Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-00000

Martin Černák

Dynamické odporúčanie

Bakalárska práca

Študijný program: Informatika
Študijný odbor: 9.2.1 Informatika
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva
Vedúci práce: prof. Ing. Pavol Návrat, PhD.

Máj 2015

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informatika
Autor: Martin Černák
Bakalárska práca: Dynamické odporúčanie
Vedúci práce: prof. Ing. Pavol Návrat, PhD.
Máj 2015

Dynamické odporúčanie v kontexte hudobných dokumentov je vďaka svojmu úzkemu zamerania a vďaka menšej komunite značne nepreskúmané. Existuje niekoľko riešení, ktoré ale nevyužívajú plný potenciál dynamického odporúčania. Jednou z možností ako tieto systémy vylepšiť je začať uvažovať starnutie ako používateľových tak globálnych preferencií. V hudobnóm odvetví môžeme častejšie ako v ostatných vidieť príchod mimoriadne populárnych nových interprétov, piesni a štýlov, ktoré rýchlo vymiznú z povedomia verejnosti, prípadne zostane okolo nich úzka skupina fanúšikov. Kontrastom k nim sú piesne, autori a hudobné štýly, ktoré pretrvávajú dlhodobo v povedomí ľudí a vypadajú, že starnutie na nich nemá vplyv.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Informatics
Author: Martin Černák
Bachelor thesis: Dynamic recommendation
Supervisor: prof. Ing. Pavol Návrat, PhD.
May 2015

Dynamic recommendation in the context of musical documents thanks to its narrow focus and with smaller community largely unexplored. There are several solutions but that don't using full potential of dynamic recommendation. One way to improve these systems is to start thinking of aging user and global preferences. In the music sector this can be more frequent than in other sectors, extremely popular new artists, songs and styles that quickly disapear from public awareness or remain around them a small group of fans. A contrast to them are songs, authors and musical styles that persist long time in the minds of people and looks like aging does not affect them.

POĎAKOVANIE

Chcel by som v prvom rade poďakovať pánu profesorovi Pavlovi Návratovi za jeho odbornú pomoc a motiváciu a za všetky konzultácie ktoré sme spoločne absolvovali. Zároveň by som rád poďakoval učástníkom jeho výskumného seminára za konštruktívnu kritiku, ktorá taktiež prispela k zdokonaleniu tejto práce.

ČESTNÉ PREHLÁSENIE

Čestne prehlasujem, že záverečnú prácu som vypracoval samostatne s použitím uvedenej literatúry a na základe svojich vedomostí a znalostí.

.....

Martin Černák

Obsah

Zoznam obrázkov

Zoznam ukážok

1 Úvod

Ako z jej názvu vyplýva, informatika je predmet zameraný na prácu s informáciami. To čo kedysi bolo najväčším problémom, teda dostať nejaké informácie k používateľom, už dávno nie je problém. Vďaka internetu sa dajú informácie dostať prakticky všade. No teraz čelíme väčšiemu problému. Naša spoločnosť dokáže za účelom zábavy, rozvoja alebo produktivity vyprodukovať neuveriteľné množstvo informácií. Precíznosť archivácie údajov je asi najväčšia v histórií, problém nastáva ak chceme nejaké údaje vyhľadať. Klasický prístup spravovania informácií už nie sú dostačujúce a jednoduché vyhľadávanie už nieje dostatočne efektívne na to aby sme boli schopní nájsť požadované informácie.

Dokonca aj vyhľadávanie ako také prestáva byť dostatočne efektívne, namiesto neho sa dostáva do popredia odporúčanie, ktoré doslova používateľovi ponúkne informácie, ktoré by ho mohli zaujímať, bez toho aby musel vynaložiť akúkoľvek námahu na hľadanie. Aby mohol systém robiť takúto predikciu potrebuje poznať používateľa a to mu umožňuje profilovanie používateľov. Profil používateľa je komplexná vec. Záujmy používateľa môžu byť ovplyvnené jeho demografickými parametrami (vek, vzdelanie, miesto pobytu), záujmami a všeobecnými novinkami ako vydanie nového albumu obľúbenej kapely alebo uvedenie nového zariadenia na trh. Do úvahy musíme brať aj udalosti v živote používateľa, napríklad narodenie potomka tiež v určitom smere ovplyvní používateľove záujmy. Z toho vyplýva že profil musí byť dynamický, a preto je potrebné nejakým spôsobom aj odoberať záujmy, o ktoré používateľ už viac neprejavuje záujem.

Cieľom tohoto projektu je vytvoriť aplikáciu ktorá bude schopná dynamicky odporúčať. Na riešenie hore spomenutých problémov existuje množstvo prístupov. Každý z týchto prístupov má mierne lepšie výsledky v iných situáciách, čiže dosť závisí od domény, pre ktorú bude systém odporúčať. V tomto projekte sa budeme zaoberať doménou hudobných dokumentov (akordy, texty, taby, preklady). Táto oblasť ešte nie je prebádaná, čo nám prináša nové možnosti ako aj nové problémy.

1.1 Použité pojmy a skratky

relácia - Sekvencia akcií vykonaná používateľom na vyhľadanie požadovaného dokumentu.

akcia - Jedna elementárna interakcia používateľa so systémom, kliknutie na odkaz, zadanie vyhľadávacieho reťazca.

hudobné dokumenty - Dokumenty, ktoré určitým spôsobom modelujú hudobné dielo v človekom čitateľnej podobe (nemusia zaznamenávať všetky vlastnosti hudobného diela), teda hlavne taby, akordy, texty a ich kombinácie.

preferencia - Akákoľvek vlastnosť hudobného dokumentu, na základe ktorej sa určuje či je daný dokument vhodný pre používateľa.

vlastnoť dokumentu - Opisná vlastnosť dokumentu, ktorú vieme spárovať na nejakú používateľovu preferenciu.

používateľov zámer - súbor preferencií ktorí chcel používateľ vyjadriť pomocou zadaného vyhľadávacieho reťazca.

2 Existujúce riešenia

Existuje veľké množstvo riešení problému s dynamickým odporúčaním. Všetky pracujú s nejakou množinou vlastností, tém alebo kľúčových slov, ku ktorým sa snažia vypočítať pravdepodobnosť, že práve daná téma, kľúčové slovo alebo vlastnosť je pre používateľa najzaujímavejšia a následne mu odporučiť vyhľadávané subjekty, z danou vlastnosťou.

Postup riešenia sa dá rozdeliť na niekoľko podproblémov, ktoré sa dajú riešiť osobitne:

- získavanie vlastnosti dokumentov,
- získavanie používateľskej spätnej väzby,
- ukladanie používateľského profilu,
- porovnávanie používateľových preferencií s dokumentami,
- starnutie preferencií,
- triedenie preferencií na krátkodobé a dlhodobé,

2.1 Kategorizácia dokumentov

Aby sme mohli odporučiť dokument, musíme získať nejaké jeho vlastnosti, ktoré budeme vedieť priradiť k používateľskému profilu.

Kategorizovanie na základe typu dokumentu

Prvou skupinou vlastností určujúcich dokument je jeho typ. Dokumenty môžu byť piatich typov, pričom sú možné aj rôzne kombinácie typov, teda môže nastať, že mám text s akordmi pričom sólo daného hudobného diela je zapísané pomocou tabov. Takže pre každý dokument d_j z množiny dokumentov $D = (d_0, d_1, \cdots, d_n)$ budem mať určené štyri premenné $d_j = (a_j, x_j, p_j, t_j, n_j)$, ktoré budú nadobúdať buď hodnotu 1 alebo 0 na základe toho, či dokument patrí do danej kategórie alebo nie. Jednotlivé premenné reprezentujú nasledovné druhy obsahu:

- a_j určuje či dokument j obsahuje akordy,
- x_j určuje či dokument j obsahuje text,
- p_j určuje či dokument j obsahuje preklad textu,
- t_j určuje či dokument j obsahuje taby,
- n_j určuje či dokument j obsahuje noty,

Kategorizovanie na základe hudobnej štruktúry

Hudobné dokumenty sa dajú ďalej deliť z hľadiska hudobnej štruktúry. Pre dokument d_j vieme určiť niekoľko časti štruktúry, Jeden hudobný dokument nemusí obsahovať všetky súčasti hudobného diela. Tak isto jedno hudobné dielo nemusí používať všetky štandardné súčasti. V zásade rozoznávame a určujem tieto štandardné časti hudobných diel [**?**]:

- predohra,
- medzihra,
- refrén,
- ukončenie (angl. Outro),
- sólo alebo inštrumentálna časť

No toto nie je najmenšie možné delenie, z hľadiska kompozície môžeme ešte hudobné dielo rozdeliť na jednotlive nástroje, ktoré vykonávajú dané prevedenie hudobného diela. Tak isto sa dané časti môžu rozlišovať variáciami motívu[**?**],

Kategorizovanie na základe prevedenia

Ďalej môžeme hudobné diela deliť na základe konkrétneho prevedenia. Niektoré hudobné diela môžu mať aj niekoľko prevedení. Prevedenia sa dajú charakterizovať na základe miesta, použitých nástrojov alebo hudobníkov, ktorí dané prevedenie zahrali.

Kategorizovanie na základe žánru

Žáner je asi jeden z najdôležitejších spôsobov kategorizovania hudobných diel. Hlavnými ukazovateľmi žánru hudobného diela sú akustické vlastnosti zvuku a téma textu. Momentálne existuje veľké množstvo hudobných štýlov a spôsobov zaradenia, avšak chýba určitá štandardizácia. Následkom toho či už automatické určovanie alebo určovanie bežným človekom dosahuje asi presnosť 70% [**?**].Tak isto z hľadiska kompozície, každá časť hudobného diela môže obsahovať iný hudobný žáner[**?**].

2.2 Získavanie používateľskej spätnej väzby

Aby sme mohli presne určiť či daný hudobný dokument vyhovuje používateľovi, je potrebné nejakým spôsobom získať jeho spätnú väzbu. V zásade existujú dva spôsoby získavania spätnej väzby:

- explicitná (používateľ vedome poskytne spätnú väzbu napr. hodnotenie dokumentu),
- implicitná (používateľ o tejto spätnej väzbe nevie, používajú sa agenti ktorí ho monitorujú napr. počítanie času stráveného na stránke),

Identifikácia používateľa

Prvým krokom pri získavaní spätnej väzby je identifikácia používateľa, najpoužívanejší spôsob identifikácie používateľa je pomocou prihlásenia, kedy používateľ pred použitím systému identifikuje sám seba podľa mena a hesla. Tento spôsob sa radi medzi spôsoby ktoré vyžadujú zásah používateľa.

Ďalšou alternatívou v tomto smere je softvérový agent, ktorého si používateľ nainštaluje u seba na počítači. Nevýhodou oproti predchádzajúcemu prístupu je že používateľ musí agenta nainštalovať na každom zariadení ktoré používa.

Alternatívy ktoré nevyžadujú používateľov zásah sú pomocou súborov cookie a pomocou relácií. Obidve tieto alternatívy trpia tím že ak sa používateľ pristupuje z iného zariadenia nebudú ho vedieť identifikovať[**?**].

	Používateľ 0	Používateľ 1	Používateľ 2
Kľúčové slovo 0	0	0	1
Kľúčové slovo 1	0	1	0
Kľúčové slovo 2	0	1	0

Obr. 1: *Ukážka profilu kľúčových slov.*

2.3 Profil používateľa

Rozlišujeme niekoľko druhov používateľských profilov, základe delenie je minimálny (angl. core) a rozšírený (angl. extended) profil. Minimálny používateľsky profil obsahuje čisto informácie o používateľových preferenciách, zatiaľ čo rozšírený používateľský profil obsahuje aj demografické informácie (vek, rodná krajina, vzdelanie, schopnosti atď.,)[**?**]

Model profilu používateľa

Je viacero spôsobov ako uložiť profil používateľa, každý má svoje špecifiká a umožňuje iným spôsobom vykonávať odhadovanie používateľových záujmov.

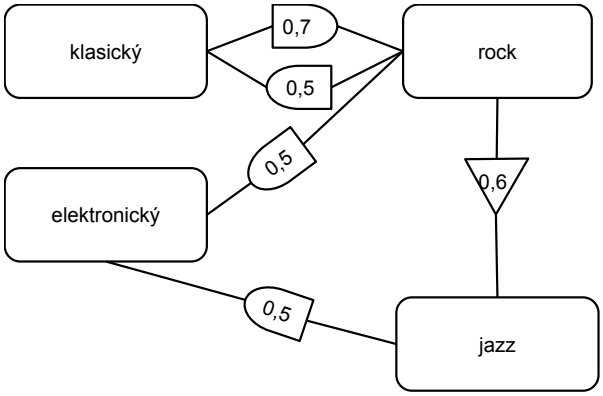
Profil kľúčových slov(angl. keywords profiles)

Profil kľúčových slov je matica o dimenziách používatelia x kľúčové slová. Na základe toho aké jednotlivé súradnice matice môžu dosahovať hodnoty 0 ak dané kľúčové slovo patrí do používateľovho profilu a 0 ak nie. Na obrázku 1. môžeme vidieť príklad v ktorom máme používateľa Používateľ 0 ktorý nemá žiadne kľúčové slová, následne Používateľa 1 ktorý používa kľúčové slová 1 a 2 a Používateľa 2 ktorý používa kľúčové slovo 1.

Profil sémantickej siete (angl. Semantic network profile)

Tento tip profilu sa používa najmä v systémoch používajúcich rozširovanie používateľských vyhľadávacích reťazcov. Základom tohoto profilu je sémantická sieť ktorej príklad môžeme vidieť na obrázku 2.

Sémantická sieť je orientovaný graf v ktorom vrcholmi sú preferencie alebo



Obr. 2: *Ukážka sémantickej siete*

vlastnosti dokumentov, zatiaľ čo hrany sú ich vzťahy. Vzťahy môžu byť niekoľkých typov:

- konjunkcia,
- disjunkcia,
- substitúcia,
- negácia,

2.4 Starnutie profilu

Keďže používateľové zaujmi sú dynamické a menia sa v čase, môžeme implementovať niekoľko modelov starnutia preferencií.

Polia-urn model

Mám množinu záujmov $D = (d_0, d_1, \cdots d_n)$, pre každý záujem d_j ukladám počet vybratí daného záujmu s_j. Pravdepodobnosť znovu vybratia záujmu dostanem pomocou:

$$\frac{s_j}{\sum_{i=1}^{n} s_i}$$

Kde n je rovné počtu záujmov. Tento model nezávisí od poradia v akom boli prejavené zaujmi.

Pravdepodobnosť že dosiahnem určitú kombináciu $S = (s_0, s_1, \cdots, s_n)$ je vyjadrená pomocou:

$$\frac{1}{\sum_{i=1}^{n} s_i}$$

[**?**]

Polčas rozpadu

Jednou z možností ako zabezpečiť starnutie používateľského profilu je využiť funkciu exponenciálneho polčasu rozpadu. Táto funkcia sa hlavne využíva pri datovaní veku uhlíka, avšak dá sa použiť aj ako funkcia starnutia používateľových preferencií. Základom tohoto prístupu je následujúci vzorec:

$$N(t) = N_0 e^{-k/t}$$

Aby sme mohli tento algoritmus použiť, musíme si určiť polčas rozpadu preferencie. Následne si na základe polčasu rozpadu vypočítam parameter k, Ten sa dá po odvodení určiť z následujúceho vzorca.

$$k = \frac{\ln \frac{1}{2}}{t_r}$$

[**?**]

Kde t_r je polčas rozpadu. Tento model umožňuje vytvorenie viacerých rýchlosti starnutia pre krátkodobe a dlhodobé záujmy[**?**].

2.5 Geometrické porovnávanie vlastnosti dokumentov z preferenciami

Pri tomto spôsobe sa relevancia dokumentu d_i pre používateľa u_j bude určovať premietnutím dokumentu do priestoru ako vektor $d_j = (p_0, p_1 \cdots p_n)$ kde $p_0, \cdots p_n$ sú vlastnosti dokumentu, následne sa do toho istého priestoru premietne aj používateľ,

pričom dimenzie budú preferencie používateľa $u_i = (p_0, p_1 \cdots p_n)$. Podobnosť týchto vektorov sa následne vyhodnotí pomocou nasledujúceho vzorca:

$$cos\theta = \frac{d_j * q}{||d_j|| * ||q||}$$

Pričom $||d_j||$ a $||q||$ sú normalizované vektory[**?**].

2.6 Kolaboratívne filtrovanie

Kolaboratívne filtrovanie je postup pri ktorom odporúčam používateľovi na základe podobnosti z iným používateľom. Kolaboratívne filtrovanie sa delí na založené na obsahu a kolaboratívne filtrovanie [**?**].

Odporúčanie založené na obsahu vychádza z dostupných informácií o diele, teda z vlastnosti diela zatiaľ čo kolaboratívne filtrovanie záleží čisto od explicitnej spätnej väzby používateľa pomocou hodnotenia.

Podľa [**?**] dosahuje kolaboratívne filtrovanie väčšiu dynamiku. Avšak hrozí problém studeného štartu, teda že novo pridaná položka nebude mať žiadne hodnotenie a preto klesne hneď na spodok odporúčaní. Kolaboratívne filtrovanie môžeme ďalej rozdeliť:

- kolaboratívne filtrovanie založené na pamäti
- kolaboratívne filtrovanie založené na modeli
- hybridné kolaboratívne filtrovanie [**?**]

Kolaboratívne filtrovanie založené na pamäti

Podobnosť medzi používateľmi sa zisťuje na základe hodnotenia dokumentov. Je použitá heuristicka metóda ktorá zisťuje chýbajúce hodnotenia porovnávaním používateľov a následne doplnenie od najpodobnejšieho používateľa.

Kolaboratívne filtrovanie založené na modeli

Pracuje z modelom ktorí vytvára hodnotenie a súčasne sa učí na existujúcich dátach.

Hybridné kolaboratívne filtrovanie

Na obídenie nedostatkov algoritmov kolaboratívneho filtrovania, niektoré aplikácie kombinujú tieto dve metódy.

3 Logistická funkcia

Logisticka funkcia sa často používa ako pravdepodobnostná funkcia. Funkcia má nasledujúci tvar:

$$f(t; a, m, n, \tau) = a * \frac{1 + me^{-t/\tau}}{1 + ne^{-t/\tau}}$$

Vo väčšine prípadov sa používa špecialny prípad tejto funkcie meno signusoida. Signusoida ma $a = 1$, $m = 0$, $n = 0$ a $\tau = 1$ čiže:

$$f(t) = \frac{1}{1 + e^{-t}}$$

Graf tejto funkcie má tvar písmena S

4 Návrh riešenia

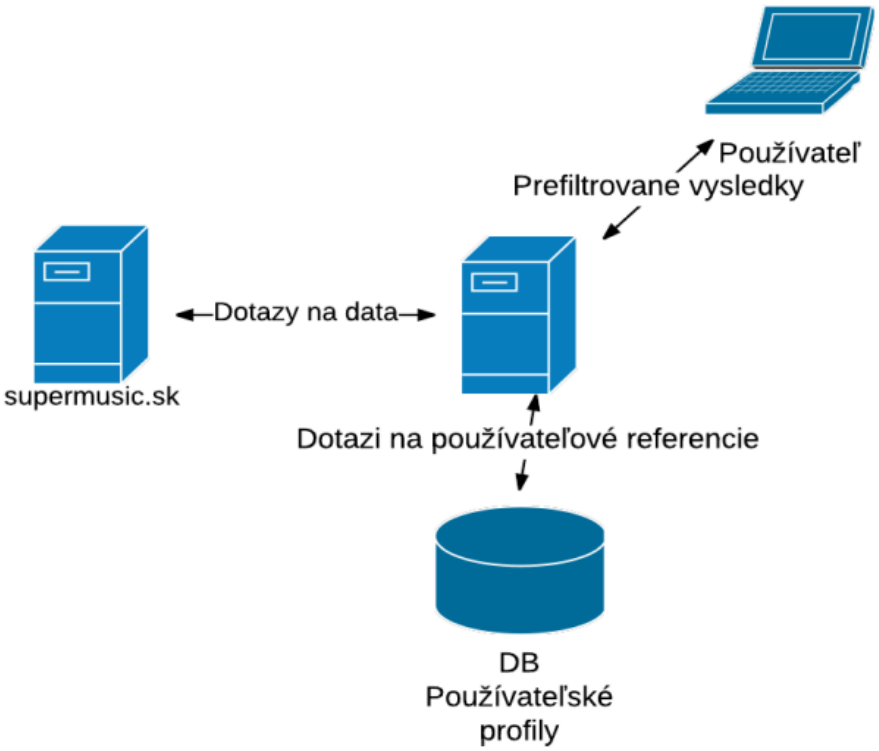
Implementácia je postavená okolo dvoch základných funkcionalít, ktoré sa budú navzájom dopĺňať, pôjde vyhľadávanie v hudobných dokumentoch ktorého hlavnou úlohou bude naplniť používateľské profily preferenciami a zostavovanie spevníkov na základe týchto preferencií.

4.1 Vyhľadávanie hudobných dokumentov

Vyhľadávanie bude pracovať nad už existujúcou databázou hudobných dokumentov. V podstate bude využívať už funkčné vyhľadávanie na tejto stránke, akurát na základe používateľových preferencií rozšíri jeho vyhľadávacie reťazce o ďalšie slová ktoré presnejšie špecifikujú používateľov zámer.

4.2 Zostavenie spevníka

Aplikácia bude podporovať funkcionalitu automatického generovania spevníka, kedy si používateľ zvoli používateľov s ktorými si chce ísť zahrať a aplikácia automatický vygeneruje spevník zložený s nejpreferovanejších hudobných diel daných používateľov.



Obr. 3: *Náčrt funkčnosti aplikácie.*