Jihočeská univerzita v Českých Budějovicích

Ekonomická fakulta

Katedra aplikované matematiky a informatiky

Bakalářská práce

Analýza a návrh doporučovacího systému

Vypracoval: Martin Hořejš

Vedoucí bakalářské práce: doc. Ing. Ladislav Beránek, CSc., MBA

Studijní program: Systémové inženýrství a informatika

Studijní obor: Ekonomická informatika

České Budějovice 2015

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě - v úpravě vzniklé vypuštěním vyznačených částí archivovaných ... fakultou elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne .....

Martin Hořejš

Poděkování

Rád bych poděkoval vedoucímu bakalářské práce, panu doc. Ing. Ladislavovi Beránkovi, CSc., MBA, za jeho pomoc, podněty, trpělivost a odborné vedení při zpracování této bakalářské práce.

Obsah

1 Úvod

1.1 Cíle bakalářské práce

Cílem bakalářské práce je analyzovat problémy existujících doporučovacích systémů a navrhnout vhodná opatření, realizovat program řešící problematiku zadání neúplných informací o uživateli. V práci budou představeny existující doporučovací systémy, metody doporučování které tyto systémy využívají, problematika záznamu uživatelské preference a algoritmy vhodné pro řešení problematiky neúplných dat.

1.2 Východiska

Mým cílem je realizovat program pro doplňování informací o uživateli, který o sobě nezadal všechny požadované informace. Pro příklad co s uživatelem který o sobě nevyplní všechny údaje? Takový uživatel má stále pro provozovatele e-commerce portálů stále hodnotný, protože se chybějící data dají odvodit za pomoci kolaborativního filtrování, tyto metody pomáhání zvyšovat přesnost doporučení a tak i šanci na doporučení zakončené nákupem

1.3 Metodiky práce

V úvodu práce rozeberu potřebu nasazení doporučovacích systémů. Popíši několik existujících portálů využívající různé metody doporučení. Poté se zaměřím na shrnutí problematiky záznamu a vyjádření uživatelské preference, jejich výhod a nevýhod a optimálním záznamem uživatelské preference. Následně sepíši nejpoužívanější typy doporučení, vysvětlím jejich použití, popíši jejich funkčnost a vyberu výhody a nevýhody. Vysvětlím algoritmy kolaborativního filtrování, které jsou vhodné pro program na doplňování chybějících dat. V prvním oddílu praktické části analyzuji několik funkčních portálů využívající různé metody doporučení, pokusím se najít nedostatky a navrhnout vhodné změny vedoucí ke zlepšení chodu jednotlivých portátů. Druhá část je zaměřena na realizaci programu, který řeší problematiku neúplných uživatelů. Program bude využívat metody zmíněné v teoretické části. Celou práci budu psát v LaTeX kvůli velmi vysoké typografické kvalitě a možnosti předdefinovat formátování dokumentu.

1.4 Doporučovací systémy v prostředí e-commerce

Počátky internetového nakupování sahají do 90. let 20. století. S růstem dostupnosti internetu a rozvojem pokročilých metod plateb počátkem 21. století se rozmohlo i obchodování na internetu. V dnešní době jen v ČR používá denně internet více, než polovina populace. Mezi hlavní důvody patří hlavně pohodlnost oproti dojíždění do kamenných obchodů, dostupnost, kterou zajišťuje neustálá

inovace výpočetní techniky jak na straně e-shopů, tak i na straně kupujících a téměř neomezený výběr zboží. Pro prodejce se stalo obchodování na internetu daleko výhodnějším, proto je stále větší důklad kladen na sběr a analyzování dat.

S tím jsou úzce spjaty doporučovací systémy, správný doporučovací systém je výhodou jak pro prodejce, který je schopen zákazníkovi nabídnout to co by mohl chtít, tak pro nakupujícího, který má zjednodušenou práci a vyhledá potřebné produkty za kratší čas.

Pro „větší“ portály je doporučovací systém samozřejmostí (amazon.com, alza.cz, facebook.com, youtube.com), avšak „menší“ portály tento systém nevlastní

Nejpravděpodobnější důvody pro absenci doporučovacího systému:

- **Neinformovanost** – Mnoho provozovatelů e-commerce ani neví o možnosti zavést doporučovací systém.
- **Neschopnost zavést systém** – Provozovatelé znají výhody, ale z časových, nebo finančních důvodů nejsou ochotni doporučovací systém zavést.
- **Neochota** – Provozovatel zná doporučovací metody, ale není přesvědčen o efektivnosti, nebo důležitosti.
- **Neefektivnost** – Pro velmi malé portály bez potenciálu na růst se implementace DS nemusí časově, nebo finančně vyplatit.

2 Doporučovací systémy v praxi

Hlavním cílem každého doporučovacího systému je poskytnout uživateli smysluplné doporučení na základě zpětné vazby.

Každé doporučení se dělí:

- **Individualizované** – Každý uživatel dostane jedinečné doporučení na základě jeho předchozích akcí.
- **Neindividualizované** – Automatické doporučení např. nejprodávanější, nejnovější produkty. Tyto doporučení jsou mnohem jednodušší na vytvoření a jsou typické pro neregistrované uživatele.

2.1 Sociální sítě (Facebook.com)

Facebook je rozsáhlý společenský webový systém sloužící hlavně k tvorbě sociálních sítí, komunikaci mezi uživateli, sdílení multimediálních dat, udržování vztahů a zábavě. Se svojí miliardou aktivních uživatelů (říjen 2012) je jednou z největších společenských sítí na světě. Je plně přeložen do šedesáti osmi jazyků. V roce 2010 vznikl americký film The Social Network, který pojednává o počátcích Facebooku.

Obrázek 1: Zakladatel Facebooku Mark Zuckenberg

Zdroj: Brooklyn Magazine

Facebook byl založen Markem Zuckerbergem, bývalým studentem Harvardovy univerzity. Původně byl tento systém omezen jenom pro studenty Harvardovy univerzity, pod doménou thefacebook.com. Během dvou měsíců byl rozšířen na některé další, které patří do tzv. Ivy League, a již do konce roku byly připojeny další univerzity. Nakonec byl přístup otevřen pro všechny uživatele s univerzitní

e-mailovou adresou, nebo pro některé zahraniční schválené univerzity, v Česku k prvním otevřeným vysokým školám patřila Masarykova univerzita. Od 27. února 2006 se začaly do systému připojovat některé nadnárodní obchodní společnosti. Od 11. srpna 2006 se může dle licence používání připojit kdokoli starší 13 let. Uživatelé se v systému mohou připojovat k různým skupinám uživatelů, kteří působí například v rámci jedné školy, firmy nebo geografické lokace.

Pro registrované uživatele nabízí Facebook hned několik doporučovacích metod:

- **Vyhledat přátele** – Systém doporučí přátele podle společných přátel, navštívených míst, vystudovaných škol, nebo podle zaměstnání. Systém je schopen analyzovat společné údaje a podle toho přátele doporučit.
- **Doporučené příspěvky** – Facebook zobrazuje aktivní stránky podle aktivity uživatelů. Stránka, která má 10 000 fanoušků a každý nový příspěvek obdrží 50 komentářů se bude procentuálně méně zobrazovat, než stránka s 2 000 fanoušky, kde každý nový příspěvek bude mít 1 000 komentářů. Facebook vyvinul funkci „propagovat“, která dokáže předem odhadnout kolik lidí váš příspěvek a za jakou cenu propagace. uvidí, nebo označí „To se mi líbí“. Otom, zda se vám propagovaná stránka zobrazí rozhodují parametry zadané při vytváření. (Wikipedie, 2015)

2.2 Server pro sdílení videa (youtube.com)

YouTube je největší internetový server pro sdílení videosouborů. Založili jej v únoru 2005 zaměstnanci PayPalu Chad Hurley, Steve Chen a Jawed Karim. V listopadu 2006 byl zakoupen společností Google za 1,65 miliardy dolarů (tehdy asi 37 miliard Kč). Od 9. 10. 2008 má YouTube i české rozhraní. Byla tak spustěná 25. služba Google v pořadí. Google kromě českého překladu serveru přinesl také spolupráci s místními partnery. Česko se stalo 22. zemí světa a desátou v Evropě, kde byl YouTube lokalizován.YouTube navštíví měsíčně 4,2 mil. unikátních českých uživatelů, podle průzkumu ho v květnu 2012 navštívilo alespoň jednou 82 % lidí připojených k internetu v České republice.

- 98 % české online populace zná YouTube alespoň podle jména
- 21 % české populace (15+) navštíví YouTube denně, 46 % týdně a 56 % měsíčně
- 80 % české online populace má o značce YouTube pozitivní mínění

Doporučovací systém na portálu YouTube.com doporučí uživateli potencionálně zajímavá videa na úvodní stránce. Doporučená videa nejsou placenou propagací, ale aktuálně populární videa.

2.3 Hudební encyklopedie (Last.fm)

Last.fm je internetové rádio, hudební encyklopedie a systém pro doporučování hudby, který se spojil se sesterským produktem Audioscrobbler v srpnu 2005. Audioscrobbler začal jako počítačový projekt Richarda Jonese při jeho studiu na Southamptonské univerzitě v Británii. Richard Jones vyvinul první plugin a otevřel API komunitě, následně bylo podporováno mnoho hudebních přehrávačů pro různé operační systémy. Audioscrobbler uměl jen zaznamenávat písničky které uživatel poslouchal a používat na ně kolaborativní filtrování a dělat z nich žebříčky.

Last.fm bylo založeno v roce 2002 Felixem Millerem, Martinem Stikselem, Michaelem Breidenbrueckerem a Thomasem Willomitzerem, všichni z Rakouska a Německa, jako internetová radiová stanice a hudební komunitní stránka používající podobné hudební profily pro generování dynamických playlistů. Tlačítka „milovat“ a „zakázat“ umožňovaly uživatelům přizpůsobovat své profily. Last.fm vyhrál Europrix 2002 a byl nominován na Prix Ars Electronica v roce 2003.

2.4 Internetový obchod (Amazon.com)

Amazon.com je internetový obchod patřící americké společnosti Amazon.com, Inc. ve státě Washington. Patří mezi nejstarší a největší obchody svého druhu. Jeff Bezos provozoval už v roce 1994, v začátcích Internetu, knihkupectví Cadabra.com, které téhož roku přejmenoval na Amazon podle řeky Amazonky. V roce 1998 koupil také Internet Movie Database (IMDb), Alexa.com. Příjmy společnosti se dnes pohybují okolo 7 miliard dolarů ročně a jako jedna z mála společností dokázala růst i v časech ekonomické krize. Firma vyrábí také vlastní čtečky elektronických knih Amazon Kindle a tablet Kindle Fire. Amazon dovoluje i prodej malovýrobců v její síti, ale bere si za to 15% provize.

4. prosince 2009 začaly evropské pobočky Amazon zasílat do ČR kromě knih, CD a DVD také zboží z ostatních produktových kategorií (elektronika, hračky atd.). V Evropě fungují k roku 2011 pobočky ve Velké Británii, Francii, Německu a Itálii. V roce 2013 se Amazon rozhodl vybudovat dvě logistická centra v Česku, každé za miliardu korun a o rozloze 95 tisíc m^2 (svou velikostí cca 13 fotbalových hřišť by se tak měl rovnat zatím největšímu skladu v Německu).

2.5 Filmová databáze (Csfd.cz)

Česko-Slovenská filmová databáze (zkracována na ČSFD) je česko-slovenská obdoba databáze IMDb. Založil ji v roce 2001 Martin Pomothy. Od IMDb se kromě lepšího pokrytí lokální scény liší i tím, že uživatelům umožňuje vést si přehled vlastní filmotéky. IMDb ale celkově uvádí více detailů o audiovizuálním díle než ČSFD.

Za dobu své existence změnil server celkem 4× svou podobu, naposledy počátkem roku 2011. ČSFD se často stává mediálním partnerem různých akcí spojených

s filmem (např. Filmasia), kino a DVD premiér (např. DVD Levných knih) atd. V roce 2011 společnost získala v soutěži Křišťálová Lupa 1. místo v kategorii zájmové weby. V kategorii All Star obsadila 4. místo.

Funkce databáze:

- Profily filmů, herců a hudebních skladatelů
- Televizní program
- Přehled Kino premiér
- Přehled DVD premiér, včetně oddělené sekce pro levná příbalová DVD
- Filmové diskuze
- Statistiky a žebříčky filmů

Uživatelské funkce:

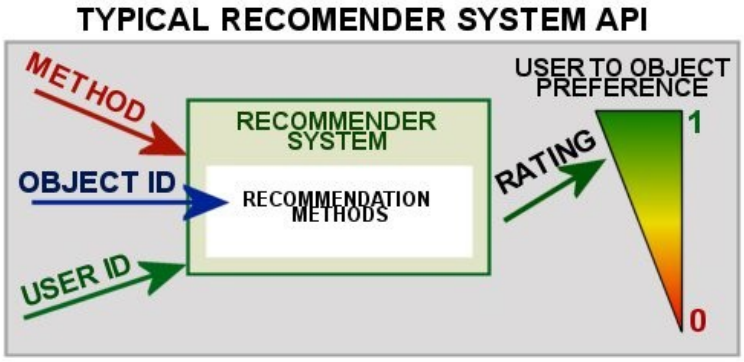
- Hodnocení filmů, případně jejich komentování
- Videotéka
- Deníček
- Budíček (v roce 2012 přejmenován na Chci vidět!)
- Bazar
- Uživatelské profily

3 Uživatelské preference

Preference je relace mezi uživatelem U (User) a objektem zájmu O (Objekt zájmu) a je obvykle definována jako funkce $PU(o) : OxU \rightarrow [0, 1]$ která vrací míru oblíbenosti – Dvouhodnotový jazyk - líbí/nelíbí V {0,1}. Zpětná vazba získána od uživatele představuje data potřebná k vyjádření preference. Pro provozovatele e-commerce portálů je důležité, aby uživatel vyjádřil co a jak moc se mu líbí, každý člověk je unikát a na každého je potřeba pohlížet zvlášť. Základní rozlišení preferencí je dlouhodobá a krátkodobá.

Nesmí se zapomenout na uživatelův osobní vývoj v čase a krátkodobou nestálost jako je aktuální duševní rozpoložení, protože člověk ve stresu, nebo s nedostatkem času se chová jinak.

Každý uživatel při práci se systémem e-commerce zanechává v systému zpětnou vazbu, ať už vědomě (hodnocením položek, vyplněním dotazníku nebo uživatelského profilu informacemi, jako jsou věk, bydliště, vzdělání nebo pohlaví) nebo nevědomě (pohyb na stránce, doba pobytu na stránce, ...). Na základě těchto informací je pak možné činit závěry či předpoklady o uživatelově preferenci vůči některému objektu. (Kortus, 2013) (Peška, 2010)



Obrázek 2: Vztah zpětné vazby a uživatelské preference

Zdroj: (Peška, 2010)

3.1 Dlouhodobá preference

Dlouhodobá preference lépe vystihuje, čím se uživatel řídí při výběru produktů. Pro příklad uživatel preferuje u stolního PC výkonnou grafickou kartu před kapacitou HDD, Smartphony s OS Android před OS Windows Mobile atd. (Vojtáš, 2010)

3.2 Krátkodobá preference

Krátkodobá preference reprezentuje aktuální uživatelův cíl. Pokud má uživatel v úmyslu koupit levný, méně výkonný notebook pro syna do školy, tak preference herního stolního PC se bude blížit 0, i přes to, že za jiných okolností se preference blíží 1. (Vojtáš, 2010)

3.3 Předmět preference

Předmět preference je vlastnost toho, o čem se uživatel rozhoduje. Každý uživatel upřednostňuje jiné atributy. Atributy objektu se nemění (kromě roku výroby) a každý má svou váhu v závislosti na preferenci uživatele. Pokud si chce uživatel koupit úsporné rodinné auto a na výkonu motoru mu nezáleží, systém přiřadí parametru spotřeba vyšší váhu, než váhu parametru výkon.(Vojtáš, 2010)

Atributy:

- **Nominální** – Barva, výrobce
- **Numerické** – Velikost, rozlišení, nosnost, výkon, rok výroby
- **Specifické** – Těžko zachytitelné atributy (tvar, design)

3.4 Identifikace uživatele

Pro jakékoli uvažování o preferencích uživatele je klíčová schopnost jej jednoznačně identifikovat. V současné době existují tři postupy, jak identifikaci uživatele provádět:

1. **Registrace uživatele** – registrace je nejspolehlivější způsob identifikace uživatele. Systém identifikuje unikátní přihlašovací jméno – umožní tedy rozeznat jednoho uživatele používajícího web z různých počítačů i více různých uživatelů používajících stejný počítač. Hlavní nevýhodou registrace je sama nutnost ji vyplnit, některé uživatele může nutnost registrovat se odradit od používání daného webu, v případně pokud registrace není povinná nezanedbatelná část uživatelů jí nevyplní.
2. **Identifikace pomocí IP adresy** – systém identifikuje unikátní IP adresu, což je zároveň jeho hlavní slabinou – stejnou IP adresu velmi často sdílí více počítačů v lokální síti a tedy pravděpodobně i více uživatelů (matka a syn v jedné rodině budou mít s největší pravděpodobností jiné preference). Systém zároveň nerozpozná stejného uživatele, který k webu přistupuje z různých lokací. Identifikaci uživatele na základě IP adresy bychom se proto měli raději vyhnout.
3. **Identifikace pomocí COOKIES** – systém identifikuje unikátní kombinaci PC + prohlížeč. Nerozpozná tedy stejného uživatele, který k webu

přistupuje z různých počítačů, ale za předpokladu, že s počítačem pracuje pouze jeden uživatel, rozpozná dobře jednotlivé uživatele (identifikace uživatelů je „jemnější“ oproti identifikaci pomocí IP). Nevýhodou systému je, že cookies, pomocí kterých je identifikace prováděna, jsou ukládány na pevném disku uživatele. Uživatel je proto může kdykoli odstranit nebo jejich ukládání zakázat a znemožnit tak svou identifikaci. I přes tyto problémy se identifikace pomocí COOKIES jeví jako lepší varianta oproti identifikaci pomocí IP.

Jako ideální řešení identifikace uživatele v prodejních webech se jeví kombinace identifikace pomocí cookies pro nepřihlášené uživatele s možností registrovat uživatele (a následná identifikace pomocí přihlášení). (Peška, 2010)

3.5 Způsob záznamu preferencí (Feedback)

3.5.1 Přímá

Přímá preference položek vyjadřuje zákazník pomocí nastavitelného rozhraní své požadavky (výrobce, barva, maximální, nebo minimální cenu atd.) Uživatel obvykle vyjadřuje přímé preference buď procházením katalogu produktů – zobrazení kategorie ¯ preference objektů této kategorie, vyhledáváním na základě klíčového slova nebo vyhledáváním podle atributů objektu. Přímé preference se obvykle přeloží jako dotaz do databáze objektů (uživateli je pak prezentován výsledek dotazu). (Peška, 2010)

3.5.2 Implicitní

Implicitní zpětné vazby jsou takové informace, které jsou získávány na základě chování uživatele na webu bez aktivní účasti uživatele na jejich podání. Mezi implicitní zpětnou vazbu můžeme zařadit například otevření stránky, přehrání hudební skladby, kliknutí na odkaz atd. Důležitým aspektem implicitní zpětné vazby je její doménová závislost – pro různé domény má smysl uvažovat různé implicitní uživatelské akce (například přehrání hudební skladby u hudebního doporučovače). (Žák, 2010)

Výhody:

- Množství získaných dat oproti explicitní metodě
- Každý uživatel zanechá implicitní zpětnou vazbu

Nevýhody:

- Těžko se interpetuje
- Težko se vyjadřuje negativní preference
- Náročné na zpracování

Pro všechny e-commerce portály lze identifikovat několik společných operací, které provádí uživatel. Identifikací takových faktorů (na základě porovnání s explicitním hodnocením) se zabýval ve své diplomové práci Vladimír Žák (Žák, 2010) a Ladislav Peška (Peška, 2010). V následující tabulce se pokusím provést jejich sumarizaci:

Faktor	**Hodnota**	**Poznámka**
Čas strávený na stránce	Dobrá	Čas na stránce byl vyhodnocen jako dobrý implicitní faktor
Počet akcí na stránce	Dobrá	Počet akcí dosáhl nejvyšší korelace s preferencí uživatele
Čas pohybu myši	Špatná	Vliv času pohybu myši na explicitní hodnocení objektu se nepotvrdil
Počet kliknutí myší	Špatná	Vliv počtu kliknutí na stránce na explicitní hodnocení objektu se nepotvrdil
Scrollování	Dobrá	Scrollování bylo identifikováno jako dobrý implicitní faktor

Tabulka 1: Implicitní faktory a jejich důležitost

Zdroj: Kombinace zdrojů (Peška, 2010) (Žák, 2010)

Dalšími možnými implicitními faktory pro prodejní weby jsou:

- Objednávka produktu
- Vložení produktu do košíku
- Odeslání příspěvku do diskuze k produktu
- Přidání na osobní seznamy (Seznam přání, sledovaný produkt aj.)
- Porovnání s podobnými produkty

3.5.3 Explicitní

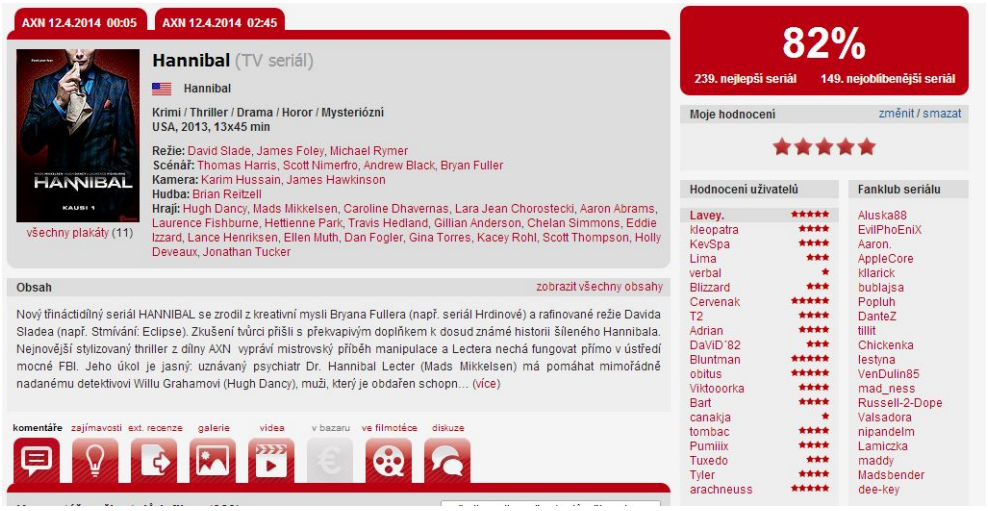
Explicitní zpětná vazba značí, že uživatel musel dobrovolně vyvinout snahu ohodnotit produkt. Produkt je hodnocen jako celek a hodnotí se podle předem daných pravidel (bodování, hodnocení hvězdičkami, procentuálně atd.). (Vojtáš, 2010)

Výhody:

- Málo námahy pro uživatele
- Přirozené

Nevýhody:

- Omezená škála hodnocení (0-5 hvězdiček)
- Hodně položek získá maximální hodnocení
- Ohodnocení více produktů může být pro uživatele zdlouhavé a frustrující
- Při větší škále (>20) je uživatel zmaten a jeho hodnocení je nekonzistentní



Obrázek 3: Explicitní hodnocení

zdroj: csfd.cz

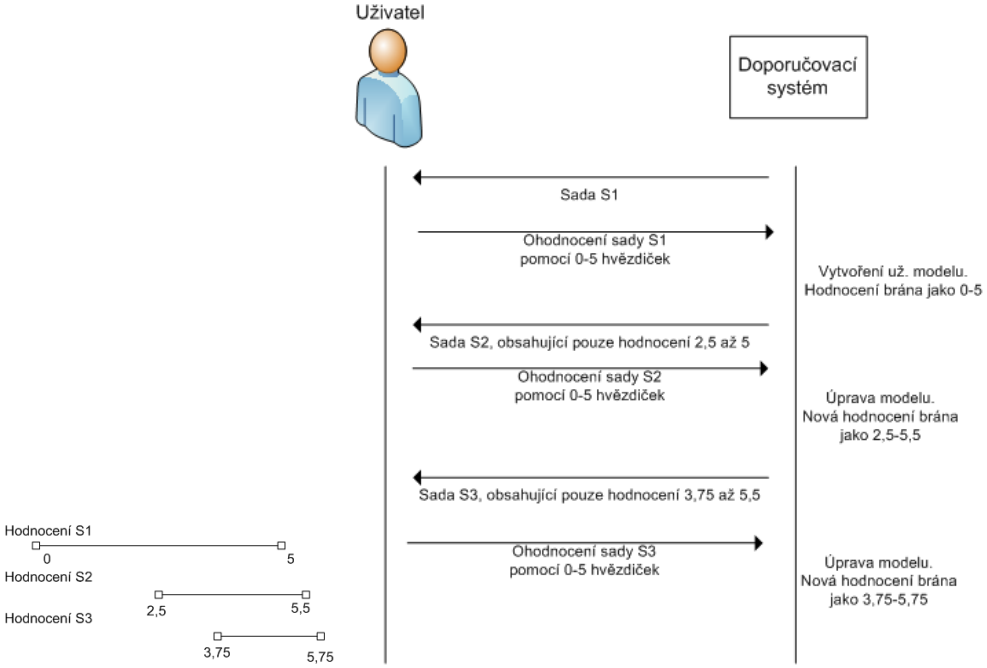
3.6 Motivace

Pro příklad vezměme explicitní zpětnou vazbu na e-shopu zabývající se prodejem fotoaparátů. První doporučená sada fotoaparátů je vybraná náhodně, nebo předem vybraná množina fotoaparátů, tak aby dobře pokryla druhy produktů. (Vojtáš, 2010)

Průběh procesu doporučení

1. Necháme uživatele, aby řekl, jak moc se mu líbí určité produkty
2. Zkonstruujeme obecný model preferencí (Nikon, rozlišení 8-10 Mpx)
3. Pomocí modelu zjistíme preference všech objektů (filtr)
4. Doporučíme 10 nejlepších objektů
5. Zpracujeme zpětnou vazbu od uživatele
6. Zpět na bod č.1

Každý další cyklus by měl vracet více relevantní produkty.



Obrázek 4: Fáze dotazování

Zdroj: (Vojtáš, 2010)

V každé fázi se posune hodnocení trochu doprava a tím zpřesňujeme hodnocení nejvíce preferovaných produktů. V ideálním případě dotazování probíhá až do stavu, kdy doporučené prukty budou mít 5 hvězdiček. (Vojtáš, 2010)

4 Typy doporučovacích systémů

Doporučovací systémy můžeme rozdělit na:

- Doporučení na základě vyhledávání
- Doporučení založené na kategoriích
- Kolaborativní filtrování
- Doporučení založené na obsahu
- Doporučení na základě omezení
- Doporučení na základě požadavků
- Hybridní techniky

4.1 Doporučení založené na vyhledávání

Hlavní výhodou tohoto typu je jednoduchost na implementaci. Zákazník zadává vyhledávací dotaz a systém vyhledá všechny položky, které odpovídají tomuto dotazu. Například uživatel zadá dotaz o zobrazení 6. nejoblíbenější knihy. Systém doporučí některou z těchto knih na základě všeobecného, neosobního hodnocení (podle prodejní pozice, popularity, atd.).

Výhody:

- Jednoduché na implementaci

Nevýhody:

- Ne příliš účinné
- Uživatel dostane jen to, na co se zeptal
- Nejedná se úplně o doporučení
- Žádná kritéria pro seřazení

4.2 Doporučení založené na kategoriích

Uživatel si vybere kategorii, která ho zajímá. Systém vybere kategorie zájmů pro zákazníka na základě aktuálně prohlížené položky, předchozích nákupů atd., a dokáže doporučit určité položky (ve slevě, nejprodávanější)

Výhody:

- Stále snadné na implementaci

Nevýhody:

- Opět se nejedná tak úplně o doporučení
- Není jasné podle čeho seřadit doporučení

4.3 Kolaborativní filtrování

Techniky kolaboratorního filtrování „porovnávají“ zákazníky na základě jejich předchozích nákupů. Na základě toho pak provádějí doporučení zákazníkům s podobnými nákupy. Tato metoda je také nazývána „sociální“ filtrování.

Kolaborativní filtrování je proces, při kterém dochází k filtrování informací na základě daných kritérií. Obvykle se používá pro velmi rozsáhlé množiny dat a pomáhá uživatelům se ve velkém množství dat lépe orientovat. Principem tohoto filtrování je vytvoření filtrovacího vzorce na základě dat získaných od velkého množství uživatelů a následné vytvoření předpovědi použitím tohoto vzorce na množinu dat, které je třeba filtrovat. (Cvengroš, 2011)

V dnešní době informační expanze stále narůstá počet položek v kategoriích např. (hudba, filmy, knihy, zájmové stránky) narostl do takového množství, že pro řadového uživatele (posluchač, divák, čtenář, kritik atd.) internetu je velmi obtížné prozkoumat, zhodnotit a vybrat si pro něj ty nejvhodnější. Pomocí kolaborativního filtrování je ale možné uživateli doporučit jen ty položky, které by pro něj měly být vhodné na základě předchozího chování tohoto i ostatních uživatelů. (Stružský, 2009)

	Kniha A	**Kniha B**	**Kniha C**	**Kniha D**	**Kniha E**
Zákazník 1	x				
Zákazník 2		x	x		
Zákazník 3					x
Zákazník 4		x	x		x
Zákazník 5					
Zákazník 6		x			

Tabulka 2: Tabulka prodeje knih

Kolaborativní přístup nevyžaduje žádnou znalost o položkách jako takových. Například o čem kniha je, nebo kdo je autorem. Jasnou výhodou tohoto systému je, že tato data o položkách nemusí být vkládána do systému nebo být v něm udržována, proto není potřeba žádná údržba (Vala, 2012)

Na příkladě v tabulce č. 2 můžeme vidět historii nákupů knih. Systém vyhledá podobné uživatele – Zákazník č. 2 se velmi podobá zákazníku č. 4 – mají společné předchozí nákupy (knihy B a C) navíc zákazník č. 4 má koupenou knihu E. Systém vyhodnotí, že kniha E by se mohla zákazníkovi č. 4 líbit a doporučí ji.

Metody kolaborativniho filtrování se mohou rozdělit do několika skupin:

- Kolaborativniho filtrování založené na paměti (Memory-based Collaborative Filtering)
  - Doporučení založené na podobnosti uživatelů
  - Doporučení založené na podobnosti položek
- Model-based Collaborative Filtering

Memory-based Collaborative Filtering

Memory-based metody jsou historicky první metody kolaborativního filtrování. Tyto metody předvídají možné vztahy, které jsou počítány na základě známých vztahů mezi objekty. Agregační funkcí může být prostý průměr nebo některé další sofistikované opatření využívající rozdíly průměrných hodnocení nebo individuální podobnosti. Jinými slovy využívají statistických metod k nalezení skupiny uživatelů známých jako sousedé, kteří mají podobnou historii cílového uživatele (tj., že buď je cena různých položek podobná, nebo mají tendenci koupit podobný soubor položek). Jakmile je sousedství uživatele vytvořeno, tyto systémy používají různé algoritmy ke kombinování preferencí sousedů k produkci doporučení pro aktivní uživatele (Cvengroš, 2011)

Doporučení založené na podobnosti uživatelů

Tato technika je založena na vybrání podmnožiny uživatelů na základě podobnosti s aktivním uživatelem. Poté se podle hodnocení této podmnožiny vypočítají doporučení pro aktivního uživatele. Postup může být shrnut do následujících kroků: (Melville P., 2010) (Cvengroš, 2011)

- Přiřazení váhy uživatelům, kteří se podobají cílenému uživateli.
- Selekce uživatelů s nejvyšší vahou podobnosti.
- Vypočítání předpovědi.

Prvním krokem je přiřazení váhy podobnosti ostatním uživatelům podle aktivního uživatele. Podobnost mezi dvěma uživateli se počítá pomocí Pearsonova korelačního koeficientu:

$$sim(a,u) = \frac{\sum_{p \in P}(r_{a,p} - \overline{r_a})(r_{u,p} - \overline{r_u})}{\sqrt{\sum_{p \in P}(r_{a,p} - \overline{r_a})^2}\sqrt{\sum_{p \in P}(r_{u,p} - \overline{r_u})^2}}.$$

kde $sim(a,u)$ je váha podobnosti, $U = \{u_1, \ldots, u_n\}$ je množina uživatelů, kde u značí uživatele a a značí aktivního uživatele. $P = \{p_1, \ldots, p_n\}$ je množina položek ohodnocena oběma uživateli, $r_{u,p}$ je hodnocení položky p uživatelem u a r_u je průměrné hodnocení uživatele u.

Druhým krokem je výběr uživatelů, kteří mají největší podobnost s aktivním uživatelem.

Ve třetím kroku se provádí výpočet předpovědi z vybraných uživatelských hodnocení. Tato předpověď se obvykle počítá pomocí vzorce:

$$pred(a,p) = \overline{r_a} + \frac{\sum_{u \in N} sim(a,u) * (r_{u,p} - \overline{r_u})}{\sum_{u \in N} sim(a,u)}$$

kde *pred(a, p)* je předpověď hodnocení aktivního uživatele a pro položku p a kde N je množina nejpodobnějších uživatelů – sousedství

Ačkoli tento přístup byl úspěšně použit v různých doménách, některé problémy přetrvávají. Problém nastane, když se tato metoda aplikuje na velké komerční webové stránky, kde musíme zpracovat miliony uživatelů a miliony položek z katalogu. Díky nutnosti kontroly velkého počtu potencionálních sousedů není možné vypočítat předpovědi hodnocení v reálném čase. Tento problém řeší metoda založená na podobnosti položek (Melville P., 2010) (Cvengroš, 2011)

Doporučení založené na podobnosti položek (Item based)

Hlavní rozdíl algoritmů doporučení založené na podobnosti položek, oproti algoritmům doporučení založené na podobnosti uživatelů, je vypočítání předpovědi použitím podobnosti mezi položkami a ne podobnosti mezi uživateli. Hlavní myšlenkou je, že pokud si uživatel koupil nějakou položku, předpokládáme, že by si mohl v budoucnu koupit položku podobnou. Analýzou historie nákupů uživatele můžeme tedy předpovědět, co si koupí v budoucnu. (Cvengroš, 2011) (Melville P., 2010)

Item-based algoritmy jsou dvoukrokové algoritmy, které se dají použít offline. V praxi to znamená rychlejší online systémy a často také kvalitnější doporučení.

Prvním krokem je zjištění podobnosti položek a, b pomocí Pearsonovi korelace vzorcem:

$$sim(a,b) = \frac{\sum_{u \in U}(r_{u,a} - \overline{r_a})(r_{u,b} - \overline{r_b})}{\sqrt{\sum_{u \in U}(r_{u,a} - \overline{r_a})^2}\sqrt{\sum_{u \in U}(r_{u,b} - \overline{r_b})^2}}$$

kde U je množina všech uživatelů, kteří hodnotili obě položky a a b, $r_{u,a}$ je hodnocení uživatele u položky a a $r_{u,b}$ je hodnocení uživatele u položky b a kde r_a a r_b je průměrné hodnocení položek a a b

Druhým krokem, poté, co jsou vypočítány podobnosti položek, je předpověď hodnocení položky a pro uživatele u pomocí váženého průměru

$$pred(u,a) = \frac{\sum_{b \in K} r_{u,b} * sim(a,b)}{\sum_{b \in K} |sim(a,b)|}$$

kde K je množina sousedních položek, hodnocených uživatelem u, které jsou nejvíce podobné položce a

4.4 Doporučení založené na obsahu (Content-based)

Na rozdíl od kolaborativního přístupu se využívá informací o zálibách uživatele a o doporučovaných produktech. V běžném životě je velmi přirozené doporučit uživateli např. seriál Battlestar Galactica, pokud víme, že Battlestar Galactica je z žánru sci-fi a uživatel má rád sci-f a líbí se mu i Star Trek. V elektronickém doporučovacím systému k takovému doporučení stačí jen hrstka informací: popis charakteristiky položky (knihy, filmu) a uživatelův profil, který popisuje jeho zájmy (oblíbené knihy, filmy a žánry, herce). I když tento přístup závisí na dalších informacích o položkách a uživatelích, nepotřebuje uživatelskou základnu. K tomu, aby vygeneroval doporučení stačí systému i jeden uživatel. (Cvengroš, 2011)

Proces doporučení je vykonán ve třech krocích a o každý se stará jiná komponenta:

1. **Analyzátor obsahu** – Hlavním úkolem této komponenty je vytvořit popis položky z informací o položce, nebo jejím obsahu, tak aby byl tento popis vhodný pro další krok.

2. **Profilová komponenta** – Tento modul shromažďuje data, která reprezentují uživatelovy volby, snaží se je generalizovat, a vytvořit tak uživatelův profil.

3. **Filtrovací komponenta** – Tato komponenta doporučuje relevantní položky na základě uživatelova profilu.

Systém vytvoří doporučení podle toho, jak moc je prozatím neviděná položka podobná těm, které se uživateli líbily v minulosti. Podobnost položek se dá vypočítat různými způsoby, stačí například, aby žánr nové knihy byl v oblíbených žánrech uživatele. Dalším způsobem je spočítat doporučení podle klíčových slov. (Cvengroš, 2011)

Podobnost mezi knihami b_1 a b_2 můžeme spočítat pomocí S?rensen–Dice koeficientu:

$$QS = \frac{2|K(b_1) \cap K(b_2)|}{|K(b_1)| + |K(b_2)|}$$

kde QS je podíl podobnosti a pohybuje se od 0 do 1. a K označuje množinu klíčových slov.

Výhodou tohoto přístupu je nezávislost na ostatních uživatelích, jediné co potřebuje, je hodnocení položek aktivního uživatele, nepotřebujeme hodnocení od ostatních. Mezi další výhodu se dá počítat i průhlednost celé metody. Je snadno vysvětlitelné jak byla položka vybrána k doporučení uživateli, ten se podle muže rozhodnout, zda je pro něj doporučení důvěryhodné. Pro uživatele to může být lepší zdůvodnění, než že neznámému uživateli s podobným vkusem se líbila stejná položka. Další výhodou je, že systém je schopen doporučit i úplně novou položku, kterou ještě nikdo předtím nehodnotil. (Vala, 2012)

Nevýhodou je, že automatické i manuální metody přiřazení rysů položkám nemusí být dostatečné k vygenerování vhodného doporučení pro uživatele. Například automatická extrakce obsahu nové stránky naprosto ignoruje estetické kvality. Dalším problémem je, že díky přílišné specializaci uživatel nedostane neočekávané doporučení, každé doporučení bude velmi podobné dříve hodnoceným položkám. Uživatel, který dostává doporučení ze žánru sci-fi nedostane doporučení s akčním filmem i přes to, že by o ní mohl mít potencionálně zájem. Dalším problémem je, že nový uživatel musí ohodnotit dostatečné množství položek, aby mu bylo doporučeno vhodná položka. (Vala, 2012) (Cvengroš, 2011)

4.5 Doporučení založené na znalostech (Knowledge-based)

Tento přístup využívá znalostí o uživatelích a položkách k vygenerování doporučení. Předchozí přístupy jsou vhodné pro doporučování produktů, (knihy, filmy, hudba atd.). Ale při doporučování například aut, počítačů, bytů nebo finančních služeb nejsou nejlepší volbou. Důvodem je, že například u pojištění je nemožné získat dostatek hodnocení, protože není dostatek stejných exemplářů, nebo že uživatel by určitě nebyl spokojený s doporučením počítače podle rok starých hodnocení. (Vala, 2012)

Doporučení založené na znalostech se nepotýká s žádným z problémů předchozích přístupů, protože k doporučení nepotřebuje žádná hodnocení. Doporučení je vypočítáno individuálně pro každého uživatele a nezávisle na ostatních. Jedná se vlastně o interaktivní filtrovací systém, kde uživatel zadá, co potřebuje a systém mu dodá doporučení. To je ale i jeho slabinou, uživatel musí zadat co chce a to taky dostane, nebude mu doporučena žádná další položka než aktuálně prohlížená. (Cvengroš, 2011)

Existují dva základní přístupy: doporučení na základě omezení nebo na základě požadavků. Oba jsou si podobné tím, že uživatel musí specifikovat své po-

žadavky a systém pak dodá výsledek. Rozdílem je, že doporučení na základě požadavků hledá co nejpodobnější položky, zatímco doporučení na základě omezení využívá znalostní základnu, která přímo obsahuje pravidla, jak se chovat vzhledem k potřebám daného zákazníka (Vala, 2012) (Cvengroš, 2011) (Melville P., 2010)

4.6 Doporučení na základě omezení (Constraints)

Tento doporučovací systém je typicky definován dvěma množinami proměnných (V_C, V_{PROD}), které popisují zákazníkovy požadavky a vlastnosti produktu, a třemi množinami omezení (C_R, C_F, V_{PROD}), které definují, jaké položky by měly být doporučeny v jaké situaci. K vysvětlení těchto proměnných používám příklady z oblasti prodeje notebooků:

- **Zákazníkovy požadavky** (V_C) – Popisuje požadavky zadané uživatelem, například počet jader procesoru, rozlišení displeje (1600 x 900 (HD+)).
- **Vlastnosti produktu** (V_{PROD}) – Popisuje vlastnosti produktu, například rozlišení displeje, frekvence procesoru (MHz), výrobce grafického čipu.
- **Omezení** (C_R) – Definuje povolené výběry požadavků uživatele, například velikost úložného prostoru větší než 750 GB, bílá barva a cena maximálně 17 000 kč.
- **Filtrovací podmínky** (V_F) – Definuje, za jakých podmínek by měl být výrobek vybrán, například notebook s 8GB RAM může být vybrán pouze pokud je jeho cena maximálně 15 000 kč.
- **Omezení produktů** (V_{PROD}) – Definuje, které produkty jsou v současnosti dostupné.

Když jsou všechny tyto proměnné vyplněny, doporučení je pak už jednoduché. Zákazník si například zadá, že chce notebook s 8GB RAM, cenu maximálně 25 000 kč, úhlopříčku displeje 17"(43 cm) a systém buď takový produkt najde, nebo napíše, z jakého důvodu nebyl žádný nalezen. (Vala, 2012)

4.7 Doporučení na základě požadavků (Cases)

V tomto přístupu jsou položky doporučovány podle toho, jak moc jsou podobné uživatelovým požadavkům. V praxi jsou některé proměnné, které chce uživatel maximalizovat, například rozlišení monitoru, a některé, které chce minimalizovat, jako například cenu. Tyto požadavky musejí být brány v potaz, takže vzorec k vypočítání podobnosti mezi atributem položky e a požadavky uživatele r pro případ, že uživatel chce danou položku maximalizovat, vypadá takto: (Cvengroš, 2011) (Melville P., 2010)

$$sim(p,r) = \frac{\phi_r(p) - min(r)}{max(r) - min(r)}$$

Podobnost mezi atributem položky p a požadavky uživatele r pro případ, že uživatel chce danou položku minimalizovat, vypadá takto:

$$sim(p,r) = \frac{max(r) - \phi_r(p)}{max(r) - min(r)}$$

Jsou zde také situace, kdy je vhodné podobnost založit pouze na vzdálenosti argumentů od sebe, například když chce uživatel určitou velikost monitoru. Tento případ spočítáme pomocí třetího vzorce:

$$sim(p,r) = 1 - \frac{|\phi_r(p) - r|}{max(r) - min(r)}$$

Pro vytvoření doporučení uživatel zadá svoje požadavky do systému a systém vyhledá produkty, které nejvíce odpovídají uživatelovým požadavkům. Když není uživatel spokojen, může modifikovat požadavky, a tím začne nový cyklus doporučování. (Vala, 2012)

4.8 Hybridní techniky

Všechny dříve zmíněné mají nějaké nedostatky a slabiny, které mohou odrazovat od jejich zavedení. Hybridní doporučovací systémy proto kombinují dvě, nebo více přístupů a snaží se tím zdokonalit výsledné doporučení. Ve většině případů se jedná o kombinaci s kolaborativním filtrováním a snaží se tak vyhnout problému s novými uživateli, nebo položkami. (Vala, 2012) (Burke, 2002) (Cvengroš, 2011)

4.8.1 Vážený hybridní doporučovací systém

V tomto přístupu je použito skóre ze všech dostupných doporučovacích technik v systému a pomocí něj se spočítá celkové doporučení. Nejjednodušší možností je použít lineární kombinaci všech výsledků, ale v jistých situacích může být vhodnější, dát některým výsledkům váhu větší, než jiným. Výhodou tohoto systému je, že potenciál všech technik může být vyžit k vygenerování doporučení a je relativně jednoduché tento hybridní systém zavést. Problémem může být, že ne všechny techniky jsou si ve všech situacích rovny, ale v tomto přístupu se často hodnotí ekvivalentně. Například kolaborativní doporučení bude mít slabší váhu při doporučování položek s malým počtem hodnocení. (Vala, 2012) (Burke, 2002)

4.8.2 Přepínací hybridní doporučovací systém

V tomto přístupu systém použije nějaké kritérium k přepínání mezi doporučovacími technikami. Například když první přístup nedokáže udělat dostatečně věrohodné doporučení, použije se druhý. Často se používá kolaborativní doporučení a doporučení podle obsahu, i když neřeší problém s novými uživateli. Kolaborativní doporučení totiž může dodat nečekané výsledky, které by se určitě podle podobnosti položek neobjevily. Nevýhodou je, že první technika je použita vždy, zatímco druhá jen když první selže. Tento přístup je také složitější, protože musí být definována přepínací kritéria. (Burke, 2002)

4.8.3 Smíšený hybridní doporučovací systém

Smíšený systém využívá doporučení z více než jedné techniky a prezentuje je dohromady. Když použijeme najednou kolaborativní doporučení a doporučení podle obsahu, vyhneme se problému nové položky, protože se s ním vypořádá druhá technika. Ale ani tato metoda se nedokáže vypořádat s problémem nového uživatele, oba zmíněné přístupy potřebují určité množství informací o uživateli. (Burke, 2002)

5 Algoritmy kolaborativního filtrování

Kolaborativní filtrování představuje často používanou metodu porovnání několika uživatelů nebo určení preferencí uživatele. To je založeno na předpokladu, že uživatelská preference uživatele u_0 pro objekt o bude stejná, jako u uživatelů $u_1, \ldots, u_k$, kteří jsou uživateli u_0 podobní. Tato podobnost uživatelů je vypočtena z podobnosti hodnocení různých objektů uživateli. Pro výpočet je nutné získat mnoho hodnocení objektů od mnoha různých uživatelů. K výpočtu podobnosti uživatelů, je třeba nashromáždit mnoho hodnocení objektů, pro přesnou předpověď vhodnosti objektu pro uživatele je třeba nalézt mnoho podobných uživatelů. Existuje několik různých algoritmů kolaborativního filtrování. (Stružský, 2009)

5.1 Algoritmus K-NN (K nearest neighbours)

První, nejjednodušší a také nejvíce intuitivní a snadno pochopitelný je algoritmus K-NN (Houdek, Svoboda, & Procházka, 2001). Pro uživatele u_0 vyberu K nejbližších sousedů $u_1, \ldots, u_k$, přičemž vzdálenost uživatelů se pro tento algoritmus určuje podle vzorce

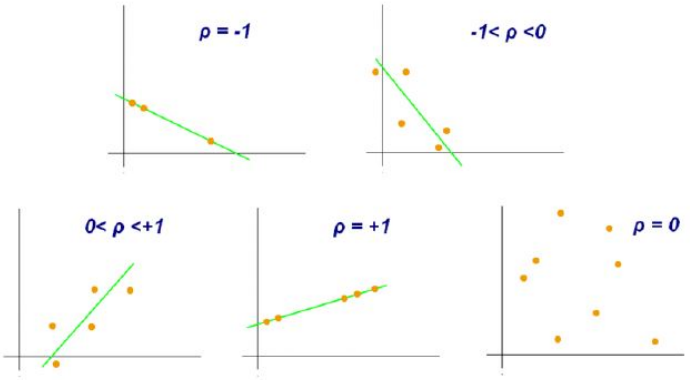
$$d(u_a, u_b) = \sqrt{\sum_{i=1}^{n} (P_a(o_i) - P_b(o_i))^2}$$

kde d je vzdálenost (distance) uživatelů u_a a u_b, n je počet porovnávaných objektů o_i a P_o je hodnocení daného objektu. Je-li tedy zvoleno K nejbližších sousedů, lze hodnocení porovnávaných objektů pro uživatele u_0 vypočítat jako aritmetický průměr hodnocení vybraných uživatelů: (Stružský, 2009) (Houdek et al., 2001)

$$P_o(o) = \frac{\sum_{i=1}^{K} P_i(o)}{K}$$

5.2 Pearsonův korelační koeficient

Pearsonův korelační koeficient je míra korelace (lineární závislosti) mezi dvěma proměnnými X a Y. Je široce používán ve vědách jako měřítko síly lineární závislosti Vyvinul jej Karl Pearson. Korelační koeficient se pohybuje mezi hodnotami 1 a -1 včetně. Hodnota 1 znamená, že lineární rovnice popisuje vztah mezi X a Y dokonale, všechny body leží na přímce a přímka je stoupající. Hodnota -1 znamená, že všechny datové body leží na přímce, pro kterou platí, že je klesající. Hodnota 0 znamená, že neexistuje žádný lineární vztah mezi proměnnými. (Houdek & Svoboda, 2001)



Obrázek 5: Příklady diagramů s různými hodnotami korelačního koeficientu (p)

Zdroj: answers.com

Těchto mezních hodnot ovšem Pearsonův korelační koeficient nabývá velmi zřídka. Hodnotu tohoto koeficientu lze určit jako podíl míry vzájemné vazby mezi dvěma veličinami (kovariance) sledovaných proměnných a jejich směrodatných odchylek. Kovariance je definována jako střední hodnota součinu rozdílu sledovaných proměnných od jejich středních hodnot. Vzorec pro výpočet je tedy následující: (Houdek & Svoboda, 2001) (Stružský, 2009)

$$r = \frac{cov(X,Y)}{s_X s_Y} = \frac{E((X_i - E(X))(Y_i - E(Y)))}{s_X s_Y} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_X}\right) \frac{(Y_i - \overline{Y})}{s_Y}$$

Parametr n je tzv. stupeň volnosti, tedy počet hodnot použitých k výpočtu (pro výpočet odchylky se ve jmenovateli vždy počítá s hodnotou n - *1*), $X_i Y_i$ jsou konkrétní hodnoty, $\overline{X}$ a $\overline{Z}$ jsou průměrné hodnoty a $s_X s_Y$ jsou směrodatné odchylky.

5.3 Spearmanův korelační koeficient

Jde o neparametrickou metodu, která při výpočtu využívá pořadí hodnot sledovaných veličin, nevyžaduje tedy normalitu dat. Spearmanův korelační koeficient používáme nejčastěji pro měření síly vztahu u takových veličin, u kterých nemůžeme předpokládat linearitu očekávaného vztahu nebo normální rozdělení sledovaných proměnných X a Y. Pro malé rozsahy n je výpočet Spearmanova korelačního koeficientu méně pracný než výpočet Pearsonova parametrického korelačního koeficientu. Proto je možno ho použít i k hodnocení lineárních závislostí, kde je jeho použití spíše orientační (využívá méně informací z dat) a na rozdíl od parametrického koeficientu je méně účinný. (Stružský, 2009) (Rozenberg, 2013)

Výpočet Spearmanova korelačního koeficientu vychází z pořadových čísel proměnných X_iY_i (korelačních dvojic) naměřených u n jedinců výběrového souboru. Jsou-li hodnoty proměnných X_iY_i seřazeny vzestupně do dvou řad a každé hodnotě je přiděleno pořadí, pak koeficient pořadové korelace je dán vztahem:

$$p = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

kde $d_i$di je rozdíl mezi pořadím hodnot x_iz_i příslušných korelačních dvojic a n je počet korelačních dvojic. (Stružský, 2009) (Rozenberg, 2013)

5.4 Euclidean distance - Euklidovská vzdálenost

V matematice je Euklidova vzdálenost vzdálenost mezi dvěma body v Euklidově prostoru, díky této vzdálenosti se Euklidovský prostor stává metrickým prostorem

Mějme dva body (x, y) se souřadnicemi $((x_1, x_2), (y_1, y_2))$, Euklidovská vzdálenost je známa z Euklidovské geometrie: (WolframMarhWorld, 2013)

$$((x_1, x_2), (y_1, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + ... + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} \sqrt{(q_i - p_i)^2}}$$

6 Problém neúplných dat

Mnoho analytiků se setkává s případy, kdy zdrojová data určená pro analýzu obsahuje chybějící údaje. Tento jev značně snižuje schopnost doporučit uživateli vhodnou položku. Je třeba odlišit, zda atribut, např. typ mobilního telefonu je chybějící korektně, nebo nekorektně. Ke korektní nenaplněnosti atributu dochází v případu, kdy uživatel tento údaj nemůže vyplnit, protože ho nevlastní (ne každý uživatel musí vlastnit mobilní telefon nebo emailovou adresu). V druhém případě daný údaj k dispozici potenciálně je, ale díky chybě je nemůže systém zpracovat:

- Pro uživatele nesrozumitelná otázka (na jaké frekvenci pracuje Vaše RAM)
- Chyba vkládání dat (uživatel špatně vyplní údaj – neexistující město/ulice)
- Chyba v přenosu dat

Součástí každé analýzy by měla být fáze přípravy dat ve které je nutné vyřešit přítomnost chybějících dat. Známe několik možností jak se s chybějícími daty vypořádat, a to je vypustit, ignorovat, nebo doplnit. V případě vypuštění dat potencionálně ztrácíme důležitou informaci pro budoucí doporučení, proto není nejvhodnější. Data málokdy chybějí náhodně, a proto pokud se rozhodneme v ponechání chybějících/nekompletních dat může dojít ke zkreslení doporučení pro uživatele. Proto nejvhodnější metodou je doplnění dat. (Magnani, 2004)

6.1 Problém studeného startu

Studený start je stav, při kterém je do doporučovacího systému přidán nový uživatel nebo nová položka o kterém nemá systém dostatek dat. Tento uživatel tedy dostane nepřesná nebo úplně chybná doporučení. Nově přidaná položka pak není doporučována.

Tento problém je obvykle řešen požádáním uživatele o vyplnění dotazníku, kde ohodnotí přiměřené množství položek pro výpočet nejvhodnějšího doporučení, nebo hybridní formou doporučení, tedy pro nového uživatele je doporučena nejoblíbenější položka. U nové položky je viditelné označení „novinka“, toho označení pomůže položce nasbírat údaje pro budoucí doporučení. (Adomavicius, 2005)

6.2 Problém řídkosti matic hodnocení

Téměř žádný uživatel nehodnotí všechny položky, není možné aby řadový uživatel viděl všechny filmy, nebo vyzkoušel všechny produkty, které jsou obsaženy v databázi, a proto je v matici hodnocení spousta prázdných míst. Jedná se o problém problém pro kolaborativní doporučovací systém, protože je těžší najít množinu uživatelů s podobným hodnocením. Tento problém se nejčastěji objevuje při spuštění nového systému, nebo se jedná o systém ve kterém se nachází mnohem více položek, než uživatelů. Řešením může být použití hybridní techniky, například využít ještě hodnocení na základě obsahu. (Burke, 2005)

6.3 Problém s novými položkami a uživateli

Problém s novými položkami se objevuje v kolaborativním doporučování, kde položka nemůže být doporučena, dokud ji nějaký uživatel neohodnotí, nebo nekoupí. Tento problém nemusí být jen s novými položkami, ale může být i s položkami méně známými. Řešením je využít pro takovéto předměty hodnocení na základě obsahu, které může zpracovat všechny položky. Těžší je vypořádat se s problémem nového uživatele, protože bez předchozích uživatelských hodnocení je nemožné najít podobné uživatele nebo položky podobné těm, které se uživateli líbily. Jedním z řešení je navrhnout mu vhodné položky k ohodnocení např. 3 filmy z každého žánru, aby měl doporučovací systém z čeho čerpat. (Burke, 2005)

6.4 Problém s podvody

Za podmínky, kdy je každý uživatel v systému poctivý bylo by kolaborativní doporučování velmi dobře funkční a kvalitní metoda doporučení, ale v reálném světě je tento systém náchylný na podvody, protože je založen na informacích od uživatele. V prodejích na online portálech jde o velké peníze a tyto systémy ověřeně ovlivňují prodeje a někteří výrobci se snaží uměle navyšovat. Nejčastějším útokem na kolaborativní systémy je profilová injekce, jedná se o množství podvodných profilů, které se snaží uměle navýšit nebo snížit hodnocení, návštěvnost, či počet odběratelů (youtube.com, facebook.com). Tímto chováním se narušuje přesnost doporučení, jedná se druh útoku na informační systém. Podle facebook.com mezi lety 2012 a 2014 bylo smazáno přes 5% celkových uživatelů, protože byli vyhodnoceni jako černé duše, tedy uživatelé, kteří vznikli jen za účelem uměle navýšit popularitu cílového profilu.

7 Analýza doporučovacích systémů

7.1 Doporučovací systém na last.fm

Doporučovací systém na portálu Last.fm vytváří profil hudebního vkusu každého uživatele, ukazuje jeho oblíbené umělce a písničky na jeho osobní stránce. Tyto informace získává ze skladeb, které uživatel poslouchal pomocí stáhnutého Audioscrobbler pluginu nainstalovaného do jeho hudebního přehrávače nebo do internetového prohlížeče pro záznam přehrávání na youtube.com. Uživatelé Last.fm si mohou vytvořit hudební profil použitím tří metod: posloucháním své hudební sbírky v jejich přehrávači s nainstalovaným pluginem, posloucháním Last.fm rádia nebo na portálu youtube.com. Přehrané písničky se ukládají do databáze, ze které se sestavují žebříčky a hudební doporučení. Uživatelské stránky také zobrazují nedávno hrané skladby, které jsou dostupné i ve formě obrázku nebo XML dokumentů, a tak se dají zobrazit i na blozích a osobních stránkách.

druh doporučení

Doporučení se počítají pomocí kolaborativního filtrovacího algoritmu, takže si uživatelé mohou prohlížet seznamy umělců, které nemají ve svém vlastním profilu, ale mají je v profilech jiní uživatelé s podobným hudebním vkusem. Last.fm také umožňuje ručně doporučovat umělce, písničky, nebo alba dalším uživatelům (pokud jsou v databázi). K přesnějšímu zařazování umělců, písniček a alb slouží značky (tagy), tagy vytváří registrovaní uživatelé. Tagovat se může podle žánru („alternative rock“), nálady („melancholy“), charakteristiky interpreta („legends“) nebo podle libovolného přání tagera („můj playlist“).

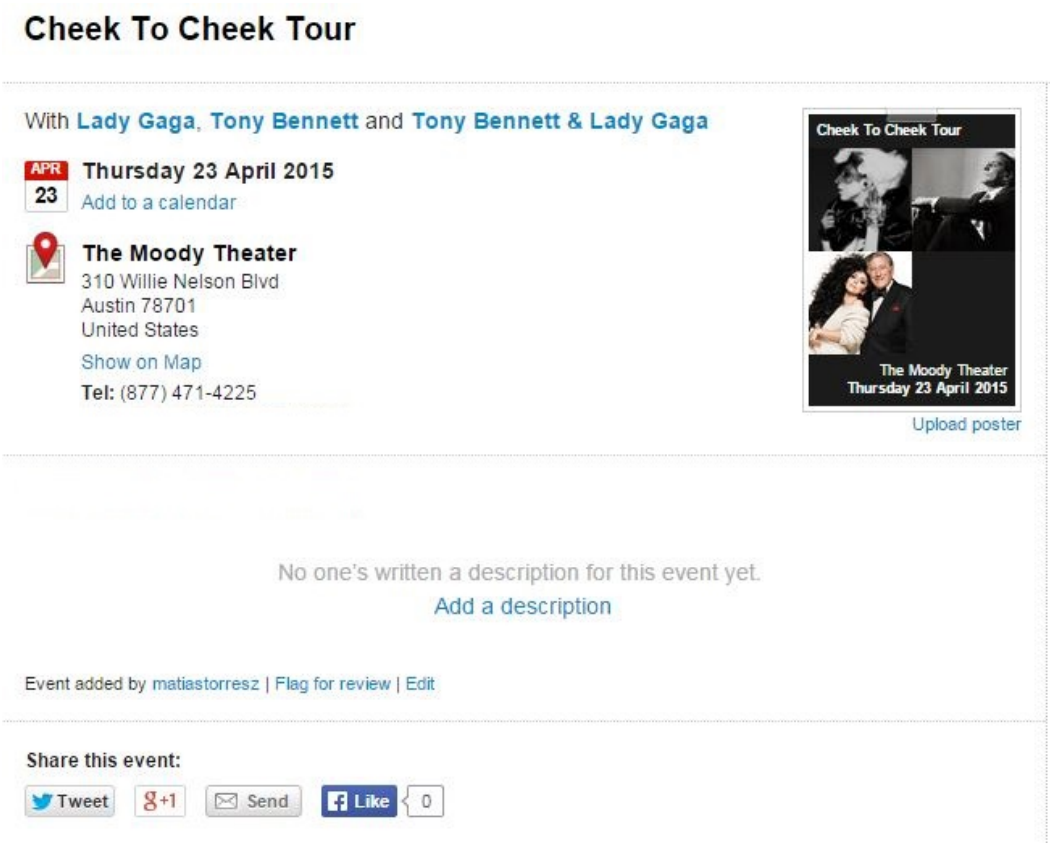
Další funkcí Last.fm je doporučení nově vydaných alb oblíbených interpretů, doporučení nadcházejících koncertů, nebo hudebních festivalů na kterých vystoupí oblíbení, potencionálně oblíbení interpreti, nebo festival spadá do uživatelovo hudebního vkusu. Last.fm nabízí i formu placeného přístupu, kdy za $3.00 měsíčně uživatel získá výhody: Prohlížení bez zobrazování reklam, zobrazení návštěvníků na jeho profilu, unikátní ikonka a přístup do beta testování nových funkcí.

Zhodnocení

Podobně jako Amazon.com je hudební katalog Last.fm dostupný pro velmi málo jazyků, v současné chvíli pro: angličtinu, němčinu, španělštinu, francouštinu, italštinu, polštinu, portugalštinu, ruštinu, švédštinu, turečtinu, japonštinu a čínštinu.

Další nedostatek shledávám v sekci hudebních akcí (obrázek č. 6), kde na stránce události jsou informace o vystupujících interpretech, času a místě události, ale nikoliv možnost dostat se přímo na webové stránky pořadatele nebo možnost zakoupit vstupenky přes předprodejní sítě. Některé události sice obsahují odkaz na webové stránky místa události, ale pokud se jedna např. o hudební festival pořádaný na letišti jsou webové stránky letiště nesouvisející a pro uživa-

tele bezcenné.



Obrázek 6: hudební událost na Last.fm

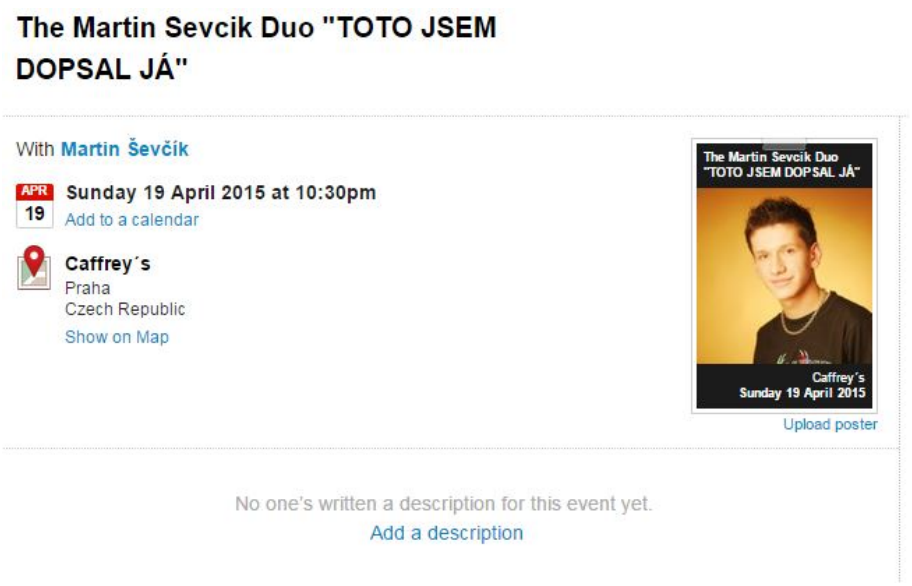
Další slabinou událostí je možnost editovat události libovolným uživatelem last.fm, muže tak lehce dojít k zlomyslnému pozměnění události (změna data, interpretů, názvu, místa události, nebo smazání celé události) a znehodnocení doporučení uživatelem záškodníkem. Pro ilustraci jsem se na obrážku č. 7 stal uživatelem záškodníkem a nakrátko jsem pozměnil název události. Tímto pozměňováním může dojít k chybnému a tudíž bezcenému doporučení.

Návrh

Podobně jako u problému slabé jazykové podpory na webu Amazon.com, tak i na webu Last.fm by řešením mohlo být analyzovat přístupy na web a u nejpočetnějších přístupů ze zemí, které nejsou jazykově podporovány vytvořit vhodné jazykové prostřední pro rozvoj webu Last.fm

Řešením nedostatečného zabezpečení proti uživatelům „záškodníkům“ by bylo nutnost každou aktualizaci události potvrdit moderátory speciálně vyčleněnými na tuto problematiku, další možností je nutnost schválení aktualizace uživatelem, který událost vytvořil u kterého je předpoklad, že je v jeho zájmu, aby údaje byly správné.

Problém nedostatečných informací u naprosté většiny událostí by byla větší provázanost se zástupci interpretů kteří ve vlastním zájmu vyšší návštěvnosti



Obrázek 7: znehodnocení události

hudebních událostí doplní potřebné marketinkové údaje. Další možností je nábor většího počtu moderátorů, kteří by se starali o hudební události v regionu, zde ale nastává problém s regiony, kde má Last.fm slabší uživatelskou základnu díky slabé jazykové podpoře.

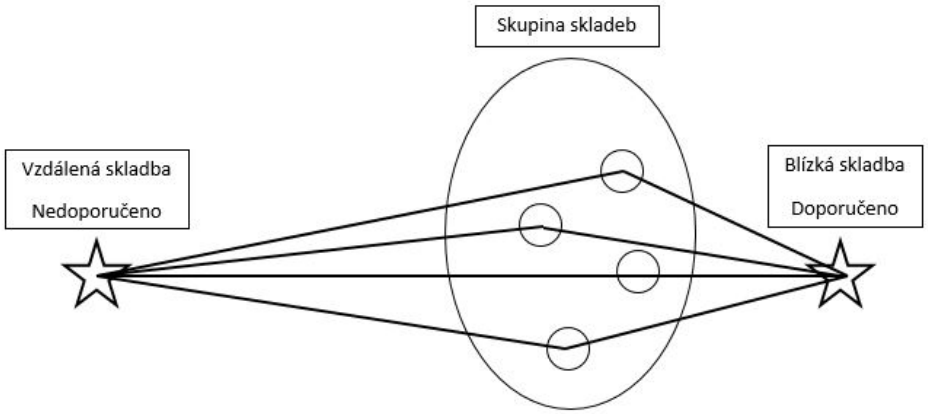
7.2 Doporučovací systém na Pandora.com

Pandora je internetové radio, obsahující také systém pro doporučení hudby, je k dispozici pouze ve Spojených státech, Austrálii a na Novém Zélandu, v ostatních zemích je pandora.com kvůli autorským zákonům nedostupná, ale díky existenci proxy serverů je běžně používána. Pandora, na rozdíl od většiny podobných služeb, nebere ohled na podobnost uživatelské preference a výsledné doporučení tvoří na základě obsahu přehrávaných skladeb, je tudíž systémem založeným na obsahu.

Music Genome Project

Pandora využívá databáze hudebních skladeb Music Genome Project, kde každá skladba je reprezentována vektorem, který obsahuje přibližně 150 charakteristik (genů). Každý gen odpovídá jedné charakteristice skladby jako: výška hlasu, množství elektrických kytar, tempo, atd. Rocková a popová hudba má 150 genů, rapová hudba - 350 a jazzová hudba má přibližně 400. Ostatní žánry hudby, jako například klasická, mají 300 až 500 genů.

Každý gen je číslo mezi 0 a 5, zlomkové hodnoty jsou povoleny. Systém počítá jednoduchou vzdálenost mezi libovolnými dvěma skladbami. Pro určenou skladbu systém nalezne skladby které jsou jí nejvíc podobné. Systém je také schopen poskytnout množinu podobných skladeb pro určenou skupinu skladeb. Obrázek 8



Obrázek 8: kandidáti na doporučení

znázorňuje vzdálenost dvou skladeb, skladba napravo je po výpočtů genů menší vzdálenost, než skladba nalevo a je výsledkem doporučení.

Zhodnocení

Doporučení na základě výpočtu podobnosti genů doporučí uživateli vždy podobnou skladbu jakou zadal při začátku přehrávání, doporučení se tedy muže zacyklit a doporučovat uživateli stále stejné interprety stále dokola. Pro příklad: při zadání kalepy System Of a Down bude přehrávač přehrávat hlavně 4 další interprety (Serj Tankian, Rage Against The Machine, KoRn, Linkin Parn), tedy kapely kterým se hudba podobá, ale nedoporučí kapelu, která má skladby na podobné téma, jako např. The Smashing Pumpkins. Obě tyto kapely se prezentují válečnou kritikou, ale jejich hudba je rozdílná. Uživatel tedy bude poslouchat stále podobné skladby, ale ve výsledku mu nemusí být doporučeno nic nového.

Návrh

Při přehrávání live radia na pandora.com je možnost hodnotit skladby pomocí funkce „thumb up“, nebo „thumb down“, neshledávám to jako dostačující opatření pro záznam uživatelské preference, z důvodu, že poslouchání radia na pandora.com je mnoho uživatelů motivováno nenutností jakkoliv zasahovat do přehrávání a doporučené skladby přepínat. Jako vhodnou alternativu shledávám při prvním zadávání kritéria např. (male vocalist, Frank Turner, Madonna) dotázat se uživatele jaké skladby si přeje přehrávat, tím systém získá rozšiřující informace o uživatelově preferenci v době kdy uživatel chce aktivně ovlivnit co mu bude doporučeno.

Rovněž by bylo vhodné rozšířit služby o možnost offline poslech hudby a doporučení playlistu čítající několik skladeb (20-30), které si uživatel muže stáhnout do svého přenosného zařízení, ideálně do aplikace od pandora.com, která bude zaznamenávat které skladby si uživatel přehrává častěji a použít tyto informace

pro budoucí přesnější doporučení.

7.3 Doporučovací systém na CSFD.cz

Doporučení na Česko-Slovenské filmové databázi se zakládá pouze na doporučení na základě kategorií, když vezmeme v úvahu, že celé fungování CSFD.cz je založené na dobrovolném hodnocení filmů, tak toho jednoduché neindividualizované doporučení (nejnavštěvovanější videa, nová videa, novinky) je nedostatečné a nevyužívá potenciálu, který poskytuje široká uživatelská základna s touhou vyjádřit svoje preference. Uživatel má možnost vyplnit své oblíbené herce, režiséry, scénáristy, ale kromě informace pro ostatní uživatele při prohlížení profilu tyto údaje nejsou nijak využity.

Návrh

Můj návrh spočívá v přechodu na hybridní doporučování, a to zachovat doporučení na základě kategoriích a doplnit o kolaborativní filtrování. Využitím kolaborativního filtrování se otevřou nové možnosti pro doporučení filmových novinek, které by se uživateli mohli potencionálně líbit a po navázání spolupráce s kiny uživateli zobrazit nejbližší kino, kde se doporučený film bude promítat. Stejně tak vidím potenciál ve spolupráci s online videopůjčovnama a doporučovat vypůjčení filmu, který uživatel nemá ohodnocen a je v uživatelovo okruhu možného zájmu. Upozornění na nový film/seriál s hercem/herečkou, kteří jsou označeni za oblíbené rovněž chybí. Implementace doporučení novinky s oblíbeným tvůrcem je poměrně snadná na implementaci, měla by proto být prvním krokem k vylepšení a k zefektivnění toho portálu.

8 Aplikace na doplňování chybějících dat

8.1 Úvod

Praktická část je zaměřena na problematiku chybějících dat, pro příklad co s uživatelem který o sobě nevyplní všechny údaje. Tento problém řeší aplikace na doplňování chybějících dat tak, že porovná vyplněná data uživatele s uživateli v db, kteří jsou mu nejbližší. K vypočtení vzdálenosti (blízkosti) jednotlivých uživatelů je využit algoritmus k-nejbližších sousedů, neboli K-NN.

Pro prezentaci bylo zvolena problematika preference filmů, kde je po uživateli požadováno vyplnění jména, věku, výběr oblíbených filmových žánrů a výběr oblíbeného způsobu sledování filmů, přičemž nevyplní alespoň 1 požadovaný údaj a program doplní nejpravděpodobnější záznam.

8.2 Analýza

Před zahájením tvorby bylo nutné zvolit jazyk pro programování. Zvolil jsem C#, kvůli předchozím zkušenostem s tímto jazykem a k tvorbě formuláře WPF. Vstup je načítán z formuláře WPF (Windows Presentation Foundation), data pro porovnání uživatelů jsou načítány z databáze kompletních uživatelů a výstupem je databáze uživatelů.

C# a WPF

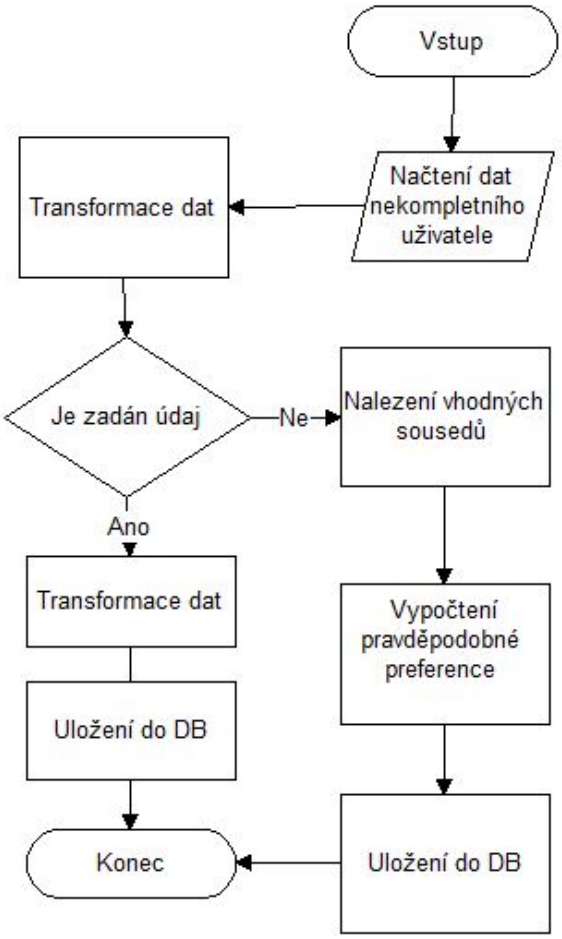
C# je programovací jazyk navržený pro vytváření různorodých aplikací, které běží v rozhraní .NET Framework. Jazyk C# je jednoduchý, výkonný, typově bezpečný a objektově orientovaný. Řada inovací v jazyce C# umožňuje rychlý vývoj aplikací a zároveň zachování expresivity a elegance jazyků stylu C. Visual C# je implementace jazyka C# společností Microsoft. Visual Studio podporuje Visual C# s kód plně vybavený editor, kompilátor, šablony projektu, návrháři, průvodcům, výkonné a snadno použitelné ladicí program a další nástroje. Knihovna tříd rozhraní .NET Framework poskytuje přístup k řadě služeb operačního systému a dalším užitečným a dobře navrženým třídám, které umožňují výrazné urychlení vývoje. (Microsoft, 2015)

WPF (Windows Presentation Foundation) je framework pro komplexní tvorbu bohatých formulářových aplikací, který je součástí .NET frameworku od verze 3.0. Disponuje širokou paletou formulářových prvků a také umožňuje bohaté stylování vzhledu aplikace. Framework nám nabízí spoustu hotových komponent, ze kterých formulář jednoduše poskládáme. Jedná se tedy o různá tlačítka, pole, posuvníky, popisky a další komponenty, které Microsoft nazval controls. Nic nám samozřejmě nebrání v tvorbě vlastních kontrolek, když by nám nějaká nestačila, ale to se nestává příliš často. Kromě WPF je v .NET frameworku stále přítomný starší formulářový framework Windows Forms. Ačkoli Microsoft Windows Forms ještě neoznačil jako zastaralý a v současné době se paralelně používají oba frameworky,

WPF je technologicky mnohem dále. I když mnoho existujících aplikací stále používá Windows Forms, nové aplikace již prakticky nemá smysl vyvíjet v ničem jiném, než právě ve WPF. (Microsoft, 2015)

8.3 Vývojový diagram

Zde jsem se pokusil o co nejjednodušší nastínění průběhu aplikace. Aplikace načítá data z formuláře, a pokud nejsou vyplněny všechny údaje (jméno může být prázdné) začne hledat k-nejbližších sousedů (k je implicitně nastavena na 5). Blízkost uživatelů je počítána na základě Euklidovské vzdálenosti. Po nalezení 5-ti sousedů program navrhne doplnění vypočtených pravděpodobných údajů a poté nabídne uložení záznamu do databáze.



Obrázek 9: Vývojový diagram

8.4 Implementace

Pro přehlednost jsou zde popsány pouze nejdůležitější části kódu. Celý kód je uložen v přílohách na CD. U nově vkládaného uživatele jsou všechny hodnoty nastaveny na -1 (u věku -1000 kvůli možné kolizi při vkládání velmi mladých uživatelů) a pokud je údaj zvolen je nastaven na 1. Dále představim několik klíčových metod pro funkci programu

8.4.1 Metoda convertAgeToPercent

Pro porovnání uživatelů podle věku bylo třeba transformovat věk z normálního formátu (25 let) na formát škály od 0 do 1, kde 0 reprezentuje minimální věk v databázi a 1 věk maximální. Z tohoto formátu nevíme kolik přesně uživateli je, ale víme, kde se nachází v porovnání s ostatními uživateli. Věk ve formátu škály je použit pro výpočty, poté je převeden a zaokrouhlen zpět na běžný formát. Pro ukládání do databáze je postup převrácen a ukládá se běžný, na celá čísla zaokrouhlený věk (25 let).

```
private double convertAgeToPercent(MySqlConnection conn, double age)
{
    int min_age = 0;
    int max_age = 1;

    string sql;
    MySqlCommand cmd;
    MySqlDataReader reader;

    sql = "SELECT min(age) as min_age, max(age) as max_age FROM users";
    cmd = new MySqlCommand(sql, conn);
    reader = cmd.ExecuteReader();

    while (reader.Read())
    {
        min_age = reader.GetInt32("min_age");
        max_age = reader.GetInt32("max_age");
    }
    reader.Close();

    age = (double)(age - min_age) / (max_age - min_age);
    age = Math.Round(age, 2);
    return age;
}
```

Obrázek 10: convertAgeToPercent

8.4.2 Metoda pro vypočtení Euklidovské vzdálenosti

Pro měření blízkosti dvou uživatelů jsem zvolil euklidovskou vzdálenost kvůli své jednoduchosti na implementaci a dobré vypovídací hodnotě o blízkosti dvou uživatelů. Bližší seznámení s Euklidovskou vzdáleností je v kapitole 5.4

```
    // euklidovska vzdalenost
    public double euclideanDistance(User u)
    {
        double sum = 0;

        if (this.age != -1000 && u.age != -1000)
            sum += Math.Pow((this.age - u.age), 2);
        if (this.scifi != -1 && u.scifi != -1)
            sum += Math.Pow((this.scifi - u.scifi), 2);
        if (this.comedy != -1 && u.comedy != -1)
            sum += Math.Pow((this.comedy - u.comedy), 2);
        if (this.action != -1 && u.action != -1)
            sum += Math.Pow((this.action - u.action), 2);
        if (this.war != -1 && u.war != -1)
            sum += Math.Pow((this.war - u.war), 2);
        if (this.drama != -1 && u.drama != -1)
            sum += Math.Pow((this.drama - u.drama), 2);
        if (this.horror != -1 && u.horror != -1)
            sum += Math.Pow((this.horror - u.horror), 2);
        if (this.historic != -1 && u.historic != -1)
            sum += Math.Pow((this.historic - u.historic), 2);
        if (this.crime != -1 && u.crime != -1)
            sum += Math.Pow((this.crime - u.crime), 2);

        return Math.Sqrt(sum);
    }
}
```

Obrázek 11: Metoda pro vypočtení Euklidovské vzdálenosti

8.4.3 Nalezení k-nejbližších sousedů

Nejvíce intuitivní a snadno pochopitelný je algoritmus K-NN. Pro uživatele u_0 vyberu K nejbližších sousedů $u_1, \ldots, u_k$, přičemž vzdálenost uživatelů se pro tento algoritmus určuje podle předešlé kapitoly. Celý algoritmus K-NN lze nelézt v kapitole 5.1

```
private List<User> findKNearestNeighbours(MySqlConnection conn, User u, int k)
{
    List<User> users = new List<User>();

    string sql = "SELECT * FROM users_transformed";
    MySqlCommand cmd = new MySqlCommand(sql, conn);
    MySqlDataReader reader = cmd.ExecuteReader();

    while (reader.Read())
    {
        User u1 = new User(reader);
        u1.distance = u.euclideanDistance(u1);
        users.Add(u1);
    }
    reader.Close();

    List<User> SortedList = users.OrderBy(o => o.distance).ToList();
    users.Clear();
    int i = 0;
    foreach (User uu in SortedList) {
        users.Add(uu);
        i++;
        if (i >= k)
            break;
    }
    return users;
}
```

Obrázek 12: Metoda pro nalezení k-nejbližších sousedů

8.4.4 Transformace dat

Pro účely výpočtu blízkých sousedů bylo nutno transformovat tabulku uživatelů do číselného formátu vyjádření preference (0 = nemá rád, 1 = má rád). U pohlaví 0 = žena, 1 = muž. Pro věk již dříve zmíněný formát škály.

id	name	gender	age	scifi	comedy	action	war	drama	horror	historic	crime	online	dvdbluray	cinema
1	Martin	1	0.13	1	1	1	0	0	0	0	0	1	0	0
2	Pepa	1	0.11	0	1	0	0	0	0	0	0	1	0	0
3	Zdenda	1	0.98	0	1	0	0	0	0	0	0	0	1	0
4	Zdena	0	0.93	0	1	0	0	0	0	0	0	0	1	0
5	Hynek	1	0.09	0	0	1	0	0	0	0	0	0	0	1
6	Míša	0	0.13	0	1	0	0	0	0	0	0	0	0	1
7	Kuba	1	0.11	0	0	1	0	0	0	0	0	1	0	0
8	uživatel 22	0	0.18	0	1	0	0	0	0	0	0	0	0	1

Obrázek 13: Transformovaná tabulka

Data se transformují z tabulek:

- User
- user_genres
- user_media

id	name	gender	age
1	Martin	M	24.00
2	Pepa	M	23.00
3	Zdenda	M	62.00
4	Zdena	F	60.00
5	Hynek	M	22.00
6	Míša	F	24.00

user_id	genre	id
1	scifi	1
1	action	129
1	comedy	128
2	comedy	2
3	comedy	3
4	comedy	4
5	action	5

user_id	medium	id
209	online	308
208	dvdbluray	307
207	online	306
206	online	305
206	dvdbluray	304
205	online	303
205	dvdbluray	302

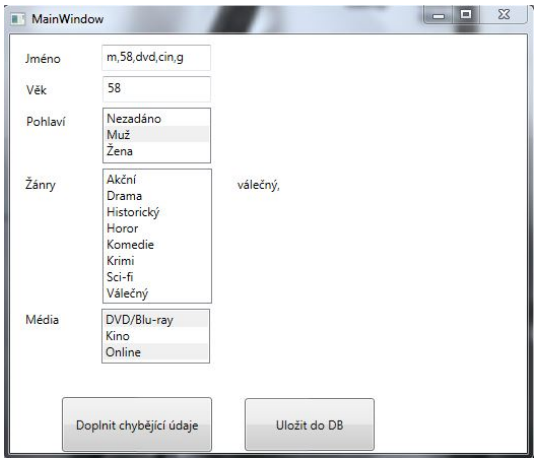
Obrázek 14: tabulkz

8.5 Algoritmus

1. Načti uživatelem zadaný informace
2. Transformuj data na potřebný formát
3. Zjisti který informace uživatel nezadal
4. Ze zadaných informací najdi vhodné sousedy
5. Zjisti nejčastější preferovanou hodnotu v sousedech
6. Doporuč hodnotu
7. Transformuj data
8. Ulož

8.6 Test

Pro testování aplikace byl zvolen muž, 58 let, který má v oblibě sledovat filmy online a z DVD/Blu-ray, ale nevyplnil svůj oblíbený žánr. Na obrázku 15 můžeme vidět doporučený žánr válečných filmů. V tabulce 3 pak můžeme vidět vypočtenou vzdálenost od našeho nekopletního uživatele.



Obrázek 15: Transformovaná tabulka

ID uživatele	205	142	148	171	202
Rozdíl	0	0,06	0,04	0,07	0,07

Tabulka 3: Vzdálenost sousedů

Na obrázku 16 vidíme doplněného uživatele (id 206) a jemu blízké sousedy pro kontrolu podobnosti údajů.

id	name	gender	age	scifi	comedy	action	war	drama	horror	historic	crime	online	dvdbluray	cinema
142		1	0.67	0	0	0	1	0	0	0	0	0	1	0
148		1	0.93	0	0	0	1	0	0	0	0	1	1	0
171		1	0.82	0	0	0	1	0	0	0	0	0	1	0
202	kkk	1	0.82	0	0	0	1	0	0	0	0	0	1	0
205		1	0.89	0	0	0	1	0	0	0	0	1	1	0
206	m,58,dvd,cin,g	1	0.89	0	0	0	1	0	0	0	0	1	1	0

Obrázek 16: Sousedi nekopletního uživatele

Test vykázal velmi malé rozdíly mezi uživateli (<0,1), kontrola uživatelů zobrazením v databázi ověřila podobnost a dá se tak prohlásit test aplikace za úspěšný.

9 Závěr

V průběhu vypracovávání této bakalářské práce jsem se seznámil a nastudoval vlastnosti, postupy a možnosti kolaborativního filtrování, které je nejvhodnější pro nasazení pro doporučovací systémy v rámci e-commerce. Cílem práce bylo seznámení se s existujícími doporučovacími systémy, analýza existujících doporučovacích systému, kriticky analyzovat a najít mezery ve funkci těchto webů a navrhnutí opatření jak tyto nedostatky zacelit a zefektivnit tak jejich fungování, druhým bodem praktické části je realizace programu pro doplnění chybějících dat v případě, že uživatel o sobě nezadal úplná data. Osobní přínos práce hodnotím velmi pozitivně, studium tématu mě zajímalo, bavilo a přiučil jsem mnoho zajímavých věcí o fungování velkých e-commerce portálů, zlepšil jsem své dovednosti v oblasti programování, které mi jak doufám pomohou ke kvalitnějšímu uplatnění v praxi po dostudování.

Abstrakt

Tématem této bakalářské práce jsou doporučovací systémy. Mým cílem je seznámit se s problematikou uživatelské preference a algoritmy kolaborativního filtrování, prozkoumat existující propracovanější doporučovací systémy, analyzovat je, najít prostor pro zlepšení a realizovat program řešící problematiku zadání nekopletních informací o uživateli.

Klíčová slova

Doporučovací systém, identifikace, uživatelská preference

Abstract

The subject of this bachelor thesis is recommendation systems. My aim is to learn about the problems of user preferences and Collaborative filtering algorithms, explore existing sophisticated recommendation systems, analyze them, find space for improvement and implement a program to solve the issue of entering incomplete user information.

Keywords

Recommendation system, identification, user preferences

Seznam obrázků

Seznam tabulek

Reference

Adomavicius, G. (2005). *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.* Retrieved from http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.107.2790&rep=rep1&type=pdf

Burke, R. (2002). *Hybrid recommender systems: Survey and experiments.* Retrieved from http://josquin.cs.depaul.edu/~rburke/pubs/burke-umuai02.pdf

Burke, R. (2005). Segment-based injection attacks against collaborative filtering recommender systems. In *In proceedings of the international conference on data mining (icdm 2005* (pp. 577–580).

Cvengroš, P. (2011). *Universal recommender system.* Retrieved from https://unresyst.googlecode.com/files/dp_final.pdf

Houdek, M., & Svoboda, T. (2001). *Pearson product-moment correlation coefficient.* Retrieved from http://www.answers.com/topic/pearson-s-correlation

Houdek, M., Svoboda, T., & Procházka, T. (2001). *Klasi?kace podle nejbližších sousedů.* Retrieved from http://cmp.felk.cvut.cz/cmp/courses/recognition/zapis_prednasky/zapis_01/4/rpz4.pdf

Žák, V. (2010). *Analýza chování uživatele na webových stránkách.* Retrieved from https://is.cuni.cz/webapps/zzp/detail/49314/

Kortus, L. (2013). *Doporučovací systém pro e-commerce.* Retrieved from http://theses.cz/id/xzoi8g/Kortus_Lukas_Bakalarska_prace.pdf

Magnani, M. (2004). *Techniques for dealing with missing data in knowledge discovery tasks.* Retrieved from http://magnanim.web.cs.unibo.it/data/pdf/missingdata.pdf

Melville P., S. V. (2010). *Recommender systems.* Retrieved from http://www.prem-melville.com/publications/recommender-systems-eml2010.pdf

Microsoft. (2015). *Visual c#.* Retrieved from https://msdn.microsoft.com/cs-cz/library/kx37x362.aspx

Peška, L. (2010). *Uživatelské preference v prostředí prodejních webů.* Retrieved from https://is.cuni.cz/studium/dipl_st/index.php?doo=detail&did=84549

Rozenberg, D. (2013). *Úprava nástroje dmvisual.* Retrieved from http://athena.zcu.cz/kurzy/spne/000/HTML/41/

Stružský, M. (2009). *Kolaborativní filtrování pro adaptivní web.* Retrieved from https://dip.felk.cvut.cz/browse/pdfcache/struzm1_2009bach.pdf

Vala, M. (2012). *E-learning – doporučovací systémy.* Retrieved from http://is.muni.cz/th/359917/fi_b/bp_final_vala.pdf

Vojtáš, P. (2010). *Modely uživatelských preferencí.* Retrieved from http://www.ksi.mff.cuni.cz/~vojtas/vyuka/NDBI021PrincipyUzivatelskychPreferenci/1112_NSWI021_DotazovaniSPreferencemi/DBI021modelyUzivatele.ppt

Wikipedie. (2015). *Facebook — wikipedie: Otevřená encyklopedie.* Retrieved from http://cs.wikipedia.org/wiki/Facebook (Online; navštíveno 23.3.2015)

WolframMarhWorld. (2013). *Euclidean distance.* Retrieved from http://mathworld.wolfram.com/Distance.html

10 Přílohy

Příloha č.1

Obsah přiloženého CD

- Test bakalářské práce ve formátu PDF
- Zdrojový kód aplikace v projektu Visual Studio 2015
- Exportovaná databáze ve formátu SQL
- Projekt LaTeX v programu TeXnicCenter