

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 Návrát, PhD.

Máj 2015

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

Študijný program: Informatika

Autor: Martin Černák

Bakalárska práca: Dynamické odporúčanie

Vedúci práce: prof. Ing. Pavol Návrát, PhD.

Máj 2015

Dynamické odporúčanie v kontexte hudobných dokumentov je vďaka svojmu úzkemu zameraniu 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 hudobnom odvetví môžeme častejšie ako v ostatných vidieť príchod mimoriadne populárnych nových interpretov, 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 disappear 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 účastníkom 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

1	Úvod	1
1.1	Použité pojmy a skratky	2
2	Existujúce riešenia	3
2.1	Odporúčanie hudobných dokumentov	3
2.2	Rôzne prístupy k odporúčaniam	5
2.3	Používateľské profily	7
2.4	Váhovanie značiek	12
2.5	Evolúcia používateľských preferencií	14
3	Návrh aplikácie	16
3.1	Platforma	16
3.2	Základné rozvrhnutie aplikácie	17
3.3	Špecifikácia služieb	18
4	Implementácia	19
4.1	Datový model	19
4.2	Indexovanie dokumentov a váhovanie	20
4.3	Filtrovanie bezvýznamných značiek (angl. stopwords)	23
4.4	Zostavenie spevníka	23
4.5	Krawler ?	23
4.6	Indexing	23
4.7	Váhovanie Dokumentu	24
	Literatúra	25

Zoznam obrázkov

1	Ukážka sémantickej siete	9
2	Ukážka bayesovej siete	10
3	Ukážka odporúčacej bayesovej siete	11
4	Štruktúra aplikácie	18
5	Zjednodušený datový model	20
6	BPMN Digram kravleru	21

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 histórii, problém nastáva ak chceme nejaké údaje vyhládať. 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 nezaujíma.

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

sedenie - Sekvencia zobrazení dokumentov ktorá je časovo.

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

dokument - Je jedno hudobné dielo reprezentované tabuľkou, textom, notami, prekladom alebo iným spôsobom nápomocným k prevedeniu hudobného diela.

vlastnosť dokumentu - Špecifický črt dokumentu ktorý môže ovplyvniť používateľové preferencie.

značka dokumentu - Otnačená vlastnosť dokumentu ktorá je rozoznávana vyhľadávacím systémom.

preferencia - Je vlastnosť dokumentu ktorú nejakým spôsobom používateľ preferuje.

užitočnosť - Vlastnosť je hodnota určujúca či je daný dokument preferovaný používateľom.

2 Existujúce riešenia

Počas analýzy som našiel niekoľko riešení problému dynamického odporúčania. Každé z týchto riešení poskytuje rôzne výhody a nevýhody pri rôznych oblastiach nasadenia. To ktorý model je najvhodnejší pre nás je ovplyvnené cieľovým artiklom a cieľovou skupinou. Pre potreby tejto práce bolo potrebné najskôr vybrať doménu, v ktorej chcem riešenie implementovať. Následne zanalizovať prístupy v danej oblasti.

2.1 Odporúčanie hudobných dokumentov

Pre potreby tejto práce som zvolil oblasť hudby, avšak nie v úplne klasickom ponímaní, zameral som sa na dokumenty umožňujúce hudobníkom naučiť sa hrať určité hudobné dielo. Za základne vlastnosti hudby sa považuje rytmus, sila a farba tónu. Na zaznamenanie týchto vlastností vzniklo viacero zápisov podľa potrieb určitých skupín hudobníkov. Odporúčanie v tejto oblasti je pomerne nové a preto sa budem skôr snažiť najst' riešenia z iných oblastí a preskúmať ich aplikovateľnosť v tejto oblasti.

Na presné odporúčanie akéhokoľvek článku, musíme najskôr najst' nejaké jeho vlastnosti na základe ktorých môžeme odporúčať. Keďže táto doména je úzko spojená z hudbou, budem sa snažiť vychádzať z nej.

Najznámejším spôsobom kategorizovania hudby je rozdelenie na žánre a podžánre. Problém pri žánroch a podžánroch je, že neexistuje jednotná definícia ani spôsob ich kategorizácie. Určovanie žánrov má nasledujúce typy pravidiel

- **formálne a technické pravidlá aplikované na obsah(sila, výška a farba tónu),**
- **semiotické pravidla** (abstraktný vopred dohodnutý koncept, napríklad politická situácia),
- **pravidla správania sa** (črty správania sa fanúčikov alebo interpretov daného žánru),

- **ekonomické a jurisdikčné pravidlá** (zákonné a právne aspekty ktoré daný žáner podporujú, napríklad český protestsong¹).

Tieto pravidlá definoval Franco Fabbri[4].

Pravidlá sú pomerne abstraktné a i napriek viacerím pokusom o vytvorenie kompletne ontológie žánrov či už z akademický alebo komerčných kruhov², neexistuje jednotná ontológia hudobných žánrov.

Pre potreby odporúčania by bolo najvhodnejšie automatické určovanie žánru, ako napríklad navrhli autori článku [9], kde analyzovali výšku a snažili sa odhaliť akcent nôt, čo im umožnilo odhaliť rytmus piesne. Následne sa zamerali na určenie jednotlivých častí hudobného diela ako predohra, hlavná časť, refrén a sloha.

Ďalším prístupom je nechať označovať vlastnosti dokumentu používateľom, tento prístup používa napríklad služba last.fm³, ktorá sa následne snaží používať najpopulárnejšie značky ako žánre. Tento prístup je bližšie opísaný v článku Paul Lamera[7] a je známy pod názvom socialne značenie (angl. social tagging).

Podobnosť odporúčania hudobných dokumentov a hudby

I keď sú tieto dve domény veľmi podobné, existujú rozdiely. Jeden z rozdielov sú ich vlastnosti dokumentov. Na presnú identifikáciu pesničky nám stačí názov piesne, autor piesne, interpret a prevedenie. Pri hudobných dokumentoch sa môžu okrem týchto vlastností líšiť aj typy dokumentov (taby, akordy, text, preklad), prípadne niektoré dokumenty môžu obsahovať iba časť daného hudobného diela (predohra, medzi-hra, refrén, sólo atď.).

Podobnosť z odporúčaním textu

Ďalšia doména ktorú je možné využiť je odporúčanie textových dokumentov. To je najčastejšie založené na analýze výskytu slov v texte. Avšak jediné typy dokumentov ktoré by sa dali takto analyzovať sú preklady a texty.

¹ www.wikipedia.sk

² <https://www.apple.com/itunes/affiliates/resources/documentation/genre-mapping.html>

³ <http://www.last.fm/home>

2.2 Rôzne prístupy k odporúčaniu

Hlavný účelom odporúčacích systémov je odhadnúť užitočnosť dokumentu pre používateľa [13], pričom v mnohých prípadoch je potrebné užitočnosť dokumentu odhadovať. Užitočnosť ako taká môže závisieť od veľkého množstva parametrov, skupina týchto parametrov sa všeobecne nazýva kontextové premenné. Na základe toho ako systém nakladá z danými údajmi, delím odporúčacie systémy na niekoľko kategórií. Hlavnou charakteristikou systému je **funkcia užitočnosti**.

Filtrovanie na základe obsahu

Pri tomto prístupe odporúčame používateľovi dokumenty podobné tým čo sa mu páčili v minulosti. Funkciou užitočnosti dokumentu je teda podobnosť dokumentu z už zobrazenými dokumentmi používateľa. Podobnosť dokumentov sa zisťuje pomocou porovnávania značiek dokumentov.

Kolaboratívne filtrovanie

Toto je najpopulárnejší spôsob implementácie odporúčania. Najjednoduchšia a originálna implementácia odporúča aktívnemu používateľovi dokumenty ktoré sa páčili ľuďom z podobným vkusom. Podobnosť používateľov je založená na histórii hodnotenia dokumentov používateľov. Takže úlohou funkcie užitočnosti je najst' najpodobnejších používateľov a vrátiť dokument ktorý mal najkľadnejšie hodnotenia od týchto používateľov.

Najväčším problémom kolaboratívneho filtrovania, tzv. problém studeného štartu. Spočíva v tom že ak pribudne do zbierky nový dokument, nemá ešte žiadne hodnotenia, takže sa nebude nikomu odporúčať.

Demograficke odporúčanie

Tieto odporúčacie systémy odporúčajú používateľovi dokumenty na základe jeho demografického profilu (vek, národnosť, jazyk atď.). Za istú formu tohoto odporúčania môžeme považovať multijazyčnosť dnešných stránok. Zaujímavým príkladom je aj domovská stránka google, ktorá sa vo významné dni zobrazuje v

rôznych krajinách rôzne¹.

Znalostne odporúčanie

Znalostné odporúčacie systémy odporúčajú na základe znalostí o tom ako nejaká vlastnosť dokumentu ovplyvňuje užitočnosť dokumentu pre používateľa. V princípe ide o systém ktorý dáva používateľovy otázky a na základe zistených faktov mu odporučí dokument. Takéto odporúčanie sa najčastejšie využíva pri zákazníkovej podpore, veľmi dobrým príkladom na tento prístup je zákaznícka podpora Microsoftu².

Komunitné odporúčanie

Toto odporúčanie je veľmi podobné s kolaboratívnym filtrovaním, avšak na rozdiel od neho upradnostňuje implicitne dané priateľstvá medzi používateľmi. Tento druh odporúčania zažíva rozkvet najmä v poslednej dobe spolu s rozkvetom používania sociálnych sietí. V dokumente[13] sa dokonca uvádza že v špeciálnych prípadoch sú efektívnejšie ako kolaboratívne filtrovanie. Tento druh odporúčania sa často nazýva aj sociálne odporúčanie. Funkcia užitočnosti v tomto prípade najskôr zistí vzťahy medzi používateľmi a preferencie priateľov používateľa a na základe ich preferencií určí užitočnosť dokumentov pre používateľa.

Hybridné odporúčanie

Toto odporúčanie kombinuje vlastnosti predchádzajúcich prístupov na vyriešenie ich vzájomných problémov. Napríklad časté riešenie je kombinovanie kolaboratívneho filtrovania s filtrovaním založeným na obsahu, kedy v podstate filtrovanie na základe obsahu rieši problém studeného štartu pre kolaboratívne filtrovanie. Tak isto sa v poslednej dobe zvikne kombinovať kolaboratívne filtrovanie s komunitným odporúčaním vďaka ich dobrým výsledkom.

¹<http://www.google.com/doodles/>

²<https://support.microsoft.com/sk-sk>

Spätná väzba

Aby odporúčacie systémy mali ako odporúčať, potrebujú zistiť preferované vlastnosti dokumentov zo korpusu odporúčaných dokumentov. Z tohto dôvodu je potrebné nejakým spôsobom zistiť ktoré vlastnosti používateľ preferuje. Na toto zisťovanie slúži spätná väzba. Spätnú väzbu v zásade delíme na dve skupiny,

- **implicitná spätná väzba** (je získavanie spätnej väzby používateľ a z jeho akcií ktoré nesúvisia priamo z hodnotením, napríklad stiahnutie dokumentu, prípadne jeho vytlačenie, hlavnou výhodou je že nevyžaduje vedomí zásah používateľ a, avšak zvyšuje technické nároky na systém)
- **explicitná spätná väzba** (je vedome ponúknutie spätnej väzby od používateľ a, napríklad ak používateľ označí že sa mu dokument páči, podľa [2] zvykne byť efektívnejšia, avšak vyžaduje vedomý zásah používateľ a).

Explicitnú spätnú väzbu môžeme ďalej deliť na základe toho, akú hodnotu nám vracia napríklad na **binárnu**, „páči sa mi“ a „nepáči sa mi“, alebo **z hodnotu** (napríklad pridelenie 1 až 5 hviezdčiek). S týmto priamo súvisí aj konštrukcia používateľských profilov.

2.3 Používateľské profily

Ďalšou dôležitou súčasťou odporúčacích systémov je spôsob akým konštruujú používateľské profily. Rôzne druhy profilov umožňujú použitie rôznych algoritmov avšak môžu mať rôzny dopad na pamäťovú či výkonovú stránku systému[2].

Binárny vektor

V tomto prípade sú preferencie reprezentované vektorom v dvojrozmernom priestore kde jeden rozmer sú značky dokumentov a druhý sú používatelia. Vektor je tvorený binárnou hodnotou kde 1 na pozícii p_{ij} , pri predpoklade že j je identifikátor j -teho používateľ a i je identifikátor i -tej značky, znamená že i -ta značka je preferencia používateľ a j . V tabuľke 1 môžeme vidieť príklad v ktorom máme používateľ a p_0 ktorý nepreferuje žiadne značky, následne Používateľ a p_1 ktorý preferuje značky z_1 , z_2 a Používateľ a p_2 ktorý preferuje značku z_0 .

z	p_0	p_1	p_2
z_0	0	0	1
z_1	0	1	0
z_2	0	1	0

Tabuľka 1: Ukážka modelu profilu ako binárneho vektora

Vahovaný vektor

Tento profil je veľmi podobný z predchádzajúcim profilom, avšak namiesto bynárnych hodnôt sú hodnotami súradníc užitočnosti daných preferencií 2.

z	p_0	p_1	p_2
z_0	0	0	2
z_1	0	2	0
z_2	0	5	1

Tabuľka 2: Ukážka modelu profilu ako váhovaného vektora.

Trojrozmerný vektor

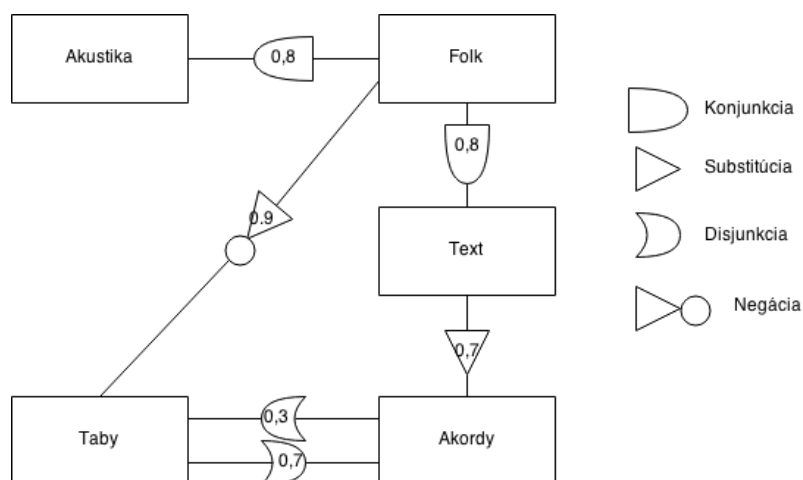
V prípade že vieme aj rozdeliť značky do domén alebo kontextov, môžeme do vektorového profilu pridať ešte jednu súradnicu ktorá reprezentuje doménu danej značky. Napríklad ak odporúčame text a značky sú reprezentované slovami ktoré sa našli v texte, tak značke ktorá pochádza z nadpisu môžeme automaticky priradiť väčšiu hodnotu.

Profil sémantickej siete

Profil sémantickej siete (angl. Semantic network profile) je semantická sieť ktorá je vybudovaná pre konkrétneho používateľa a vyjadruje vzťahy medzi značkami ktoré používateľ preferuje.

Sémantická sieť [17] je orientovaný graf v ktorom sú vrcholmi značky, zatiaľ čo hrany sú ich vzťahy. napríklad v [2] sú použité vzťahy typov.

- konjunkcia,



Obr. 1: Ukážka sémantickej siete

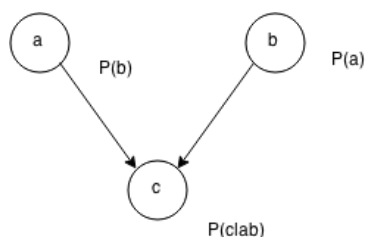
- disjunkcia,
- substitúcia,
- negácia.

Takéto sémantické siete sa používajú najmä pri dopĺňaní slov do vyhľadávacích fráz, kde v prípade krátkej vyhľadávacej frázy môžeme zredkovať jej viacznačnosť doplnením slov ktoré majú v sémantickej sieti najsilnejšiu pozitívnu väzbu. Na obrázku 1 môžeme vidieť príklad takejto siete.

Teda ak by mi prišiel od používateľa výraz *folk*, tak môžem výraz rozšíriť slovami *akustika* a *text*. Keď budem pokračovať, slovo *text* sa dá nahradiť slovom *akordy* (keďže v podstate *akordy* sú *text* doplnení o skratky *akordov*). Následne však *taby* už nemôžem doplniť aj keď majú disjunktný vzťah z *akordami*, pretože majú silnú negatívnu väzbu z *folk*om. Teda výsledný výraz by bol *folk a akustika akordy*.

Bayesova sieť

Ďalšou možnosťou ako ukladať používateľské dáta je bayesová sieť. Bayesová sieť vychádza z bayesovej teóremy, o ktorej využití v kontexte odporúčania sa budeme zaoberať neskôr. Bayesová sieť slúži na výpočet pravdepodobnosti hypotézy



Obr. 2: Ukážka bayesovej siete

pri zmene jej evidencie. Jej vrcholmy sú latentné premenné (závisia od ostatných premenných) a jej hranamy ich vzťahy. Takže v prípade že sa mi zmení hodnota jednej premennej v grafe, po hranách viem upraviť hodnoty všetkých premenných ktoré od nej závisia.

Napríklad na obrázku 2 môžeme vydiť bayesovú sieť zloženú z troch premenných a , b a c . Tieto sú vo vzájomnom vzťahu. V prípade zmeny hodnoty a alebo b sa hodnota c automaticky prepočíta.

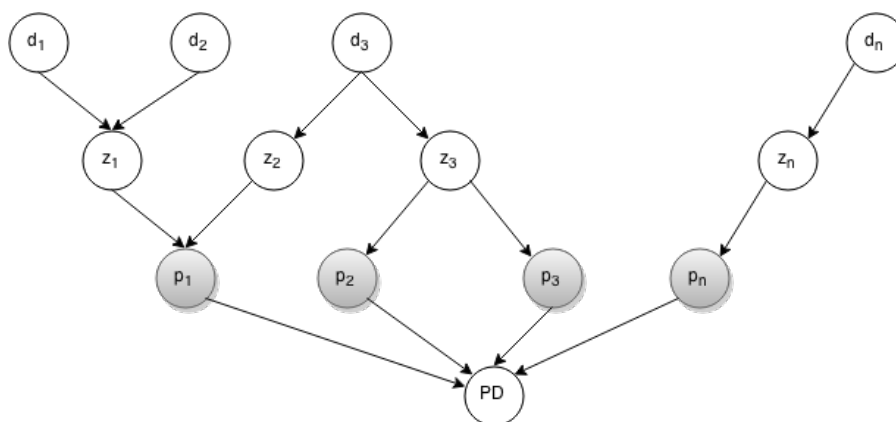
Ako príklad ako využiť takúto sieť môžem uviesť zjednodušený príklad z prezentácie [5], kde najhornejšie vrcholy (tie ktoré nezávisia od iných premenných) boli dokumenty $D = d_1, d_2 \dots d_n$ a pod nimi boli ich značky $Z = z_1, z_2 \dots d_n$, následne ak sme chceli odporúčať (príklad je aplikovaný na vyhľadávací dotaz, avšak dotaz môže byť nahradený používateľským profilom), na najspodnejšie body sme pripojili preferencie používateľa $P = p_1, p_2 \dots p_n$. Náčrt siete môžeme vydiť na obrázku 3.

Pričom značky a dokumenty si môžem držať predpočítané v pamäti, a tak isto aj používateľské preferencie. V prípade odporúčania ich iba poprepájam.

Bayesov theorem

Bayesová teoréma sa zaoberá vplyvom nových poznatkov na existujúce domienky o určitej hypotéze. Vďaka nej vieme kombinovať nové dáta z existujúcimi poznatkami. Matematicky je táto teoréma vyjadrená v kontexte vyhľadávania pomocou rovnice 1 a rovnice 2.

$$P(P|d) = \frac{P(d|P) * P(P)}{P(d)} \quad (1)$$



Obr. 3: Ukážka odporúčacej bayesovej siete

$$P(NP|d) = \frac{P(d|NP) * P(NP)}{P(d)} \quad (2)$$

Rovnice obsahujú,

- pravdepodobnosť že dokument je užitočný P ,
- pravdepodobnosť že dokument nie je užitočný NP ,
- pravdepodobnosť že vrátený dokument d je užitočný $P(P|d)$,
- $P(d|P)$ pravdepodobnosť že ak je vrátený užitočný dokument, je to dokument d ,
- pravdepodobnosť vrátenia užitočného dokumentu $P(P)$,
- pravdepodobnosť výberu dokumentu $P(d)$,
- pravdepodobnosť neužitočnosti dokumentu $P(NP|d)$,
- pravdepodobnosť $P(d|NP)$ že ak je vrátený neužitočný dokument, je to dokument d ,
- $P(NP)$ je pravdepodobnosť že je vrátený neužitočný dokument,
- pravdepodobnosť $P(d)$ vrátenia dokumentu d .

To či je dokumentu užitočný sa následne určuje tým, či je $P(P|d)$ väčšie ako $P(NP|d)$.

2.4 Váhovanie značiek

Ak mám dokument alebo profil reprezentovaný značkami, je potrebné zistiť ako veľmi sú tieto značky preferované. Nie všetky značky majú rovnakú váhu, takže potrebujeme dosiahnuť aby užitočnosť dokumentu závisela od unikátnosti značiek v ňom. Základy váhovania sú popísané v prezentácii Hinrich Schütze [15].

Frekvencia pojmov

Frekvencia pojmov (angl. Term Frequency ďalej TF), je jeden z najjednoduchších a najstarších prístupov k váhovaniu značiek, pri tomto prístupe sa jednoducho počíta počet výskytov značiek v dokumente. Existuje niekoľko druhov tohoto váhovania.

Binárne váhovanie znamená že napríklad, nepočítam počet výskytov slova v texte ale beriem iba či sa v texte nachádza alebo nie. Matematicky vyjadrené rovnicou ??.

$$w_{TFBIN}(t_i) = \begin{cases} 1 & \text{ak } t_i \in d_j \\ 0 & \text{ak } t_i \notin d_j \end{cases} \quad (3)$$

Kde d_j je dokument a t_i je slovo.

Čistá frekvencia sa dá tiež použiť, v tomto prípade sa váha slova určuje podľa počtu jeho výskytov v texte. Matematicky vyjadrené rovnicou 4

$$w_{TFRAW}(t_i) = t_{i_{d_j}} \quad (4)$$

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

Logaritmická váha sa taktiež používa najmä kôľ tomu, že relevancia dokumentu nerastie proporcionálne z počtom výskytov slova v dokumente. Toto môžem matematicky vyjadriť napríklad rovnicou ??.

$$w_{TFLOG}(t_i) = \begin{cases} 1 + \log 10 t_{i_{d_j}} & \text{ak } t_{i_{d_j}} > 0 \\ 0 & \text{inak} \end{cases} \quad (5)$$

Existuje ešte viac spôsobov ako sa dá vyhodnotiť frekvencia pojmov, medzi ktoré patrí napríklad dvojita normalizácia 0.5 (angl. double normalization 0.5) alebo k-dvojita normalizácia (angl. double normalization K) [15].

Frekvencie pojmov, inverzná frekvencia dokumentov

Frekvencia pojmov, inverzná frekvencia pojmov (angl. Term Frequency, Inverse Document Frequency d'alej TF*IDF) je prístup, pri ktorom zahrňujeme do užitočnosti aj počet dokumentov, ktoré majú danú značku. Základom je zníženie užitočnosti často sa vyskytujúcim značkám. Toto znižovanie je reprezentované rovnicou 6 d'alej IDF.

$$idf_i = \log_{10} \frac{N}{dt_i} \quad (6)$$

Kde dt_i je počet dokumentov v ktorých sa pojem t_i nachádza.

Výsledná rovnica TF*IDF je rovnica 7 ktorá je vlastne súčin TF a IDF.

$$w_{TF*IDF} = w_{TF} * \log_{10} \frac{N}{dt_i} \quad (7)$$

Presonalizované BM25 váhovanie

Tento model je jeden zo štatistických modelov. Tu uvedená verzia je jeho modifikácia podľa S. Cronen a spol. [3] ktorá je matematicky reprezentovaná rovnicou 8.

$$w_{BM25}(t_i) = \log \frac{(r_{t_i} + 0.5)(N - n_{t_i} + 0.5)}{(n_{t_i} + 0.5)(R - r_{t_i} + 0.5)} \quad (8)$$

,

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 navštívil obsahujúcich pojem t_i

2.5 Evolúcia používateľských preferencií

Preferencie používateľ a sa z časom menia, môžu vzniknúť nové a zaniknúť staré, prípadne sa vracajú predošlé. Na základe toho môžeme používateľské preferencie rozdeliť na

- krátkodobé preferencie,
- dlhodobé preferencie,
- sezónne preferencie.

Odhalenie krátkodobých záujmov je pomerne triviálne, stačí agregovať používateľové záujmy za časové obdobie ktoré považujeme za „krátku dobu“ a vrátiť značky ktoré používateľ preferoval najčastejšie.

Dlhodobé záujmy

Problém dlhodobých záujmov je o dosť komplikovanejší, väčšina riešení ktoré som preskúmal používala na tento problém kombináciu rôznych váhovacích algoritmov a štatistických metód. Napríklad v článku [?] , kde použili už spomínané váhovacie algoritmy.

Používateľský profil je reprezentovaný trojrozmerný váhovaným vektorovým profilom a zoznamom zobrazených dokumentov (navštívené url). Následne sa najskôr vytvorí zoznam adekvátnych dokumentov zo značiek. Následne na porovnávanie značiek dokumentov a profilov sa používajú tri rôzne algoritmy.

Jednoduché porovnávanie kde sa vlastne spočítajú váhy značiek ktoré majú spoločné používateľ a dokument podľa vzorca 10.

$$u_j(d_i) = \sum_{t=1}^N z_t f(z_t) * u(z_t) \quad (9)$$

Vysledkom je funkcia užitočnosti pre dokument i $u_j(d_i)$, N_{z_i} je počet unikátnych značiek v dokumente d_i , $f(z_t)$ je počet výskytu značky z_t v dokumente a $u(z_t)$ je vypočítaná užitočnosť značky z_t .

Porovnávanie unikátnych značiek, teda zanedbávame váhu určenú váhovacími algoritmami a iba spočítame unikátne značky podľa vzorca ??.

$$u_u(d_i) = \sum_{t=1}^N z_i u(z_t) \quad (10)$$

Jazykový model (angl. Language model) ktorý generuje unigramovy jazykový model (angl. unigram language model) kde užitočnosť značiek je použitá ako ako frekvencia značiek v rovnici 11.

$$u_{lm}(d_i) = \sum_{t=1}^N z_i \log u(z_t) + 1 \sum_{j=1}^N z_i \quad (11)$$

Ďalej sa ešte výsledne dokumenty poskytnuté jedným z týchto algoritmov ešte filtrujú algoritmom PClick, ktorý pracuje z históriou navštívených dokumentov. Tento algoritmus vracia iba dokumenty ktoré používateľ v histó často navštevoval. Matematický je to vyjadrené rovnicou 12. Tento algoritmus berie do úvahy aj vyhľadávací dotaz.

$$u_{pclick}(d_i) = \frac{|Zobrazenia(d_i, u_j, q_n)|}{|Zobrazenia(q_n, u_j)| + \beta} \quad (12)$$

Kde $Zobrazenia(d_i, u_j, q_n)$ je počet zobrazení dokumentu d_i a $Zobrazenia(q_n, u_j)$ je celkový počet zobrazení dokumentu d_i pre vyhľadávací dotaz q_i .

Ďalším možné riešenie som našiel v článku Ferida Achemoukh a spol. [?]. Toto riešenie využíva hlavne štatistické metódy. Používateľ a modeluje ako bayesovú sieť ktorú modeluje pre niekoľko sedení ktoré sa skladajú z niekoľkých používateľových interakcií zo systémom.

3 Návrh aplikácie

V tejto kapitole sa budeme zaoberať zvolenými technológiami a abstraktným návrhom aplikácie.

3.1 Platforma

Prácu som sa rozhodol vypracovať vo forme webovej aplikácie. Túto formu som zvolili najmä pre väčšiu dostupnosť pre koncového používateľa. Tak isto veľkým plus je že celá aplikácia beží na strane tvorca takže prípadne zmeny v systéme nebude potrebné synchronizovať naprieč viacerými zariadeniami. Ďalšou výhodou je redukovanie problémov zo synchronizáciou databáza medzi koncovými zariadeniami.

Tak isto protabilita riešenia je pomerne dobrá, nepotrebujem sa zaoberať rôznymi verziami operačných systémov. Proti tomuto by sa dalo argumentovať rozdielmy medzi internetovými prehliadačmi a potrebu responzívneho dizajnu pre mobilné zariadenia, ale toto všetko sa dá jednoducho vyriešiť voľbou správny grafických knižníc.

Keďže sa jedná o webovú aplikáciu, bude bežať čiastočne vo forme klient server ako html, css a javascript kód bežiaci vo vyhľadávачi používateľ a a php kód bežiaci na strane.

Klient

Klientska časť je postavená na JavaScripte, HTML a CSS3. Aby som zabezpečil responzívny dizajn a portabilitu naprieč prehliadačmi, rozhodol som sa použiť knižnicu Bootstrap³, ktorá je postavená na CSS3 a jQuery. Vďaka nej nemusím kontrolovať vizuálnu stránku v rôznych prehliadačoch a vytvorenie dizajnu, ktorý dobre spolupracuje s rôznymi rozmermi obrazoviek a rozlíšení, ma stojí pomerne málo námahy. Tak isto plánujem použiť jQuery⁴ v prípade potreby implementácie logiky na klientskej strane a CSS3⁵ v prípade potreby hlbších zásahov do grafiky.

³<http://getbootstrap.com/>

⁴<https://jquery.com/>

⁵<http://www.css3.info/>

Server

Na strane servera som sa rozhodol použiť jazyk PHP5⁶ najmä z dôvodu že s ním mám predošlé skúsenosti, tak isto pre tento jazyk sa dá ľahko vyhľadať hosting a má veľmi rozsiahlu dokumentáciu. Avšak v tomto jazyku je nedostatok štruktúry čo niekedy môže viesť k veľkým neudržateľným aplikáciám, preto som sa rozhodol použiť aplikačný rámec (angl. framework) Yii2⁷ ktorý má sadu odporúčaných praktík vďaka ktorým je aplikácia dlhodobo udržateľná. Ďalšou výhodou je že tento aplikačný rámec má v sebe zapracované všetky hore uvedené klientské knižnice, čím uľahčuje prácu s nimi. Tak isto nám tento aplikačný rámec umožňuje používať rôzne databázy vďaka svojmu modulu DAO⁸ pomocou ktorého modeluje SQL dotazy. Avšak niektoré zložitejšie dotazy sa v ňom nedajú namodelovať. Ako databázu som zvolil MySQL⁹ vďaka jej rýchlosti, popularite ale najmä dostupnosti.

Vývojová platforma

Ako správcu verzií som sa rozhodol použiť git¹⁰. Tento nástroj je pomerne jednoduchý na ovládanie a spoľahlivý. Tak isto existuje množstvo internetových služieb kde sa pomocou neho dá zálohovať. Verziovanie dopĺňam správou balíkov pomocou manažéra balíkov Composer¹¹

Na testovanie som sa rozhodol použiť na PHPUnit postavený Codeception, ktorý je integrovaný do Yii2.

Ako editor používam Vim z množstvom prídavných modulov.

3.2 Základné rozvrhnutie aplikácie

Keďže ide o webovskú aplikáciu, na najvyššej úrovni bude postavená na vzore Klient-Server. Avšak z dôvodu bezpečnosti sa bude väčšina úloh vykonávať na servery. Aplikácia bude získavať dáta z už existujúcej databázy hudobných doku-

⁶<http://php.net/>

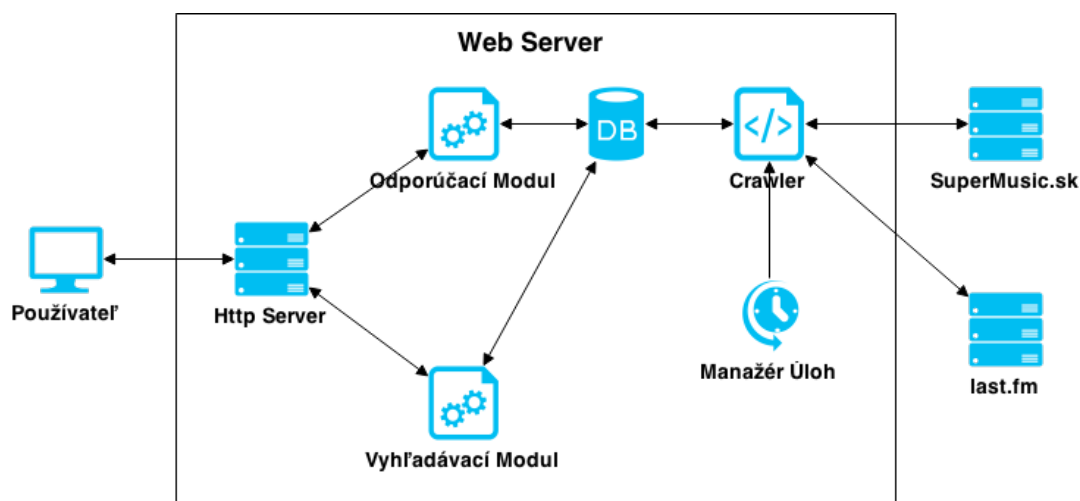
⁷<http://www.yiiframework.com/>

⁸<https://github.com/yiisoft/yii2/blob/master/docs/guide/db-dao.md>

⁹<https://www.mysql.com/>

¹⁰<http://git-scm.com/>

¹¹<https://getcomposer.org/>



Obr. 4: Štruktúra aplikácie

mentov¹². Tieto dáta chcem doplniť o sociálne tagy zo služby last.fm¹³. Nad týmito dátami sa bude vykonávať vyhľadávanie, ale najmä odporúčanie. Základný náčrt aplikácie môžeme vydiť na obrázku 4.

Aplikácia sa bude skladať z dôch hlavných oddelených subsystémov

- odporúčanie (spúšťané používateľom cez prehliadač),
- indexovanie dokumentov (séria úloh spúšťaná manažérom úloh).

3.3 Špecifikácia služieb

Aplikácia by mala poskytovať niekoľko základných služieb

- vyhľadávanie dokumentov,
- nájdenie podobných dokumentov aktuálne zobrazenému dokumentu,
- odporúčanie dokumentov,
- zostavenie spevníka.

¹²<http://www.supermusic.sk/>

¹³<http://www.last.fm/home>

Prvé dve funkcionality sú vytvorené najmä z dôvodu že v danej doméne sa ťažko získavajú preferencie, ich jedinou úlohou je získanie základných preferencií používateľa na základe ktorých budeme vytvárať odporúčania.

Pre zabezpečenie týchto funkcií bude potrebované udržiavať databázu dokumentov ktorá sa bude pravidelne aktualizovať. Toto bude zabezpečovať Kravler (angl. crawler) ktorý bude vykonávať činnosti

- prehľadávanie supermusic po nových dokumentoch,
- prehľadávanie supermusic po nových interpretoch,
- sťahovanie dokumentov zo supermusic,
- sťahovanie značiek z last.fm,
- váhovanie značiek.

Kravler bude implementovaný ako aplikácia pre príkazový riadok aby sa dal použiť z manažérom úloh.

Odporúčanie

Na odporúčanie by som chcel využiť filtrovanie na základe obsahu spolu z vlastným algoritmom na určenie dlhodobých, krátkodobých a sezonných záujmov.

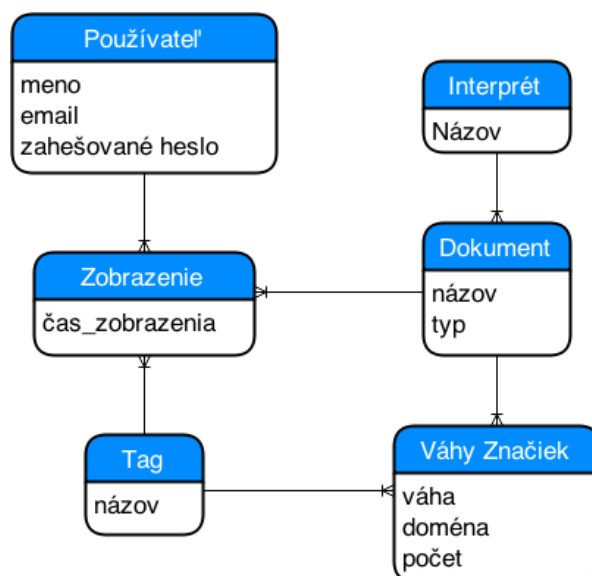
4 Implementácia

4.1 Datový model

Model dokumentu

Dokument bude reprezentovaný svojím názvom, typom a značkami, pričom značky budú mať doménu na základe toho či vznikli z názvu dokumentu, jeho typu, názvu interpreta alebo alebo boli získane zo služby last.fm.

Čiže dokument bude uložený v trojrozmernom vektorovom priestore, kde prvá súradnica budú dokumenty $D = d_1, d_2, \dots, d_n$, ďalšia súradnica budú značky $Z = z_1, z_2, \dots, z_n$ a posledná súradnica budú domény $D = d_{názov}, d_{interpret}, d_{typ}, d_{tag}$.



Obr. 5: Zjednodušený datový model

Model používateľa

Používateľov profil bude reprezentovaný históriou navštívených značiek. Bude sa ukladať pre každú navštívenú značku a jeden riadok bude obsahovať informácie $row = d_i, u_j, z_l, čas$ kde d_i je zobrazený dokument, u_j je používateľ ktorý dokument zobrazil, z_l je značka ktorá bola zobrazená.

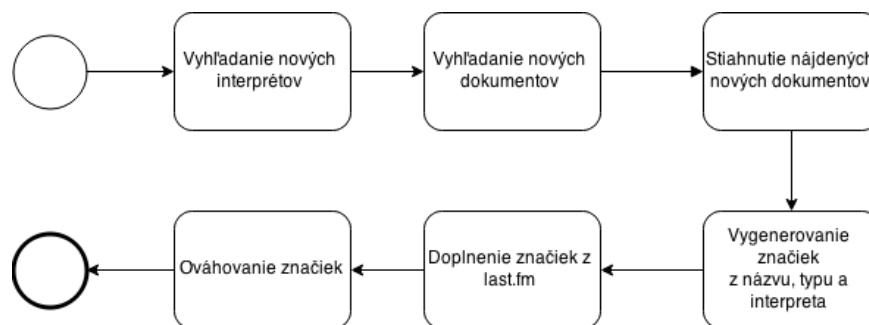
Celú túto schému môžeme vydiť na obrázku 5.

4.2 Indexovanie dokumentov a váhovanie

Kravler

Kravler (angl. crawler) je program slúžiaci na prehľadávanie webu za účelom indexovania¹⁴. Kravler vykonáva niekoľko operácií v rámci aplikácie, je implementovaný ako jednoduchý program príkazového riadku. Každá z jeho operácií sa v prípade potreby dá vykonať osobitne. Toto je hlavne potrebné ak by sa zmenili parametre indexovania. Na obrázku 6 môžeme vydiť digram akcií kravleru.

¹⁴http://en.wikipedia.org/wiki/Web_crawler



Obr. 6: BPMN Diagram kravleru

Vyhľadanie nových interpretov

Vyhľadanie nových dokumentov

Pri tejto akcii sa využíva spôsob triedenie piesni na supermusic. Piesne sú triedené podľa abecedy a na ich zobrazenie slúži adresa „[http://www.supermusic.sk/piesne.php?od=\\$param](http://www.supermusic.sk/piesne.php?od=$param)“. pričom param je reťazec na ktorý majú piesne začínať. Experimentálne som overil že na zobrazenie skoro všetkých dokumentov mi stačí použiť dva znaky.

Prvý znak je anglická abeceda rozšírená o niekoľko slovenský a českých interpunkčný písmen plus hviezda. Teda v prvom rade sú písmená z množiny A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Z, Č, Ď, Ľ, Ř, Š, Ť, Ž, v druhej rade už je anglická abeceda.

Následne sa snaží získať zoznam piesni zo stiahnutej stránky, za týmto účelom používam tri XPath výrazy

- „`./table[@width=740]//td/a/text()`“ na vyhľadanie všetkých názvov piesni,
- „`./table[@width=740]//td/a/@href`“ na získanie univerzálnneho identifikátora piesne,
- „`./table[@width=740]//td/amt/@src`“ na získanie typu dokumentu,
- „`./table[@width=740]//td/text()`“ na získanie názvu interpreta.

Tieto parametre sa následne ďalej spracúvajú.

Stiahnutie nových dokumentov

Vygenerovanie značiek

Doplnenie značiek z last.fm

Váhovanie značiek dokumentov

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

V MySQL je tento prístup implementovaný pomocou nasledujúcej rovnice¹:

$$w_d = \frac{\log(dt f_d) + 1}{\sum_{i=1}^t \log(dt f_i) + 1} \cdot \frac{U}{1 + 0.0115 * U} \cdot \log \frac{N}{nf}$$

- $dt f_d$ je sila (množstvo koľko krát sa nachádza pojem v text v prípade analýzy textu) vlastnosti vyhodnocovaného dokumentu
- $dt f_i$ sila i -tej vlastnosti
- U počet unikátnych vlastností 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 priemerná dĺžka dokumentu, jeho relevancia stúpa. [16]

¹<http://dev.mysql.com/doc/internals/en/full-text-search.html>

Inverzná frekvencia

Zabezpečuje že menej časté pojmy majú vyššiu váhu.

4.3 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 slov z projektu TODO: Ako ? `google code stop-words`

4.4 Zostavenie spevníka

Aplikácia bude podporovať funkcionality automatického generovania spevníka, kedy si používateľ zvolí používateľov s ktorými si chce ísť zahrať a aplikácia automaticky vygeneruje spevník zložený s najpreferovanejších hudobných diel daných používateľov.

4.5 Krawler ?

Tento komponent prehľadáva databázu ktorá je cieľom môjho odporúčača, využíva k tomu abecedne zobrazenie záznamov databázy. Databáza sa nedá zobrazit' 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 stránok na 3380.

4.6 Indexing

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

Priama tagovacia tabuľka

Vytvoril som tabuľku tagov, kde bol každý tag fyzický priamo vložený spolu z id dokumentu ku ktorému sa viaže, tento prístup ale nebol dostatočne rýchly na

vygenerovanie, ani na vyhľadávanie. Pri vyhľadávaní nad 118989 značkami označujúcimi 47002 dokumentov zabral 44.6984 sekúnd. Nepomohlo ani zindexovanie podľa mena.

4.7 Váhovanie Dokumentu

$$w(d_j) = \sum_{i=1}^N w(t_i)$$

- N Počet značiek v dokumente,
- t_i i -ty pojem v dokumente,
- d_j j -ty dokument.

Literatúra

- [1] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. august 2011.
- [2] Peter Brusilovsky. User profiles for personalized information access. School of Information Sciences University of Pittsburgh, USA, 2009.
- [3] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. 2002.
- [4] Franco Fabbri. A theory of musical genres: Two applications. 1980. [navštívené 27 apríla 2015].
- [5] *information_retrieval. Probabilistic information retrieval parti : Survey.* [navštívené 10 januara 2015].
- [6] Karin Kosina. Music genre recognition. 2002.
- [7] Paul Lamere. Social tagging and music information retrieval. 2009. [navštívené 24 apríla 2015].
- [8] DIK L. LEE. Document ranking and the vector-space model. 1997.
- [9] Namunu C. Maddage, Li Haithou, and Mohan S. Kankanhalli. Music structure analysis statistics for popular songs. November 2009.
- [10] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. júl 2002.
- [11] Gašpar Peter. Odporúčanie s využitím osobných vyjadrení. 2014.
- [12] Yves Raimond, Samer Abdallah, and Mark Sandler. The music ontology. 2007.
- [13] Francesc Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *RECOMMENDER SYSTEMS HANDBOOK*, chapter Recommendation Techniques, pages 11 – 14. Springer New York Dordrecht Heidelberg London. [navštívené 1 máj 2015].

- [14] Arnold Schoenberg. *Fundamentals of musical composition*. Faber and Faber Limited, 1967.
- [15] Hinrich Schütze. Introduction to information retrieval. [navštívené 6 januara 2015].
- [16] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. 1996.
- [17] John F. Sowa. Semantic networks. [navštívené 4 februára 2015].
- [18] Kiryl Tsishchanka. Exponential growth and decay. 2010.
- [19] Tong Zhu. Nonlinear p´olya urn models and self-organizing processes. 2009.