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 [7]:

- 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[11],

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% [5].Tak isto z hľadiska kompozície, každá časť hudobného diela môže obsahovať iný hudobný žáner[10].

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[3].

	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,)[3]

Model profilu používateľa

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

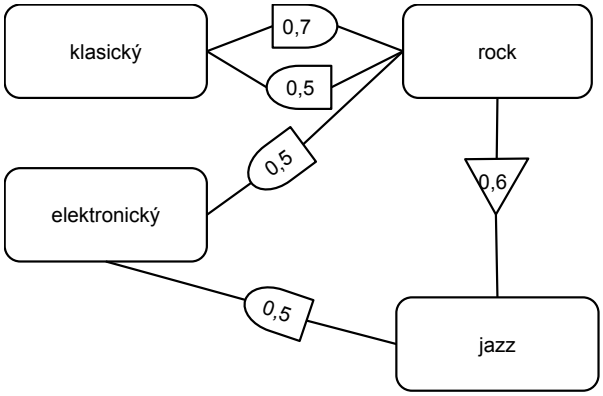
Profil kľúčových slov(angl. keywords profiles)

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:

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

[15]

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.

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

[14]

Kde t_r je polčas rozpadu. Tento model umožňuje vytvorenie viacerých rýchlosti starnutia pre krátkodobe a dlhodobé záujmy[2].

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[6].

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 [9].

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 [9] 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 [8]

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 Váhovanie značiek

3.1 Frequencia pojmov (angl. Term Frequency(TF))

Toto je jeden z najjednoduchších a najstarších prístupov k váhovaniu pojmov, pri tomto prístupe sa jednoducho počíta počet výskytov pojmu v dokumente. Existuje niekoľko druhov tohoto váhovania.

Binárne váhovanie

znamená že nepočítam počet výskytov slova v texte ale beriem iba či sa v texte nachádza alebo nie, Teda:

$$w_{TF_BIN}(t_i) = \begin{cases} 1 & \text{ak } t_i \in d_j \\ 0 & \text{ak } t_i \notin d_j \end{cases}$$

Kde d_j je dokument a t_i je slovo.

Čistá frequencia

sa dá tiež použiť, v tomto prípade sa váha slova určuje podľa počtu jeho výskytov v texte. Teda:

$$w_{TF_RAW}(t_i) = t_{i_{d_j}}$$

Kde $t_{i_{d_j}}$ je počet výskytov slova t_i v dokumente d_j.

Logaritmická váha

sa taktiež používa najmä kôly tomu, že relevancia dokumentu nerastie proporcionalne z počtom výskytov slova v dokumente.

$$w_{TF_LOG}(t_i) = \begin{cases} 1 + \log 10 t_{i_{d_j}} \text{ ak } t_{i_{d_j}} > 0 \\ 0 \qquad\qquad\qquad \text{inak} \end{cases}$$

Existuje ešte viac spôsobov ako sa dá vyhodnotiť frequencia pojmov, medzid ktoré patrí napríklad dvojita normalizácia 0.5 (angl. double normalization 0.5) alebo k-dvojita normalizácia (angl. double normalization K) [12]

3.2 Frequencie pojmov, inverzna frequencia dokumentov (angl. Term Frequency, Inverse Document Frequency (TF*IDF))

Pri tomto prístupe zahrňujeme do váhovania aj to v akom počte dokumentov sa daný pojem nachádza, ak sa pojem vyskytuje príliš často, znížime jeho váhu. Toto znižovanie je reprezentované inverznou frequenciou pojmu:

$$idf_i = log10\frac{N}{dt_i}$$

Kde dt_i je počet dokumentov v ktorý sa pojem t_i nacháza.

Výsledna rovnica je potom:

$$w_{TF*IDF} = w_{TF} * log10\frac{N}{dt_i}$$

[12]

3.3 Presonalizované BM25 váhovanie

Tento model je jeden zo štatystických modelov. Tu uvedená verzia je jeho modifikácia podľa [4]

$$w_{BM25}(t_i) = \log \frac{(r_{t_i} + 0.5)(N - n_{t_i} + 0.5)}{(n_{t_i} + 0.5)(R - r_{t_i} + 0.5)}$$

,

kde N je počet všetkých dokumentov, n_{t_i} je počet dokumentov obsahujúcich pojem t_i, R je počet dokumentov ktoré používateľ už navštívil a r_{t_i} je počet dokumentov ktoré už používateľ navštíciľ obsahujúcich pojem t_i

4 Určovanie adekvátnych dokumentov

Párovanie na základe unikátnych pojmov

Párovanie na základe váh

spočíva v sčítaní váh slov ktoré sú vyhľadávané a zároveň sú v dokumente:

$$w_{d_i} = \sum_{z=1}^{N_{d_t}} f_{t_z} * w(t_z)$$

Jazykový model(angl. Language Model)

5 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

6 Starnutie záujmov profilu

6.1 Krátkodobé záujmy

6.2 Dlhodobé záujmy

7 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.

7.1 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.

7.2 Zostavenie spevníka

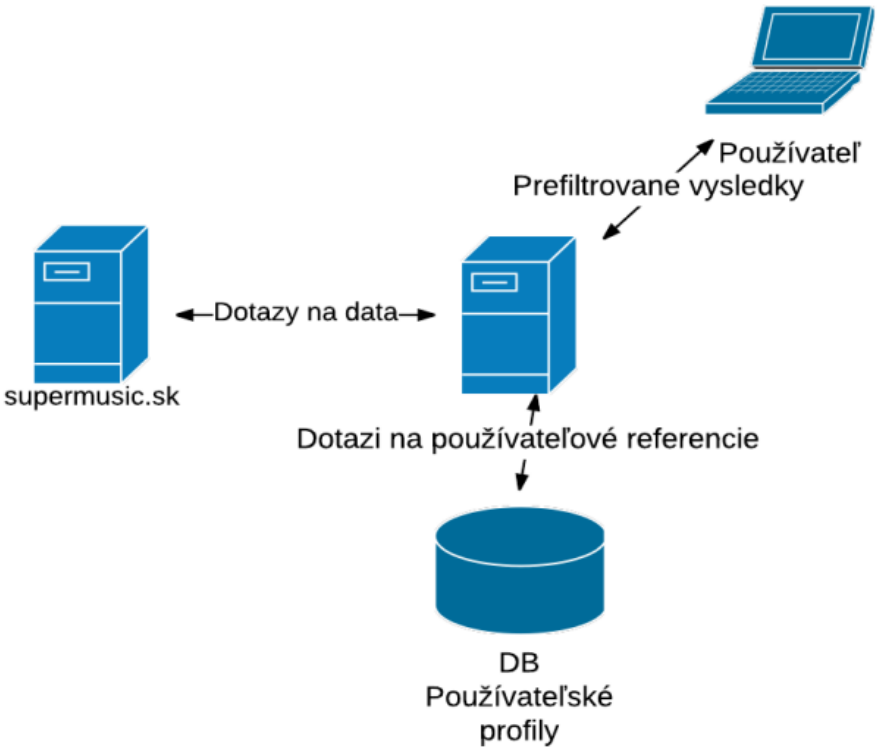
Aplikácia bude podporovať funkcionalitu automatického generovania spevníka, kedy si používateľ zvoli používateľov s ktorými si chce ísť zahrať a aplikácia automatický vygeneruje spevník zložený s nejpreferovanejších hudobných diel daných používateľov.

7.3 Krawler ?

Tento komponenet prehľadáva databázu ktorá je cieľom môjho odporúčača, využíva k tomu abecedne zobrazenie záznamov databázy. Databáza sa nedá zobraziť od do, takže granularitu zobrazenie stránok som musel určiť pokusom, najskôr som si zobrazoval všetky troj písmenkove názvy, čo bolo 30*26*26 zobrazení (20280), čo ale trvalo príliš dlho, tak som v tretej sade prehľadával iba každé štvrté písmenko, čo zredukovalo počet stranok na 3380.

7.4 Indexing

Jestvuje veľa spôsobov ako sa dá označovať a vyhľadať obsah, ja som sa počas prieskumu zameral na try:



Obr. 3: *Náčrt funkčnosti aplikácie.*

Priama tagovacia tabuľka

Vytvoril som tabuľku tagov, kde bol každý tag fyzický priamo vložený spolu z id dokumentu ku ktoremu sa viaže, tento prístup ale nebol dostatočne rýchli na vygenerovanie, ani na vyhľadávanie. Pri vyhľadávaní nad 118989 značkamy označujúcimi 47002 dokumentov zabral 44.6984 sekúnd. Nepomohlo ani zindexovanie podľa mena.

Model vektorovho priestoru (angl. vector space model

Pri použití tohto modelu zabral dotaz 0.006 sec.

Tento model sa v MySQL nazýva model prírodzeného jazyka (angl. Natural Language Model), ktorý porovnáva vlastnosti dokumentov na základe abstrakcie priestoru, v ktorom sú jednou dimenziou vlastnosti jedného dokuemntu a druhou vlastností druhého dokumentu, prípade vyhľadávacieho reťazca alebo používateľský profil. Následne sa vracajú dokumenty ktoré majú najpodobnejší smer vektora k požadovanej fráze.

V MySQL je tento prístup implementovaný pomocov nasledujúcej rovnice [1]:

$$w_d = \frac{\log(dtf_d) + 1}{\sum_{i=1}^{t} \log(dtf_i) + 1} * \frac{U}{1 + 0.0115 * U} * \log \frac{N}{nf}$$

- dtf_d je sila (množstvo koľko krát sa nachádza pojem v text v prípade analízy textu) vlastnosti vyhodnocovaného dokumentu
- dtf_i sila i-tej vlastnosti
- U počet unikátnych vlastnosti dokumentu
- N počet všetkých dokumentov
- nf je počet dokumentov ktoré obsahuje danú vlastnosť

Rovnica sa dá rozdeliť na tri časti.

Základná časť

Je to primárna rovnica určujúca váhu pojmu.

Normalizačný faktor

Spôsobý, že ak je dokument kratší ako preiemerná dĺžka dokuemntu, jeho relevancia stúpa. [13]

Inverzná frequencia

Zabezpečuje že menej časté pojmy majú vyššiu váhu.

7.5 Filtrovanie bezvýznamných značiek (angl. stopwords)

Niektoré slová sú pri vyhľadávaní a indexovaní zbytočné. Síce sa dá použiť tf*idf ktorý redukuje váhu slov na základe ich unikátnosti, ale tieto slová aj tak musí systém spracovať, ja som sa rozhodol použiť kombináciu českých, anglických a slovenských slóv z projektu TODO: Ako ? google code stop-words

Literatúra

[1] *Mysql developers documentation, 10.7 Full-Text Search*.

[2] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. august 2011.

[3] Peter Brusilovsky. User profiles for personalized information access.

[4] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. 2002.

[5] Karin Kosina. Music genre recognition. 2002.

[6] DIK L. LEE. Document ranking and the vector-space model. 1997.

[7] Namunu C. Maddage, Li Haithou, and Mohan S. Kankanhalli. Music structure analysis statistics for popular songs. November 2009.

[8] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. júl 2002.

[9] Gašpar Peter. Odporúčanie s využitím osobných vyjadrení. 2014.

[10] Yves Raimond, Samer Abdallah, and Mark Sandler. The music ontology. 2007.

[11] Arnold Schoenberg. *Fundamentalls of musical composition*. Faber and Faber Limited, 1967.

[12] Hinrich Sch¨utze. Introduction to information retrieval. 2011.

[13] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. 1996.

[14] Kiryl Tsishchanka. Exponential growth and decay. 2010.

[15] Tong Zhu. Nonlinear p´olya urn models and self-organizing processes. 2009.