



**INSTYTUT BADAŃ SYSTEMOWYCH  
POLSKIEJ AKADEMII NAUK**

**TECHNIKI INFORMACYJNE  
TEORIA I ZASTOSOWANIA**

Wybrane problemy  
Tom 1(13)

*poprzednio*

**ANALIZA SYSTEMOWA W FINANSACH  
I ZARZĄDZANIU**

Pod redakcją  
Jerzego HOŁUBCA

Warszawa 2011



**INSTYTUT BADAŃ SYSTEMOWYCH  
POLSKIEJ AKADEMII NAUK**

# **TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA**

Wybrane problemy  
Tom 1(13)

*poprzednio*

**ANALIZA SYSTEMOWA W FINANSACH  
I ZARZĄDZANIU**

Pod redakcją  
Jerzego HOŁUBCA

Warszawa 2011

Wykaz opiniodawców artykułów zamieszczonych  
w niniejszym tomie:

Dr hab. inż. Przemysław GRZEGORZEWSKI, prof. PAN

Prof. dr hab. inż. Jerzy HOŁUBIEC

Dr inż. Tatiana JAWORSKA

Dr hab. inż. Wiesław KRAJEWSKI, prof. PAN

Dr hab. inż. Maciej KRAWCZAK, prof. PAN

Dr hab. Michał MAJSTEREK

Dr hab. inż. Andrzej MYŚLIŃSKI, prof. PAN

Prof. dr hab. inż. Witold PEDRYCZ

Dr hab. inż. Ryszard SMARZEWSKI, prof. KUL

Prof. dr hab. inż. Andrzej STRASZAK

Dr Dominik ŚLĘZAK

Prof. dr hab. inż. Stanisław WALUKIEWICZ

© Instytut Badań Systemowych PAN  
Warszawa 2011

**ISBN 9788389475336**

# SYSTEMY REKOMENDACYJNE – ICH NATURA I OGRANICZENIA

***Przemysław Kusaj***

*Studia Doktoranckie IBS PAN*

*Katolicki Uniwersytet Lubelski Jana Pawła II*

**Abstract.** *Together with the Internet becoming a collaborative space, open to all users to create, share and rate online data, the amount of data available has started growing to enormous quantities. In order to find the interesting and useful to a user pieces of information (or better to say – objects, due to different types of data- e.g. text-based data, multimedia data) the user had to go through an effort- and time-consuming searching and assessing process. The need to support a user in the process of discovering interesting information emerged. As a response experts in data analysis and classification have started developing special systems based on various filtering techniques that would recommend online objects to a user according to his/her preferences and expectations – so in a personalized manner.*

## 1. Wstęp

W ciągu ostatnich 10 lat, wraz z rozwojem Internetu i wzrostem jego znaczenia w życiu codziennym, powstaniem portali internetowych i społecznościowych utrzymanych w trendzie Web 2.0, oferujących różnorakie treści generowane i udostępniane przez użytkowników, oraz pojawieniu się coraz większej ilości sklepów internetowych, przyrost dostępnych informacji stał się wręcz przytłaczający. Przed 2001 rokiem, użytkownik starając się natrafić na nieznaną mu jeszcze informację i multimedia o interesującej go tematyce, czy też produkty danego typu, zmuszony był podejmować czasochłonne poszukiwania. W dobie Web 2.0, kiedy to Internet stał się otwartą na użytkowników platformą swobodnego współtworzenia informacji i współdzielenia wiedzy, mając do dyspozycji nowe kanały komunikacji z innymi użytkownikami, np. korzystając z portalu zrzeszającego osoby o podobnych zainteresowaniach, osoba może łatwo skorzystać z podpowiedzi i rekomendacji innych. W związku z szybko przyrastającą ilością treści online szybko dostrzeżono potrzebę filtrowania in-

formacji tak, aby użytkownik otrzymywał natychmiastowy dostęp przede wszystkim do tego, co może go zainteresować. By zautomatyzować i spersonalizować takie filtrowanie, zaczęto masowo wdrażać ideę systemów rekomendacyjnych, które w obecnej chwili są już elementarną częścią wszystkich portali, czy to tych zorientowanych na treści konkretnego typu, czy społecznościowych umożliwiających komunikację z przyjaciółmi, czy też w e-biznesie - sklepów internetowych i portali aukcyjnych. Na bazie pewnych dostępnych informacji wspomagają one użytkowników w ich poszukiwaniach sugerując treści czy produkty podobne do tych przez nich przeglądanych, kupowanych, lubianych.

W niniejszym artykule w paragrafie drugim opisane zostały typy systemów rekomendacyjnych przez wzgląd na ich działanie i analizowany typ informacji dla uzyskania poszukiwanych wyników oraz ich główne zastosowania. W paragrafie trzecim wskazano ich ograniczenia dotyczące rekomendacji utworów muzycznych oraz przywołano badania nad analizą dźwięku zmierzającą do przezwyciężenia ograniczeń i trudności związanych z sugerowaniem utworów muzycznych.

## **2. Systemy rekomendacyjne**

Systemy rekomendacyjne (w dalszej części nazywane SR) to takie, które u celu swego działania mają w zautomatyzowany i spersonalizowany sposób dostarczyć użytkownikowi informacje i wskazania o treściach, które mogą być dlań interesujące, poprzez specjalny proces filtrowania wszystkich dostępnych informacji w danej przestrzeni informacyjnej (np. w obrębie danego portalu internetowego czy zakupowego). Jak zauważają del Olmo i Gaudio [1] spojrzenie na działanie i cele SR zmieniło się w czasie. Świadczyć o tym mogą przywołane przez nich dwie definicje SR. Pierwsza z nich, zaproponowana przez Resnicka i Variana w 1997 roku [2], mówi o SR jako o takich, gdzie danymi wejściowymi są rekomendacje dostarczone przez ludzi (użytkowników), które następnie są agregowane przez system i wysyłane, jako dane wyjściowe, do odpowiednich odbiorców. W 2002 Burke [3] przedstawił bardziej ogólną definicję, ale zawierającą w sobie główny cel stosowania SR. Tym celem i głównym założeniem SR jest „nakierowywanie użytkownika na interesujące lub przydatne obiekty w ogromnej przestrzeni możliwych opcji w spersonalizowany sposób”.

Struktura systemu rekomendacyjnego składa się ze źródła danych wejściowych, metody rekomendującej oraz danych wyjściowych – w postaci gotowej rekomendacji lub (w większości przypadków) zhierarchizowanej listy rekomendacji [4]. Źródło danych wejściowych to profil użytkownika zawierający

np. jego preferencje dotyczące poszukiwanych treści, poszukiwane cechy produktów, historię przeglądanych i zakupionych produktów, oceny obiektów (stron WWW, treści internetowych, artykułów, obiektów multimedialnych), słowa kluczowe, dane demograficzne itp. [5]. Część z tych danych pobierana jest do SR niejawnie, np. poprzez śledzenie klikanych linków stron WWW (np. podstron sklepu internetowego dotyczących interesujących użytkownika produktów), śledzenie zapytań wpisywanych do wyszukiwarki internetowej (Google), wyszukiwanie słów kluczowych w wiadomościach e-mail (gmail.com).

Dzięki tego typu danym każdy użytkownik, poprzez analizę zestawu wyodrębnionych jego cech i preferencji, zostaje przyporządkowany do grupy użytkowników o zbliżonych cechach. Szerokim zastosowaniem do tego celu cieszy się *Clustering Analysis*, u podłoża której leży zasada przyporządkowywania użytkowników do klastrów o charakterystyce zbiorów ostrych (różnice cech charakterystycznych (odległości) pomiędzy użytkownikami w obrębie danego klastra minimalne; różnice (odległości) pomiędzy klastrami maksymalne), przechowywanych off-line w bazie danych np. portalu, w obrębie którego się porusza się analizowany użytkownik [6]. Poprzez analizę nowego użytkownika i odpowiednie przypisanie go do grupy, a potem analizę właściwości i cech tej grupy, można starać się przewidzieć jego zachowania czy preferowane przez niego treści czy produkty i zarekomendować najczęściej preferowane/zakupywane przez większość w tej grupie treści.

Z uwagi na fakt, że powyższe rozwiązanie nie pozwalało na rekomendowanie obiektów innych niż te preferowane przez grupę w [7] zaproponowano zastosowanie reguł skojarzeń pomiędzy elementami dużych baz danych (*association rules*). Do tego, dodatkowo, tworzone były też reguły rekomendacyjne w takiej ilości, która pozwoliłaby na obsługę różnych zachowań użytkownika (metoda ta była ukierunkowana na rekomendacje w sklepach internetowych, więc w przywoływanym tu artykule rozpatrywano zachowania związane z zakupami). Na bazie wcześniejszych zachowań / transakcji użytkownika algorytm poprzez reguły skojarzeń wyznaczał reguły rekomendacyjne, a te z kolei określały listę nie znanych jeszcze użytkownikowi obiektów. Poprzez reguły skojarzeń wyniki były sortowane i N znajdujących się najwyżej w rankingu było wybieranych jako rekomendowany zestaw. Rozwiązanie to jednak, przy bardzo dużej liczbie transakcji i reguł rekomendacyjnych, okazywało się nieprecyzyjne.

Aby usprawnić i doprecyzować system selekcji obiektów do rekomendacji potrzebna jest sprawnie działająca technologia filtrowania informacji. Internet, będąc platformą swobodnej wymiany informacji, zrzeszającej podobnych sobie użytkowników w społecznościach, zapewnia ciągłą weryfikację i ocenę

treści przez samych użytkowników. Oprócz tego, daje im również możliwość klasyfikacji obiektów i ich opisywania przy użyciu tzw. *Tagów*. W związku z tym tzw. „wisdom of crowds”, czyli charakterystyczne dla społeczności internetowych Web 2.0 współdziałanie i wymiana informacji mogły służyć filtrowaniu użytecznych i interesujących obiektów oraz wspomaganie podejmowania decyzji przez użytkowników należących do danej społeczności. W oparciu o to zaproponowane zostały dwie koncepcje filtrowania obiektów, implementacje których stosuje się obecnie SR.

## 2.1 Collaborative filtering (CF)

Bazując na relacjach i stopniu podobieństwa użytkowników stron i portali WWW, Montagner et al. (2003) [8] zaproponowali technologię filtrowania obiektów w sieci zwaną *Collaborative filtering*. Polega ona na przewidywaniu preferencji użytkownika w oparciu o preferencje innych, którzy oceniają obiekty lub kupują produkty tworząc w ten sposób rankingi. Danymi, które określają moc relacji i skalę podobieństwa pomiędzy użytkownikami, jest ilość wystawionych takich samych ocen/opinii tym samym obiektem lub zakupionych takich samych przedmiotów przez poszczególnych użytkowników. Danymi wejściowymi dla SR działającego w oparciu o tą technologię są więc informacje zebrane w profilach użytkowników – dane dotyczące tworzonych ocen/rankingów i historia zakupów. Rekomendacja nowych interesujących obiektów odbywa się przez pryzmat ocen wystawionych tym obiektem przez innych, podobnych użytkowników.

Technologia ta sprawdza się przy rozpatrywaniu dobrze znanych, ocenionych przez wiele osób obiektów. Problem pojawia się jednak, gdy nowy obiekt zostaje dodany do zbioru. Obiekt taki nie mając żadnych ocen lub taki, który jeszcze nigdy nie został zakupiony przez nikogo z danej grupy w obrębie społeczności, prawdopodobnie nigdy nie zostanie zarekomendowany.

## 2.2 Content-Based filtering (CB)

Powyższy problem częściowo rozwiązuje kolejna technologia, oparta na filtrowaniu przez pryzmat zawartości obiektów – *Content-Based Filtering* [9]. Aby dopasować obiekty do preferencji i zainteresowań użytkownika (zapisanych w profilu) niezbędna jest analiza cech zawartości obiektu (np. częstotliwość występowania poszukiwanego przez użytkownika terminu w tekście lub na stronie WWW) oraz/lub metadanych, którymi opisane są obiekty (autor, reżyser, aktorzy, twórca, gatunek, tematyka). Analiza to prowadzi do indeksowania obiektów, by wyodrębnić te najbardziej odpowiednie do zarekomendo-

wania. Technologia ta sprawdza się w szczególności, gdy rozpatrywane obiekty tekstowe – dokumenty, treść stron WWW, itp., w przypadku innych mogą wystąpić trudności z indeksowaniem. Przykładem obiektów trudnych do zindeksowania są obiekty muzyczne lub video.

Dlatego też, jak słusznie zauważają autorzy w [4] technologia ta nie nadaje się do wszystkich zastosowań. Obecnie tworzone są jednak systemy hybrydowe, łączące w sobie zalety obu opisanych w 2.1 i 2.2 podejść.

### 3. Problem rekomendacji utworów muzycznych

Nawet łącząc zalety wyżej wspomnianych technologii, rekomendacje utworów muzycznych czy filmów są problematyczne, gdyż w dużej mierze zależą od ocen przyznawanych tym obiektom przez użytkowników. Co za tym idzie, tego typu obiekty nowe, nie ocenione, nigdy nie zakupione przez kogośkolwiek zostają automatycznie odrzucane przy rozpatrywaniu obiektów do rekomendacji.

Próbując rozwiązać problem rekomendacji utworów muzycznych w [10] zaproponowano hybrydowy system na oparty na analizie polifonicznych tembrów dźwięku w obiektach muzycznych w połączeniu, bazując na modelu systemu rekomendującego dokumenty tekstowe poprzez analizę lingwistyczną Popescule'a (bag-of-words). Jest to metoda, która niskim kosztem pozwala wyekstrahować z sygnału audio pewne cechy charakterystyczne dla struktury utworu, ponieważ nie przewiduje ona rozdziału badanego dźwięku na poszczególne instrumenty, analizuje natomiast brzmienie „mieszanki” dźwięków w danym momencie. Jednakże sami autorzy przyznali, że klastrowanie materiałów audio przez pryzmat podobieństwa tembrów dźwięku wydajnościowo jest ograniczone. W związku z tym wspomagają je „wiedzą tłumów”, czyli tym, co charakterystyczne dla *Collaborative filtering*'u. Zalety CF, niestety, pociągają za sobą również wady tej technologii – znów nowo wprowadzone czy nieznanne utwory nie są brane pod uwagę przy tworzeniu listy obiektów do zarekomendowania.

#### 3.1 Przykładowe badania mogące pomóc w rozwiązaniu problemu

Aby rozwiązać problem nowych utworów właściwym wydaje się być odejście od korzystania z CF i.



Aby ograniczyć dźwiękowe obiekty analizy warto zwrócić uwagę na badania Karydis et al. (2006) [11]. Podjęli oni starania zmierzające do wyszukania w utworach muzycznych powtarzający się wzorzec dźwiękowy o maksymalnej długości. Zaproponowany przez nich algorytm M<sup>2</sup>P przeszukuje graf skierowany reprezentujący sekwencję dźwięków w utworze używając operacji string-join i wykazuje on wydajność większą od swojej konkurencji – algorytmu HLC. Niestety, algorytm był testowany tylko na formacie dźwiękowym MIDI, więc w rzeczywistości na symbolicznym zapisie dźwięków utworu. W większości jednak materiały dźwiękowe, chociażby w przypadku sklepów internetowych, dostępne są w formacie MP3. Metoda ta jednak pozwala wyszukać w całości utworu najdłuższy, nie-trywialny jego fragment, co można uznać za motyw przewodni utworu i na jego bazie prowadzić rozważania dotyczące całości np. na potrzeby systemów rekomendacyjnych, celem redukcji ilości danych do analizy. Dalsze badania nad algorytmem M<sup>2</sup>P zmierzają właśnie w tym kierunku.

Jak wskazują badania autorów [12] klasyfikacja utworów muzycznych i grupowanie ich w gatunki (nazwane przez ekspertów grupy obiektów wykazujących wspólne cechy celem określenia ich stylistyki) jest nie tylko bardzo istotnym elementem metadanych opisujących materiały audio, ale i jednym z najczęstszych zapytań kierowanych do wyszukiwarek. W ogromnej ilości udostępnianych w sieci utworów można dopatrzeć się jednak nieprecyzyjności i subiektywizmu w kategoryzacji utworów i błędów w metadanych, co również wpływa na problemy z wyszukiwaniem i rekomendacją utworów muzycznych. Dlatego też wykorzystują oni uczenie maszynowe do automatycznej klasyfikacji gatunków muzycznych. Poprzez dekompozycję czasowo-przestrzenną sygnału audio ekstrahują oni możliwe do pogrupowania cechy, tworząc w rezultacie 30-wymiarowy wektor cech. Poprzez zastosowanie klasyfikatorów i procedur decyzyjnych utwory przypisywane są ostatecznie do określonej klasy.

Cechy wyekstrahowane w opisywanych w [12] badaniach pogrupowane zostały w kategorie cech:

- związanych z rytmem
- związanych z barwą dźwięku
- związanych z tonacją

Klasyfikacja wydaje się przyporządkowywać utwory do konkretnych gatunków na zasadach grupowania ostrego.

Dla pewnych, relatywnie prostych w stosunku do całego problemu, zastosowań, jak choćby z doбором rytmicznie pasujących do siebie utworów w obrębie bazy danych sklepu internetowego, można by przechowywać w bazie danych lub jako metadane plików muzycznych niektóre, najbardziej znaczące dla określania stylistyki utworu cechy.

Zastanawiające jest, czy zastosowanie grupowania rozmytego na potrzeby rekomendacji utworów muzycznych mogących zainteresować użytkownika pozwoliłoby na uzyskanie ciekawych, zróżnicowanych rezultatów. Teoretycznie, jeżeli każdy utwór należałby do każdego z gatunków z różnym stopniem przynależności, wówczas możliwe, że poprzez wyszukanie najbliższych sąsiadów można by otrzymać zróżnicowaną listę rekomendowanych utworów, które wg podziału ostrego należałyby do różnych zbiorów / gatunków.

## Podsumowanie

Systemy rekomendacyjne stały się integralną częścią praktycznie każdego portalu internetowego. Filtrując ogromne ilości dostępnych treści wspomagają użytkowników w wyszukiwaniu interesujących ich, użytecznych obiektów w spersonalizowany sposób. O ile w przypadkach, gdzie obiekty są danymi tekstowymi wydajność ich jest zadowalająca. Jeżeli natomiast szukane obiekty są danymi multimedialnymi a meta dane je opisujące są nieprecyzyjne lub obiekty te są nowe, jeszcze nieznanne i nie ocenione w żaden sposób przez innych użytkowników, szanse odnalezienia ich czy to przez użytkownika czy przez systemy rekomendacyjne są znikome. By zaradzić temu problemowi, prowadzone są liczne badania i opracowywane są nowe technologie i algorytmy. Kilka wymienionych w tym artykule dotyczy analizy danych dźwiękowych. Jednakże, przez złożoność tego problemu, są one bardzo znikomą częścią wszystkich prowadzonych prac, często charakteryzujących się zupełnie odmiennymi spojrzeniami na analizę tego typu danych. Ilość różnego rodzaju diametralnie różnych podejść do analizy sygnału audio i ekstrakowania cech charakterystycznych a także ich reprezentacji cyfrowej oraz algorytmów porównywania jest tak duża, że niestety niemożliwym jest przywołania i porównania tutaj nawet znacznej części z nich. Można zatem wywnioskować, że ten dział wiedzy jest wciąż bardzo aktywnie badany.

## Literatura

- [1] del Olmo, F. H., & Gaudioso, E. (2008), Evaluation of recommender systems: A new approach. *Expert Systems with Application* (35), pp. 790-804.

- [2] Resnick, P. & Varian, H. (1997), Recommender Systems. *Communications of the ACM* vol. 40(3).
- [3] Burke, R. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*. 12(4), pp 331-370.
- [4] Wang, H.-F. & Wu, C.-T. (2009), A mathematical model for product selection strategies in a recommender system. *Expert Systems with Application* (36), pp. 7299 – 7308.
- [5] Schafer, J.B., Konstan, J.A. & Riedl, J. (2001), E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5, pp 115-153.
- [6] Theodoridis, S. & Koutroumbas, K. (2003), Pattern recognition. *Academic Press*.
- [7] Wang, H.-F. & Hong, W.K., (2006), Managing customer profitability in a competitive market by continuous data mining. *Industrial Marketing Management Journal* (35), pp. 712-723.
- [8] Montaner, M., López, B., de la Rosa, J. Ll. (2003), A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review*, Kluwer Academic Publishers. Vol.19(4), pp. 285-330.
- [9] Mooney, R.J. & Roy, L. (2000), Content-Based Book Recommending Using Learning for Text Categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 195-204.
- [10] Yoshii, K., Goto, M., Komatani, K. & Ogata, T. An Efficient Hybrid Music Recommender System Using an Incrementally Trainable Probabilistic Generative Model. *IEEE Transaction on Audio, Speech, and Language Processing*, Vol. 16(2), pp. 435-447.
- [11] Karydis, I., Nanopolos, A., Manolopoulos, Y. (2006), Finding maximum-length repeating patterns in music databases. *Multimed Tools Appl* (32), pp. 49-71.
- [12] Silla Jr., C.N., Koerich, A.L. & Kaestner, C.A.A. (2008), A Machine Learning Approach to Automatic Music Genre Classification. *Journal of the Brazilian Computer Society*, 14(3), pp. 7-18.
- [13] Rutkowski, L. (2006), *Metody i techniki sztucznej inteligencji*. PWN.

ISBN 9788389475336

