Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta
Katedra aplikované matematiky a informatiky

Teze bakalářské práce

Doporučovací algoritmy

Vypracoval: Veranika Kuliashova

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

České Budějovice

2023

Klíčová slova: doporučovací algoritmy, doporučení, e-komerce, Python
Anotace: Cílem práce je analýza běžné známých doporučovacích algoritmů a jejich využití v e-commerce. Tento výzkum poskytuje přehled o současném stavu oboru, pokud jde o využití doporučovacích systémů v e-shopech a na streamovacích platformách. V práce se použity široce používané doporučovací systémy, jako jsou Kolaborativní Filtrovaní, Content-Based Filtrovaní, and Hybridní filtrovaní. Zde jsou uvedeny teoretická východiska výše uvedených systémů, jakož i jejich problematika, jedinečné vlastnosti, reálné aplikace a jejich rozdíly. Výsledkem praktické části je kolaborativní doporučovací algoritmus napsaný v jazyce Python autorkou této práce se zaměřením na přesnost a efektivitu, s použitím kterého byli udělaný finální doporučeni filmů při uzívaní data setu MovieLens 10M.

Úvod a přehled literatury

Současný svět je charakterizován rychlými změnami a pokrokem, což je dáno technologiemi a globalizací, způsobuje to, že svět se pohybuje rychleji než kdykoli předtím. Počítače a další zařízení jsou nyní všudypřítomné a lidé je používají nejen pro práci, studium nebo zábavu, ale také pro komunikaci, nákupy, sledování filmů a seriálů, poslech hudby a čtení knih. Díky internetu má téměř každý neomezený přístup k informacím, zboží a službám, které v minulosti nebyly k dispozici, což vyžaduje, aby lidé byli proaktivní a přizpůsobiví, aby drželi krok s měnící se dobou.

Doporučovací algoritmy jsou v současnosti velmi důležité a jsou využívány na mnoha webových stránkách, včetně elektronických obchodů, sociálních médií, zpravodajských webových stránek a služeb pro streamování, kde hrají klíčovou roli při zapojení a udržení uživatelů. Tyto algoritmy se stávají stále sofistikovanějšími s využitím strojového učení a umělé inteligence, ale stále se setkávají s mnoha problémy, jako je řídkost dat, škálovatelnost a obavy o soukromí.

Toto téma je relevantní pro mnoho oblastí a je důležité pro vývoj efektivních doporučovacích algoritmů a systémů. Z tohoto důvodu jsem si vybrala toto téma pro svou bakalářskou práci, která se skládá ze 5 hlavních kapitol. První kapitola se zabývá úvodem a významem práce a jejími cíli. Druhá kapitola se zaměřuje na problematiku doporučovacích algoritmů a systémů, včetně jejich výhod a použití v praxi. Třetí kapitola popisuje metodiku praktické části, zatímco kapitola čtyři se zaměřuje na vytvoření kolaborativního algoritmu pro filmovou databázi MovieLens. Poslední kapitola obsahuje závěry a důsledky práce a možná zlepšení pro další výzkum.

Při psaní této bakalářské práce byla použita především literatura jako knihy a články z vědeckých časopisů a internetových zdrojů. Použity byly také některé zdroje, které se zaměřovaly nejen na technické vlastnosti a popisy algoritmů, ale také na jejich vliv na svět kolem nás a způsoby, kterými nám usnadňují nebo ztěžují život. Chtěla bych zmínit několik knih, které mi nejvíce pomohly pochopit problematiku doporučovacích algoritmů a jejich typů, jak fungují a co dobrého přinášejí světu. Například jsem našla knihu "Recommender Systems: the textbook" od Charu C. Aggarwal velmi užitečnou. Poskytuje ucelený přehled o oblasti doporučovacích systémů pro knihy. Kniha se zabývá tradičními i moderními technikami pro vytváření doporučovacích systémů, včetně kolaborativního filtrování, filtrování na základě obsahu, maticové faktorizace, hlubokého učení a dalších, probírá důležitá témata, jako jsou metriky hodnocení.

Cíl práce

Cílem práce je získat a shrnout poznatky o doporučovacích algoritmech z odborné literatury, analyzovat co je doporučovací algoritmus a každý typ zvlášť, a jejich současné využiti na internetu v e-commerce, včetně e-shopů a streamovacích služeb. A na základě teoretických informací vytvořit vlastní kolaborativní doporučovací algoritmus pro filmovou databázi MovieLens s záměrem na přesnost.

Metodika

Jako kompilátor kódu jsem použil Google Colab. To je cloudové vývojové prostředí (IDE), které umožňuje vývojářům a vědcům pracovat s daty a programovat v různých jazycích, včetně Pythonu. Jedná se o zdarma dostupnou službu, která využívá výpočetní sílu a infrastrukturu Google Cloud. Je ideální pro vývoj aplikací s použitím strojového učení,

hlubokého učení a umělé inteligence, protože umožňuje používat GPU a TPU pro výpočty. To znamená, že lze provádět náročné výpočty mnohem rychleji, než by bylo možné na běžném počítači. Prostředí Google Colab umožňuje uživatelům pracovat s daty uloženými v Google Drive, importovat knihovny a sdílet svůj kód a projekty s ostatními. Kromě toho, Google Colab je také ideální pro vzdělávání a školení v oblasti programování a datové analýzy, protože poskytuje prostředí pro jednoduché vytváření a sdílení kódu.

Použila jsem tento kompilátor kódu kvůli snadnému procesu instalace knihoven pro její používání si získal místo v mém výzkumu. Navíc, z důvodu že můj počítač není v současné době nejmodernější ani nejvýkonnější, takže možnost používat cloudovou platformu pro mě byla ideální možnost.

K vytvoření doporučovacích algoritmů byl použit programovací jazyk Python a knihovna Surprise.

Python je vysokoúrovňový interpretovaný programovací jazyk, který se hojně používá pro programování pro obecné účely (Pilgrim, 2014). Je navržen tak, aby se snadno četl a psal, což z něj činí oblíbenou volbu pro začátečníky i zkušené programátory. Jednoduchost jazyka Python je dána jeho minimalistickou syntaxí, která umožňuje vývojářům psát kód rychle a efektivně.

Jednou z klíčových výhod jazyka Python je jeho univerzálnost. Lze jej použít pro širokou škálu aplikací, včetně vývoje webových stránek, analýzy dat, umělé inteligence, strojového učení, vědeckých výpočtů, vývoje her a dalších. Tato všestrannost je dána rozsáhlými knihovnami a rámci jazyka Python, které vývojářům poskytují hotová řešení pro běžné programátorské úlohy. Vybrala jsem si ho hlavně z těchto důvodů.

Surprise je scikitová knihovna pro Python určená k vytváření a analýze doporučovacích systémů. Poskytuje jednoduché a snadno použitelné rozhraní pro vytváření doporučovacích algoritmů s kolaborativním filtrováním. Kolaborativní filtrování je metoda vytváření doporučení předpovídáním hodnocení nebo preferencí, které by uživatel dal položce na základě preferencí jiných uživatelů s podobným vkusem.

Surprise poskytuje řadu vestavěných algoritmů, například SVD (Singular Value Decomposition), NMF (Non-Negative Matrix Factorization) a podobně, a také užitečné funkce pro načítání, rozdělování a vyhodnocování datových sad (Hug, 2020). Podporuje také vlastní algoritmy a datové formáty.

Křížová validace, či cross validation, je technika používaná k hodnocení výkonnosti doporučovacích algoritmů, včetně SVD (Singulární rozklad), NMF (Faktorizace nezáporných matic) a PMF (Pravděpodobnostní faktorizace matic). Běžně se používá také ve strojovém učení k testování zobecňováních schopností modelu na neznámých datech.

V kontextu doporučovacích algoritmů zahrnuje křížová validace rozdělení dat na více složek nebo podmnožin. Jedna podmnožina se použije k testování a ostatní podmnožiny se použijí k trénování modelu. Tento proces se několikrát opakuje, přičemž každá podmnožina se použije k testování jednou, a výsledky se zprůměrují, aby se získala celková míra výkonnosti algoritmu. umožňuje testovat algoritmus na různých částech dat, což poskytuje robustnější posouzení jeho výkonnosti.

Hodnoty, podle kterých bude vyhodnocena přesnost algoritmů a vybrán nejlepší z nich – MAE a RMSE.

Průměrná absolutní hodnota nebo Mean Average Error (MAE) je jednou z mnoha metrik pro shrnutí a hodnocení kvality modelu strojového učení. Chyba se zde vztahuje k odečtení předpovídané hodnoty od skutečné hodnoty, jak je vidět na vzorci níže.

$$Chyba\ Předpovědi = Skutečná\ Hodnota - Předpovězená\ Hodnota$$

Chyba predikce se bere pro každý záznam a poté se všechny chyby převedou na kladné. Toho dosáhneme tak, že pro každou chybu vezmeme absolutní hodnotu, jak je uvedeno níže:

Absolutní chyba → | *Chyba Předpovědi* |

Nakonec vypočítáme průměr pro všechny zaznamenané absolutní chyby (průměrný součet všech absolutních chyb).

$$MAE = \frac{\sum_{i=1}^{n} |\mathrm{yi} - \mathrm{xi}|}{n}$$

Zde jsou:

Zde yi jsou předpokládaná hodnota.

xi jsou pozorovaná hodnota.

n je počet pozorování.

RMSE (Root-Mean-Square Deviation či směrodatná odchylka chyb) je standardní způsob měření chyby modelu při předpovídání kvantitativních dat. Formálně je definována jako následující vzorec:

$$MSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{\mathrm{y}}\mathrm{i} - \mathrm{yi})^2}{n}}$$

Zde jsou:

Zde $\hat{y}1$, $\hat{y}2$, ..., $\hat{y}n$ jsou předpokládané hodnoty.

$y1$, $y2$, ..., yn jsou pozorované hodnoty.

n je počet pozorování.

RMSE měří průměrný rozdíl mezi předpovídanými hodnoceními a skutečnými hodnoceními, která uživatelé udělili souboru položek. Čím nižší je hodnota RMSE, tím algoritmus lépe předpovídá hodnocení položek uživateli, a proto je přesnější.

Důležité je také zmínit dataset, která bude analyzována v praktické části práce. Datová sada MovieLens je oblíbenou srovnávací datovou sadou pro výzkum doporučovacích systémů. Jedná se o sbírku hodnocení filmů a metadat o filmech, která byla v průběhu let shromážděna z webových stránek MovieLens. Tato datová sada byla široce využívána při výzkumu doporučovacích algoritmů a sloužila jako měřítko pro hodnocení výkonnosti doporučovacích modelů. Datová sada MovieLens byla poprvé zveřejněna v roce 1997 výzkumnou skupinou GroupLens Research z Minnesotské univerzity (Harper & Konstan, 2015). Datová sada byla v průběhu let několikrát aktualizována, přičemž poslední verzí je MovieLens 25M, vydaná v roce 2019. Data set obsahuje různé velikosti, včetně původního data setu se 100 000 hodnoceními a dalších s 1M, 10M, 20M a 25M hodnoceními.

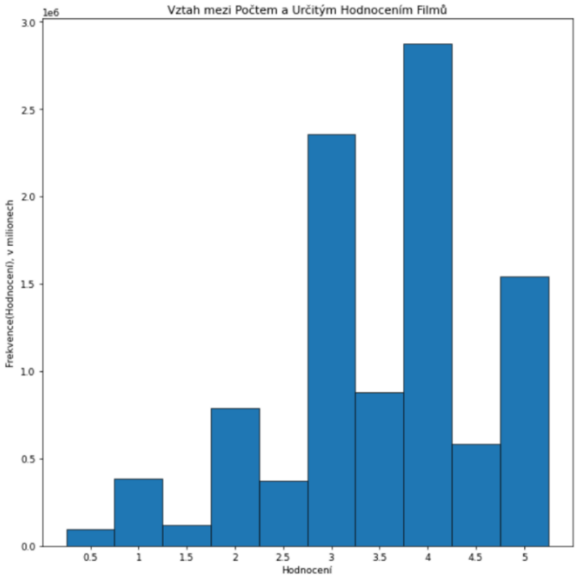
Soubor dat MovieLens 10M obsahuje přibližně 10 milionů hodnocení filmů, která poskytlo zhruba 70 000 uživatelů k víc než 10 000 filmům. Hodnocení jsou na stupnici od 1 do 5, přičemž 5 je nejvyšší hodnocení, s půlhvězdičkovými přírůstky (0,5 hvězdičky - 5,0 hvězdičky). Soubor dat obsahuje další informace o filmech, například název filmu, rok vydání a žánry. Data byla shromážděna v období od ledna 1995 do září 2015, a v databázi jsou použity pouze údaje o uživatelích a filmech, které mají 20 nebo více hodnocení.

Výsledky

Před zahájením práce na algoritmu byla provedena analýza datové sady MovieLens. Autor práce po zformátování dat do požadovaného formátu pomocí knihovny Pandas, kterou lze

analyzovat pomocí knihovny Surprise comment, vytvořil grafy v programovacím jazyce Python díky knihovně Matplotlib.

Graf 1. Vztah mezi počtem filmů a hodnocením



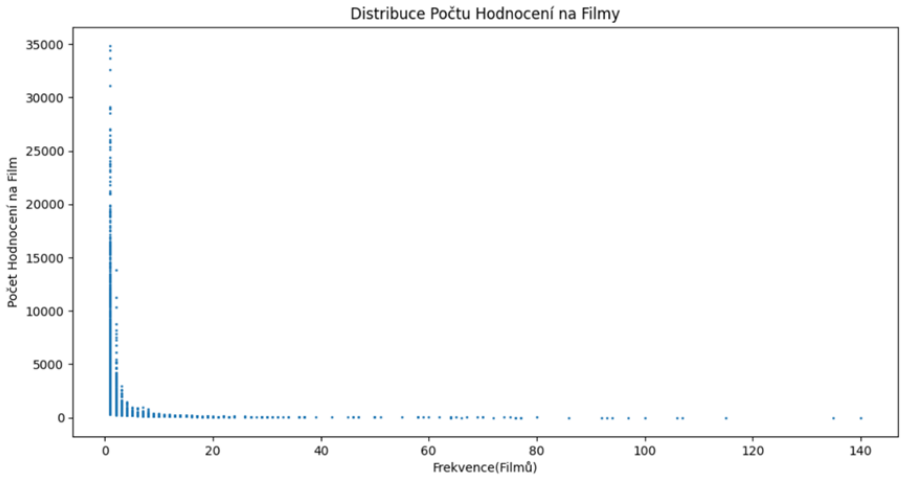
Zdroj: vlastní tvorba

Graf 1 ukazuje vztah mezi počtem a hodnocením filmů. Osa x představuje vše existujici možnosti hodnocení – od 0,5 do 5 hvězdiček, zatímco osa y představuje počet určitym hodnocení filmů. Z grafu vyplývá, že většina filmů má hodnocení 4 a 3 hvězdičky, menší počet filmů má hodnocení 5 nebo 3,5 hvězdiček. Kromě toho je méně hodnocení 0,5 až 1,5 mají nejmenší počet hodnocení.

Díky tomu, že je v databázi tolik různých recenzí, s větším počtem pozitivních, si můžeme být jisti, že každá z těchto hodnocení bude hrát důležitou roli při jejich doporučování filmů uživatelům. Lze říci, že většina filmů v tomto souboru dat má hodnocení 4 hvězdičky, což znamená, že oni byly diváky obecně dobře přijaty.

Dalším grafem 2 bych chtěla ukázat distribuce počtu hodnocení na filmy. Skvěle nám ukazuje, že ve skutečnosti často existuje problém dlouhého chvostu, protože vidíme, že maximální počet recenzí má malý počet filmů. Dále se nejčastěji vyskytují počty recenzí na jeden film v hodnotách mezi 50 a přibližně 4000.

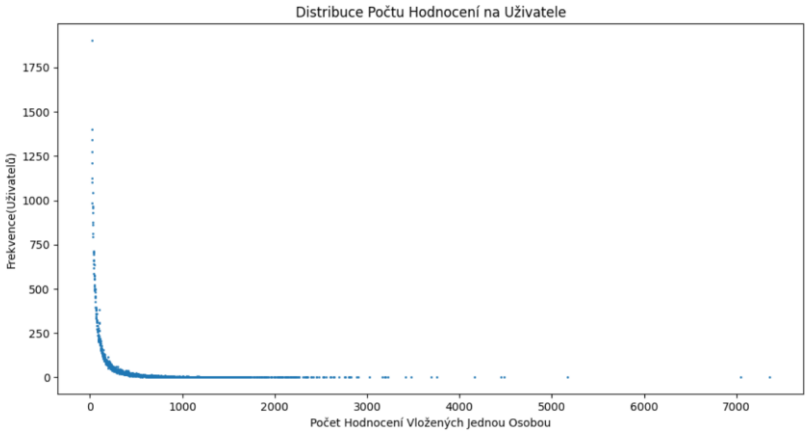
Graf 2. Distribuce Počtu Hodnocení na Filmy



Zdroj: vlastní tvorba

Pomocí grafu 3 bych chtěla ukázat distribuce počtu hodnocení na uživatele. Díky 3. grafu můžeme opět vidět graf dlouhého chvostu, což znamená, že uživatelé v průběhu let udělili velké množství hodnocení populárním filmům, pak průměrné populární filmy máji většinu hodnocení (od cca 100 do cca 500) a většina filmů má opět nízké hodnocení.

Graf 3. Distribuce Počtu Hodnocení na Uživatele



Zdroj: vlastní tvorba

Nyní můžeme přistoupit ke zvážení výsledků algoritmů SVD, NMF a PMF.

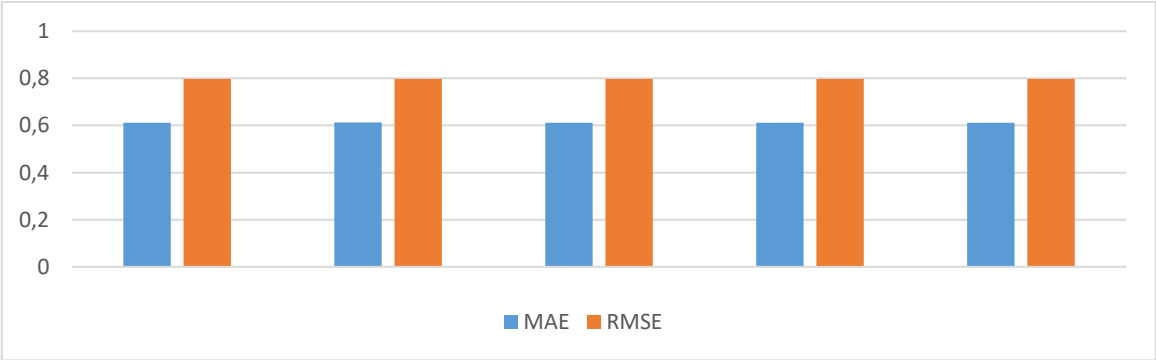
SVD je zkratka pro singulární rozklad nebo Singular Value Decomposition v angličtině. Jedná se o techniku faktorizace matice, která rozkládá matici na další matice. V kontextu doporučovacích algoritmů se SVD často používá k faktorizaci matice interakce mezi uživatelem a položkou na dvě méně rozměrné matice: jednu matici obsahující latentní faktory, které představují preference uživatele, a druhou matici obsahující latentní faktory, které představují atributy položky. v tabulce 1 vidíme výsledky algoritmu SVD a pomocí hodnot MAE a RMSE vytvořit graf 4.

Tabulka 1. Hodnoty MAE a RMSE SVD algoritmu

	Složení 1	Složení 2	Složení 3	Složení 4	Složení 5	Průměr	STD
MAE	0.6120	0.6120	0.6121	0.6115	0.6120	0.6119	0.0002
RMSE	0.7983	0.7981	0.7980	0.7974	0.7982	0.7980	0.0003

Zdroj: vlastní vypočet

Graf 4. RMSE a MAE doporučovacího algoritmu, založeného na SVD



Zdroj: vlastní tvorba

Jak vidíme, hodnoty MAE a RMSE je malé, což znamená, že pravděpodobnost chyby doporučení v tomto algoritmu je mala. Tyto údaje jsme získali díky velikosti použité databáze. Z důvodu, že obsahuje 10 milionů hodnocení přesnosti, je přesnější.

NMF (Non-negative Matrix Factorization) je technika faktorizace matic, kterou lze použít k rozkladu nezáporné matice na dvě nezáporné matice. Cílem NMF je najít dvě nezáporné matice s nízkým řádem, které se po vynásobení přibližují původní nezáporné matici.

V kontextu doporučovacích algoritmů lze NMF použít k faktorizaci matice hodnocení uživatelů a položek na dvě nezáporné matice s nižší hodnotou, z nichž jedna představuje vložené hodnoty uživatelů a druhá vložené hodnoty položek.

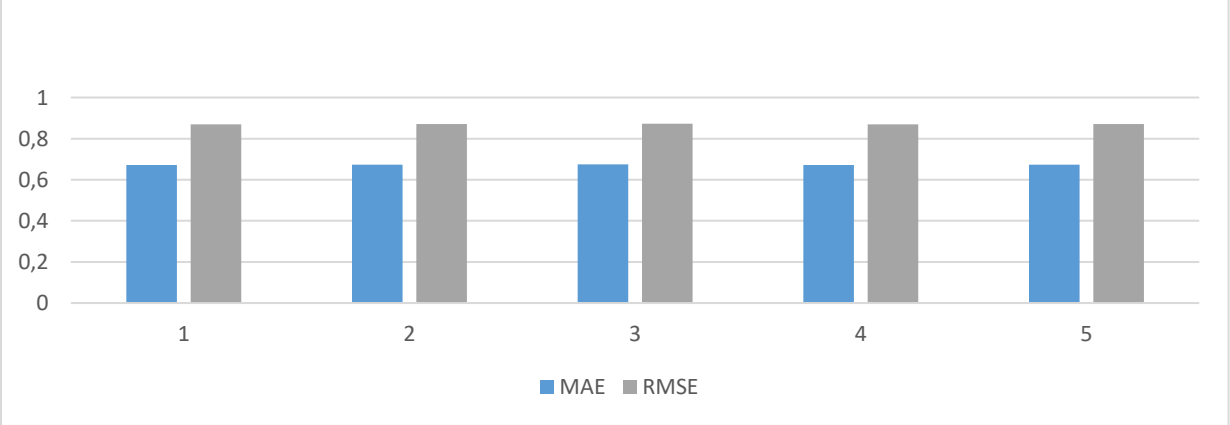
Tabulka 2. Hodnoty MAE a RMSE NMF algoritmu

	Složení 1	Složení 2	Složení 3	Složení 4	Složení 5	Průměr	STD
MAE	0.6717	0.6736	0.6744	0.6717	0.6726	0.6728	0.0011
RMSE	0.8693	0.8716	0.8724	0.8690	0.8706	0.8706	0.0013

Zdroj: vlastní vypočet

Jak je vidět z výsledků algoritmu zaznamenaných v tabulce 2, algoritmus NMF je také dost přesný, protože jeho průměrná hodnota RMSE je 0,8706. Není to perfektní hodnota, ale je stále dobrá. Obecně platí, že RMSE 0,87 není špatná, ale v závislosti na kontextu může, ale nemusí být dobrá. Pracuji jsem na doporučovacím algoritmu, který má velký počet položek a uživatelů a hodnocení nejsou velmi řídká, může být RMSE 0,87 považována za normální.

Graf 5. RMSE a MAE doporučovacího algoritmu, založeného na NMF



Zdroj: vlastní tvorba

Pravděpodobnostní maticová faktorizace (PMF) je doporučovací algoritmus, který se používá k předvídání, jak by uživatel hodnotil položku, kterou ještě nehodnotil. PMF

předpokládá, že v pozadí matice hodnocení uživatele a položky jsou určité latentní faktory nebo vlastnosti, které nejsou přímo pozorovatelné. Tyto latentní faktory se pak používají k rozkladu matice hodnocení na matice nižších rozměrů, které reprezentují vlastnosti uživatele a položky v latentním prostoru.

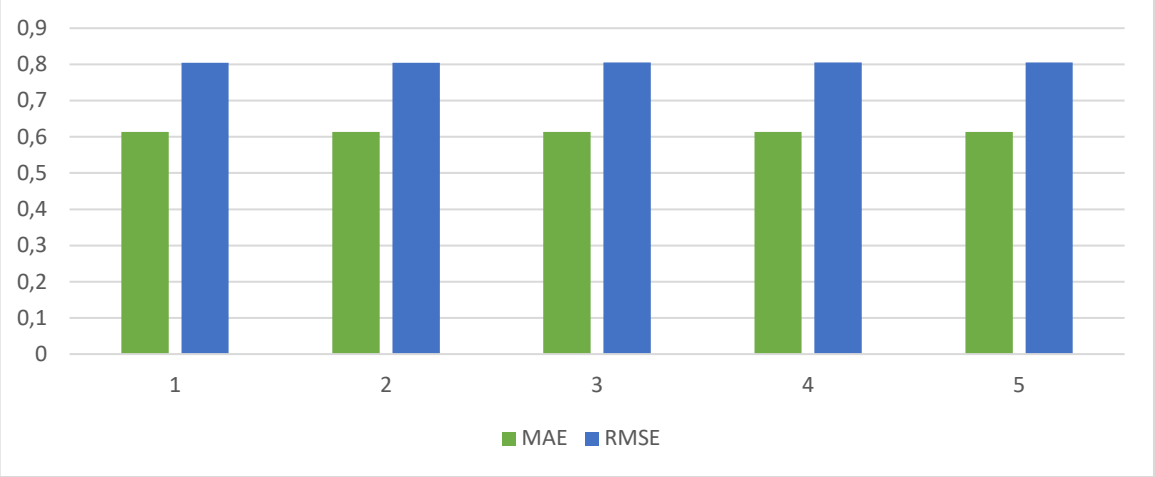
Tabulka 3. Hodnoty MAE a RMSE PMF algoritmu

	Složení 1	Složení 2	Složení 3	Složení 4	Složení 5	Průměr	STD
MAE	0.6132	0.6131	0.6138	0.6133	0.6133	0.6134	0.0002
RMSE	0.8048	0.8044	0.8051	0.8052	0.8051	0.8049	0.0003

Zdroj: vlastní vypočet

Průměrné hodnoty MAE 0,6134 a RMSE 0,8049 ukazují přesnost algoritmu PMF při předpovídání hodnocení filmů. Nižší hodnoty těchto ukazatelů naznačují lepší přesnost, takže tyto hodnoty naznačují, že algoritmus PMF je ve svých předpovědích středně přesný.

Graf 6. RMSE a MAE doporučovacího algoritmu, založeného na PMF



Zdroj: vlastní tvorba

Díky tomu, že vše algoritmy byly úspěšně spuštěny a hodnoty chyb byly získaný, můžeme je porovnat na základě jejich hodnot MAE a RMSE, abychom našli ten nejpřesnější algoritmus, který použiji pro doporučování filmů v další kapitole.

Pokud chceme porovnat výkonnost a přesnost SVD, NMF a PMF v doporučovacím algoritmů, aby najit nejpřesnější algoritmus, můžeme k vyhodnocení přesnosti předpovídaných hodnocení použít metriku RMSE. Jedná se o běžně používanou metriku hodnocení doporučovacích algoritmů. RMSE je užitečná metrika, protože bere v úvahu velikost chyb a větší chyby jsou více penalizovány. Kromě toho se poměrně snadno interpretuje a její hodnoty lze porovnávat mezi různými modely a soubory dat.

Porovnáním hodnot RMSE různých modelů lze posoudit, které z nich jsou pro daný problém a datovou sadu výkonnější. To vám může pomoci vybrat nejlepší model pro váš doporučovací systém a zlepšit jeho přesnost.

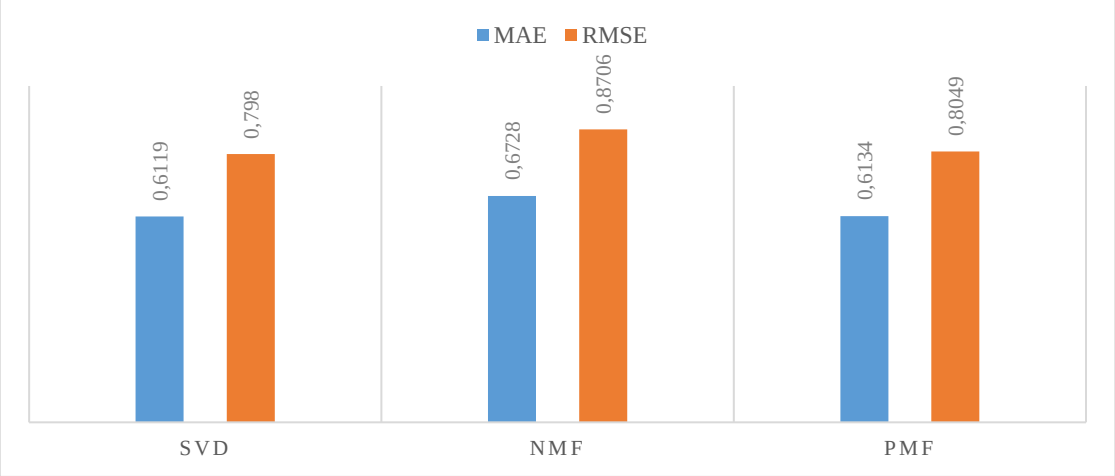
Pro porovnání algoritmů musím nejdříve vytvořit tabulku, která bude obsahovat průměrné hodnoty MAE a RMSE u obou použitých algoritmů. Tyto údaje můžete vidět v tabulce 5, pomocí které je vytvořen graf 7.

Tabulka 4. Průměrné hodnoty MAE a RMSE algoritmů SVD, NMF a PMF

	SVD	NMF	PMF
MAE	0,6119	0,6728	0,6134
RMSE	0,7980	0,8706	0,8049

Zdroj: vlastní vypočet

Graf 7. Porovnání průměrných hodnot MAE a RMSE algoritmů SVD, NMF a PMF



Zdroj: vlastní tvorba

Po vyhodnocení výkonnosti dvou algoritmů faktorizace matic, SVD, NMF a PMF, jako doporučujícího algoritmu a jejich srovnávaní pomoci tabule a grafu jsem zjistila, že SVD a PMF je přesnější než NMF. Je to z důvodu, že průměrné hodnoty RMSE pro SVD a PMF byly 0,7980, resp. 0,8049, když NMF má hodnotu 0,8706. Je známo, že čím menší je hodnota RMSE, tím je algoritmus přesnější. Také můžeme vidět, že hodnota MAE algoritmu SVD je nižší, pokud se podíváme na graf 8. Ona také znamená, že čím nižší bude, tím že pravděpodobnost chyby je menší. ale hodnoty RMSE jsou obvykle přesnější než, a z toho důvodu to oni jsou srovnáváni.

Metrika RMSE měří průměrnou vzdálenost mezi předpovídanými hodnoceními a skutečnými hodnoceními v testovací množině. Nižší hodnota RMSE znamená lepší výkonnost modelu, protože odráží, jak dobře dokáže algoritmus předpovídat hodnocení uživatelů. Hodnoty RMSE 0,7980, 0,8049 a 0,8706 tedy naznačují, že SVD v tomto konkrétním problému a datové sadě funguje lépe než NMF a PMF.

Konečně můžeme použít nejpřesnější algoritmus SVD pro funkci, která na základě ID daného uživatele vytváří doporučení filmů.

Obrázek 1. Výsledek kolaborativního algoritmu

Doporučení na základě žánrů pro uživatele : 20

	movie_id	estimated_rating	title	actual_rating	genres
1293	318	4.419858	[Shawshank Redemption, The (1994)]	[]	[Drama]
133	527	4.355029	[Schindler's List (1993)]	[]	[Drama|War]
213	50	4.342639	[Usual Suspects, The (1995)]	[]	[Crime|Mystery|Thriller]
8631	8684	4.332486	[Man Escaped, A (Un condamnÃ© Ã mort s'est Ã©...	[]	[Adventure|Drama]
4689	44555	4.323449	[Lives of Others, The (Das Leben der Anderen) ...	[]	[Drama]
5632	6896	4.320666	[Shoah (1985)]	[]	[Documentary|War]
2056	2324	4.310301	[Life Is Beautiful (La Vita Ã¨ bella) (1997)]	[]	[Comedy|Drama|Romance|War]
1926	4973	4.305226	[Amelie (Fabuleux destin d'AmÃ©lie Poulain, Le...	[]	[Comedy|Romance]
34	858	4.301807	[Godfather, The (1972)]	[]	[Crime|Drama]
8142	8516	4.293884	[Matter of Life and Death, A (Stairway to Heav...	[]	[Comedy|Fantasy|Romance]

Zdroj: vlastní tvorba

Na obrázku 1 vidíme 10 doporučení filmů pro konkrétního uživatele, v našem případě s id 20. Každý z těchto filmů má odhadované hodnocení pro uživatele. V tomto seznamu doporučení to není možné vidět, ale teoreticky se může stát, že jeden z doporučených filmů je již hodnocen. několikrát se to stalo během testování kódu, že se někdy v nejlepších doporučeních objevili filmy, které již mají existující hodnoceni od uživatele. Z mého hlediska není to velký problém, protože takto fungují například algoritmy Netflixu. Na některé uživatele možná zafunguje připomenutí dříve zhlédnutého oblíbeného filmu a rozhodnou se jej zhlédnout znovu – v tomto případě to není nevýhoda algoritmu.

Závěr

V závěru lze říci, že byly úspěšně splněny všechny cíle této bakalářské práce. Získané výsledky splnily očekávání stanovená na začátku studie. Autorkou práce byla provedena analýza doporučovacích algoritmů a jejich typů s použitím dat získaných z odborné literatury.

Jedním z cílů této studie bylo poskytnout přehled o použití doporučovacích algoritmů v elektronickém obchodě. Hloubková analýza literatury a webových stránek služeb prokázala, že tyto algoritmy se staly nezbytnou součástí platforem elektronického obchodování a poskytují uživatelům personalizované a relevantní návrhy produktů. Výsledky společností Amazon, Netflix a Spotify, tří nejúspěšnějších společností, které využívají doporučovací algoritmy, poskytují důkazy o jejich účinnosti při zvyšování uživatelského komfortu a podpoře prodeje. Poznatky z této části studie mohou být užitečné pro podniky, které chtějí zlepšit své strategie elektronického obchodování a poskytnout personalizovanou a uspokojivou uživatelskou zkušenost.

Cílem praktické části bylo vytvořit doporučovací algoritmus, který bude mít co nejvyšší přesnost parametru a doporučení. K tomuto účelu jsem vytvořila a analyzovala kolaborativní algoritmy pomocí knihovny Surprise v jazyce Python, které využívají techniky křížového ověřování. Těmito algoritmy jsou SVD – Singulární rozklad, NMF – Faktorizace nezáporných matic a PMF – Faktorizace pravděpodobnostních matic. Všechny tyto algoritmy jsou jedinečné, mají své silné a slabé stránky a určité zvláštnosti. K určení přesnosti algoritmů byly použity hodnoty MAE – střední absolutní chyba a RMSE – kořenová střední kvadratická odchylka. Čím nižší je tato hodnota, tím je algoritmus přesnější. Po porovnání všech průměrných hodnot RMSE a MAE u těchto 3 algoritmů bylo zřejmé, že nejpřesnějším algoritmem je pro nás algoritmus SVD. Ten pomáhá doporučení filmů bylo provedeno pro konkrétního uživatele.

Potenciál pro budoucí výzkum existuje v každé práci nebo studijním oboru vzhledem k neustálému vývoji znalostí a chápání, stejně jako ke vzniku nových otázek a oblastí zkoumání.

Například, v oblasti doporučovacích algoritmů v kontextu umělé inteligence má obrovský potenciál pro další výzkum vzhledem k tomu, roste poptávka po sofistikovanějších a přesnějších doporučovacích algoritmech, které dokáží zpracovávat větší objemy dat, lépe předpovídat a poskytovat personalizovanější uživatelské prostředí. Je také možnost zkoumat nové metody sběru dat, jako jsou aktivní metody učení, které umožňují algoritmu učit se od uživatelů během interakce s nimi.

Bez ohledu na přesnost algoritmu je v každé úloze prostor pro další vylepšení. V mém případě lze v budoucí práci dosáhnout zlepšení zvýšením přesnosti výpočtů, například použitím novějších a rozsáhlejších verzí databáze, konkrétně současných MovieLens 20M a MovieLens 25M. Kromě toho by s výše uvedenými databázemi mohlo lze bychom bylo bez problémů pracovat s použitím modernějšího a výkonnějšího počítače, který by data zpracovával mnohonásobně rychleji. V budoucnu lze tento algoritmus také kombinovat s existujícími algoritmy s otevřeným zdrojovým kódem či s použitím open source zdrojů, nebo vyvíjet k tomu nové algoritmy a rozšířit takém způsobem doporučení.

Seznam literatury použité v tezích

Plný seznam použité literatury je uveden v textu bakalářské práce.

Aggarwal, C. C. (2016*). Recommender Systems – The Textbook*. Springer. ISBN: 978-3-319-29659-3.

Hug, N. (2020). *Surprise: A Python library for recommender systems.* Columbia University. DOI: 10.21105/joss.02174.

Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* (TiiS), 5(4), 19. DOI: 10.1145/2827872.

Pilgrim, M. (2014). *Dive into Python*. Apress.

Další bibliografické údaje

Type of thesis: bachelor

Author: Veranika KULIASHOVA

Academic Year: 2022/2023

Department: Department of Applied Mathematics and Computer Science, Faculty of Economics, University of South Bohemia in Ceske Budejovice, Czech Republic

Thesis supervisor: doc. Ing. Ladislav Beránek, CSc., MBA

Number of pages: 72

Language od thesis: CZ

Title of Thesis: Recommendation Algorithms

Annotation: The purpose of the bachelor thesis is to analyse common known recommendation algorithms and their applications in E-Commerce. This research provides an overview of the current state of the field as for the usage of recommendation systems by electronic commerce, in other words, online shopping facilities and streaming services. The study investigates three widely used recommendation systems such as Collaborative Filtering, Content-Based Filtering, and Hybrid recommendation system. The theoretical basis of the aforementioned systems, as well as their problematics, unique features, real-life applications and their differences are presented. The result of the empirical part is a conclusion of utilising the chosen database using a recommendation algorithm written in Python by the author of this thesis with a focus on accuracy and efficiency.

Key words: recommendation algorithm, recommendations, e-commerce, Python.
