania w wybranym sklepie internetowym. Rozwiązanie zostało wykorzystane w sklepie posiadającym kilkadziesiąt tysięcy użytkowników i bardzo dobrze poradziło sobie z wygenerowanymi przez nich danymi. Zaproponowany system porównywany był z już wcześniej istniejącym systemem rekomendacji opartym na prostej analizie koszykowej historycznych zamówień. Testy równoległe A/B, pozwoliły wyeliminować czynniki sezonowości sprzedaży i uzyskać porównywalne, niemal identyczne warunki pracy dla obu systemów rekomendacji. Potwierdzeniem tego mogą być niemal identyczne wyniki dla niektórych z badanych wskaźników. Dodatkowo uzyskane dane zostały przefiltrowane pod kątem występowania anomalii. Stąd można wnioskować, że są wysoce wiarygodne. Dane zostały przedstawione w różnych perspektywach, pod kątem ilości produktów w koszyku, przychodów, skuteczności w poszczególne dni tygodnia, zmiany skuteczności w czasie testów oraz globalnie. Wskazują również na istotę procesu nauczania algorytmu, który z czasem osiągnął lepsze rezultaty działania.

Chociaż uzyskane wyniki są zadowalające można pokusić się o dalszy rozwój proponowanego systemu rekomendacji. Rozwiązanie mogłoby zostać wzbogacone o uwzględnianie czasu dokonywanych ocen, co pozwoliłoby lepiej dostosować się do zmiennych w czasie gustów odbiorców. System mógłby zostać rozbudowany o interfejs programistyczny (API), umożliwiający wdrożenie rozwiązania również w sklepach wykorzystujących inne niż PHP technologie programistyczne. Dobrym pomysłem wydaje się również przetestowanie systemu, w przypadku opisanego sklepu internetowego, poza kartami produktów. Rozwiązanie może być bowiem stosowane na dowolnej stronie, również nie związanej z danym produktem oraz w systemach posprzedażowych, np. w e-mail marketingu do klientów. W przypadku dalszych testów należałoby więcej uwagi poświęcić na generowaniu rekomendacji dla danego produktu, nie tylko dla danego użytkownika. Takie rekomendacje można też dowolnie ze sobą łączyć w celu uzyskania jeszcze lepszych rezultatów. Należałoby się również zastanowić nad sposobem prezentacji rekomendacji. Istnieją duże szanse, że rekomendacje lepiej wyeksponowane lub z zastosowaniem zdjęcia miałyby jeszcze wyższą skuteczność w stosunku do zwykłych odnośników tekstowych.

7. Bibliografia

- [1] Segaran T., Programming Collective Intelligence, O'Reilly, 2007
- [2] Su Xiaoyuan, Khoshgoftaar Taghi M., A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence, 2009
- [3] Shafer J., Konstan J., Riedl J., E-Commerce Recommendation Applications, Data Mining and Knowledge Discovery, 5, 115-153, 2001
- [4] Kazienko P., Kołodziejski P., Personalized Integration of Recommendation Methods for E-commerce, International Journal of Computer Science & Applications 3(3), 12-26, 2006
- [5] Kobsa A, Koenemann J, Pohl W., Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships, The Knowledge Engineering Review 16(2), 111-155, 2001
- [6] Adomavicius G., Tuzhilin A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions On Knowledge and Data Engineering, 17(6), 2005
- [7] Sarwar B., Karypis G., Konstan J., Riedl J., Application of Dimensionality Reduction in Recommender System -- A Case Study, IN ACM WEBKDD WORKSHOP, 2000
- [8] Szczepański L., Rekomendacyjny interfejs użytkownika systemu zarządzania zawartością multimedialną, praca magisterska, Politechnika Wroclawska, wydział Informatyki i Zarządzania,
- [9] Lemire D., Maclachlan A., Slope One Predictors for Online Rating-Based Collaborative Filtering, in SIAM Data Mining, 21-23, 2005
- [10] Pu Wang, HongWu Ye, A Personalized Recommendation Algorithm Combining Slope One Scheme and User Based Collaborative Filtering, Intelligent Information Systems, 152-154, International Conference on Industrial and Information Systems, 2009
- [11] Hellwig Z., Elementy rachunku prawdopodobieństwa i statystyki matematycznej, PWN, 1998