Зборник радова Факултета техничких наука, НовиСад

УДК: 4.9
ДОИ: https://doi.org/10.24867/34BE35Djanic

Примена факторизације матрица у системима препорука

Application of Matrix Factorization in Recommender Systems

Леополдина Ђанић, *Факултет техничких наука, Нови Сад*

Студијски програм – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *У овом раду се разматра и изучава примена медода факторизације матрица у системима препорука. Проучавају се различите методе факторизације матрица. Циљ је да се побољшају тачности и персонализација препорука. Проблем који се решава се односи на што прецизније предвиђање оцена које би корисник дао филмовима. Технике коришћене у експерименту су СВД, СВД++ и ТимеСВД++.*

Кључне речи: *системи препорука, факторизација матрица, СВД, СВД++, ТимеСВД++*

Abstract – *This paper examines and explores the application of matrix factorization methods in recommender systems. Various matrix factorization techniques are studied with the goal of improving the accuracy and personalization of recommendations. The main problem addressed is achieving more precise prediction of the ratings that users would assign to movies. The techniques used in the experiment are SVD, SVD++ and TimeSVD++.*

Keywords: *recommendation system, matrix factorization, SVD, SVD++, TimeSVD++*

НАПОМЕНА: Овај рад проистекао је из мастер рада чији ментор је био др Драган Иветић, ред. проф.

1. УВОД

Развојем интернета у последњих пар деценија повећала се и количина информација која се налази на интернету. Корисник сам не може да се снађе у овако великом обиму информација и да пронађе оно што га занима. Ова велика појава информација се назива још и преоптерећење информацијама (*енг. Information overload*). Системи препорука уведени су да олакшавају корисницима претрагу чинећи је бржом и омогућавајући корисницима да пронађу што више релевантних података. Најкоришћенији тип система препорука који се користе су системи препорука засновани на сарадњи [1]. Поред предности које су увели системи препорука, јављају се и проблеми у њиховом коришћењу као што су проблем реткоће информација (*енг. Sparsity*) и проблем хладног старта (*енг. Cold start problem*). У циљу решавања ових проблема уведени су системи препорука базирани на моделу а најпопуларнији овакав систем је систем препорука који користи факторизацију матрице. Факторизација матрице је добила знатно на популарности након такмичења које је организовала компанија Нетфликс 2006. године [2]. Циљ овог рада је да се истраже различити модели факторизације матрице и да се испитају њихове предности и мане.

2. ФАКТОРИЗАЦИЈА МАТРИЦЕ

Факторизација матрице је модел латентних фактора [2]. У моделима факторизације матрице је циљ да се и корисницима и филмовима придруже латентни фактори [3]. Латентни фактори се добијају учењем и тренирањем модела и служе да прикажу колико је корисник или филм повезан са појединим фактором. Код филмова латентни фактори могу бити жанрови иако их рачунар и модел неће видети у семантичком смислу као конкретан жанр, повезаће их са корисницима на исправан начин. Факторизација матрица је постала веома доминантна техника у примени у системима препорука [2]. Разлог увођења факторизације матрица је да се реше основни проблеми система препорука као што су реткоћа информација. Реткоћа информација је појава која настаје због недостатака информација о кориснику, корисници не оцењују све филмове које погледају, корисници су одгледали само јако мали део филмова од укупног броја филмова који постоје. То доводи до тога да матрица корисник-филм има веома велики број празних ћелија, што доводи до беспотребног заузећа меморије.

Основна идеја факторизације матрица је да почетну матрицу корисник-филм раздвоји на две нове матрице нижег ранга: корисник-фактор и филм-фактор [4]. У изразу (1) је приказана ова формула при чему је $\hat{A}$ матрица приближна почетној матрици, U^{T} је транспонована матрица корисник-фактор, а V је матрица филм-фактор.

$$\hat{A} \approx U^{T}V \tag{1}$$

Рачунање предвиђене оцене по овом основном моделу би се рачунало као што је приказано у изразу (2), где је u корисник који оцењује, i је филм који се оцењује, $\hat{r}_{ui}$ је предвиђена оцена, p_u је ред из матрице корисник-фактор, док је q_i ред из матрице филм-фактор.

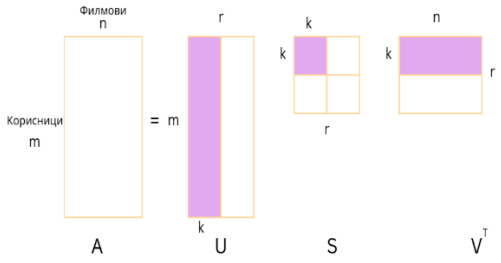
$$\hat{r}_{ui} = p_u q_i \tag{2}$$

2.1. Декомпозиција сингуларне вредности

Декомпозиција сингуларне вредности (*енг. Singular Value Decomposition, скр. СВД*) се први пут појавила као модел 2000. године у системима препоруке [5]. СВД модел, за разлику од основног модела факторизације матрице разлаже почетну матрицу A на 3 матрице: U, S и V. Димензије ових матрица су редом $m \times n$, $m \times r$, $r \times r$, $r \times n$, где су m број корисника, n је број филмова, r је ранг матрице A. У изразу (3) приказана је формула коју користи СВД модел.

$$A = USV^T \tag{3}$$

Циљ СВД модела је да смањи димензионалност почетне матрице и то ради тако што изабере константу k па из израчунате матрице U узме првих k колона, из матрице S узме исечак $k \times k$, а из матрице V узме првих k колона. На слици 1 приказана је ова редукција на нижи ранг.



Слика 1. Пример смањивања ранга [4]

2.2. СВД++

СВД модел није узимао у обзир ништа осим оцена корисника. Такође, СВД модел има ману проблема хладног старта. Овај проблем се јавља приликом регистрације новог корисника на систем. Не постоје подаци о кориснику, он још није оценио ни одгледао ниједан филм, па систем не зна шта би могао да му препоручи. Овај проблем се дешава и када се појави нови филм у систему који нема ниједну оцену. Како би се решио проблем хладног старта настао је нови модел СВД++ [6]. Разлика у односу на основни СВД модел је што СВД++ уводи посматрање корисниковог понашања, односно имплицитне податке о кориснику [7]. Поред имплицитних података СВД++ посматра и предрасуде корисника и предрасуде самог филма.

Предрасуда корисника према неком филму се јавља када корисник зна да ће у филму бити добар глумац или глумац којег воли, ако је филм режирао режисер који је пре тога увек режирао добре филмове. Ови фактори ће утицати на корисникову коначну оцену коју да филму након што га одгледа, а та оцена не мора да буде реална, односно може бити већа или мања у односу на стварну вредност оцене којом би филм требао бити оцењен.

Предрасуда филма показује колико је филм прецењен или потцењен у односу на реалну оцену. У предрасуду филма се убрајају и популарност филма у одређеном временском периоду, као што су новогодишњи филмови популарни у зимском периоду, хорор филмови су популарнији у периоду око празника Ноћи вештица и у том периоду је већа шанса да филм буде оцењен бољим оценама него у било ком другом периоду године.

У изразу (4) приказана је формула коју СВД++ модел користи за рачунање предикције оцене коју би корисник u дао филму i. У формули $\hat{r}_{ui}$ је израчуната предвиђена оцена, μ је укупна просечна оцена свих филмова, b_u је предрасуда корисника u, b_i је предрасуда филма i, q_i је вектор латентних фактора филма i, p_u је вектор латентних фактора корисника u, y_j су имплицитни фактори корисника са филмовима које је одгледао, $N(u)$ је скуп филмова са којима је корисник имао интеракцију (оценио их је, претраживао, одгледао).

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T \left(p_u + \frac{1}{\sqrt{|N(u)|}} \sum_{j \in N(u)} y_j\right) \tag{4}$$

2.3. ТимеСВД++

Како би се побољшао СВД++ модел и повећала тачност у предикцијама уведен је и временски фактор чиме је креиран ТимеСВД++ модел који је предложен 2010. године [5]. ТимеСВД++ модел узима у обзир и временски тренутак када је корисник оценио филм. Корисников укус и интересовања могу да се промене током времена, њега могу данас интересовати филмови који припадају жанру акција, а да за пар месеци или година промени интересовање и да га више занимају филмови који припадају жанру комедија. Расположење корисника може утицати на оцену коју ће дати одгледаном филму, већина корисника је обично уморна радним данима и ако погледају филм увече може се десити да га нису доживели исто као што би то био случај да су филм гледали викендом или када су одморни. ТимеСВД++ модел посматра корисниково понашање током читавог времена, а не само у блиској прошлости [3]. Он може уочити патерне корисниковог оцењивања.

Параметар корисниковог предрасуда b_u се мења током времена, јер филм који оцени једном оценом, након неког времена може оценити већом или мањом оценом, што се често дешава јер корисник у једном тренутку крене да упоређује филмове и оцене које је дао филмовима. Корисник може постати строжи у оцењивању током времена.

Параметар предрасуда филма b_i се мења током времена због утицаја популарности самог филма. ТимеСВД++ модел може уочити када је филм мање или више популаран на основу генералних оцена које филмови добијају коришћењем података о оценама и времену када су оцене додељене филму. Популарност новог филма може да порасте нагло, ако се зна да ће познати глумци глумити у филму. Може се десити да након неког времена популарност филма опадне, јер корисници стекну утисак да ли је филм добар или није и после неког времена га оцењују реалнијим оценама.

У изразу (5) приказана је формула коју користи ТимеСВД++ модел да израчуна оцену коју би корисник u дао филму i. Ова формула представља проширење формуле коју користи СВД++ модел, а која је приказана у изразу 4. тиме што је уведено да предрасуд корисника $b_u(t)$ зависи од времена,

предрасуд филма $b_i(t)$ зависи од времена и латентни фактори корисника $p_u(t)$ зависе од времена.

$$\hat{r}_{ui}(t) = \mu + b_u(t) + b_i(t) + q_i^T(p_u(t) + \frac{1}{\sqrt{|N(u)|}}\sum_{j\in N(u)} y_j) \quad (5)$$

3. МЕТРИКЕ ЕВАЛУАЦИЈЕ

За мерење грешке коју су модели направили у експериментима коришћене су средња апсолутна грешка и корен средње квадратне грешке.

3.1. Средња апсолутна грешка

Средња апсолутна грешка је грепка која се рачуна као однос суме апсолутних вредности разлика стварне и израчунате вредности и броја рачунања [9]. У изразу (6) је приказана формула рачунања средње апсолутне грешке, при чему n представља укупан број предвиђених вредности, y_{ti} је стварна оцена, док је y_i предвиђена оцена.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{ti} - y_i| \quad (6)$$

3.2. Корен средње квадратне грешке

Корен средње квадратне грешке се рачуна као корен збира квадрата разлике између израчунате предвиђене оцене и стварне оцене, подељене са бројем предвиђених оцена [10]. У изразу (7) је приказана формула рачунања корена средње квадратне грешке означене са *RMSE*, при чему n представља број предвиђених оцена, y_{ti} представља тачну вредност оцене, док y_i представља предвиђену вредност оцене.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ti} - y_i)^2} \quad (7)$$

4. ЕКСПЕРИМЕНТИ

За експерименте, њихову имплементацију и евалуацију резултата је коришћен програмски језик *Python.* Скуп података који је коришћен за тренирање различитих модела факторизације је *MovieLens* који је преузет са сајта *Kaggle* [8]. Овај скуп података се састоји од 2 табеле. Прва табела састоји се од 3 колоне, при чему је прва колона идентификатор филма, друга колона је назив филма и трећа колона представља жанрове филма. Друга табела се састоји од 4 колоне, где је прва колона идентификатор корисника, друга колона је идентификатор филма који је корисник оценио, трећа колона је вредност оцене коју је корисник дао филму и четврта колона је временски тренутак када је корисник оценио филм. Прва табела састоји се од преко 62 хиљаде филмова, а друга се састоји од преко 25 милиона података.
Како би се олакшало рачунање и избегли проблеми који настају због непопуњености матрице, подаци су филтрирани тако што су узети само они филмови који имају више од 50 оцена и само они корисници који су оценили више од 20 филмова. Овом филтрацијом добијена је матрица од преко 162 хиљаде корисника и преко 13 хиљада филмова.
Модели који су тренирани су модели описани у претходном поглављу: СВД, СВД++ и ТимеСВД++. Модели су тренирани над узорком података од 10 000 редова из матрице корисник-филм, за тренинг је узето 80% узорка, док је за тестни скуп узет остатак од 20% узорка. Метрике коришћене за евалуацију су средња апсолутна грешка наведена у изразу (6) и корен средње апсолутне грешке наведене у изразу (7).
Резултати су приказани у табели 1 из које се види да највећу грешку даје СВД модел, где је PMCE 2.4552, а MAE 2.1379. Много бољи резултат постигао се коришћењем СВД++ модела чије грешке су PMCE: 0.8805, а MAE: 0.6778. Најбољи резултат и најмању грешку дао је ТимеСВД++ модел који је за нијансу бољи у односу на СВД++ модел, PMCE: 0.8787, MAE: 0.6752.

Табела 1. Поређење резултата грешака различитих модела

Грешка / Модел	СВД	СВД++	ТимеСВД++
PMCE	2.4552	0.8805	0.8787
MAE	2.1379	0.6779	0.6752

5. ЗАКЉУЧАК

Факторизација матрице је увела велико побољшање у системима препоруке. Резултати који су добијени у експерименту су показали да када модел има више информација о корисницима и филмовима може да израчуна тачнију предикцију и тачније одреди који филм би се кориснику више свидео.
Поред података коришћених у експерименту додатни подаци који би могли да се користе су локација где је сниман филм, локација корисника, старосна доб корисника, пол корисника. Све ове податке би систем могао да искористи да направи прецизније предикције.
Због своје велике користи, системи препорука ће постати незаобилазна ставка у свим великим апликацијама, сајтовима и платформама и због тога будућа истраживања треба да се посвете не само побољшању употребе факторизације матрица у системима препоруке, истицању њених предности и умањивању њених мана, већ и проналазак нових техника које ће можда дати боље резултате у системима препорука.

6. ЛИТЕРАТУРА

[1] X. Zhou, J. He, G.Huang, Y. Yhang, “SVD based incremental approacher for recommender systems”, Joiurnal of Computer and System Sciences, Vol. 81, pp. 717-733, June 2015.

[2] R. Bell, C. Volinosky, “Matrix factorization techniques for recommender systems”, Computer, pp. 30-37, August 2009.

[3] Y. Koren, “Collaborative filtering with temporal dynamics”, Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 447-456, June 2009.

[4] F. O. Isinkaye, “Matrix Factorization in Recommender Systems: Algorithms, Applications and Peciuliar Challenges”, IETE Journal of research, pp. 6087-6100, November 2021.

[5] R. Mehta, K. Rana, “A Review on Matrix Factorization Techniques in Recommender Systems”, CSCITA, pp. 269-274, April 2017.

[6] M. Jallouli, S. Lajmi, I. Amous, “When contextual information meets recommender systems: extended SVD++ models”, International Journal of Computers and Applications, pp. 349-356, May 2020.

[7] B. Zhang, X. Zhou, J. Li, L. Li, “Recommendation Algorithm Based on Matrix SVD with Exponential Correction”, CIPAE 2020, pp. 71-75, October 2020.

[8]https://www.kaggle.com/datasets/parasharmanas/movie-recommendation-system, (pristupljeno u oktobru 2025.)

[9] S. Jiang, J. Li, W. Zhou, “An Application of SVD++ Method in Collaborative Filtering”, IEEE, pp. 192-197, January 2021.

[10] https://www.datacamp.com/tutorial/rmse, (pristupljeno u oktobru 2025.)

Кратка биографија:

Леополдина Ђанић рођена је у Врбасу 2001. године. Завршила је Електротехничку школу „Михајло Пупин“ у Новом Саду, 2020. године. Факултет техничких наука у Новом Саду је уписала 2020. године. Дипломирала је на Факултету техничких наука на одсеку Рачунарство и аутоматика 2024. године са просечном оценом 9.68. Године 2024. је уписала мастер академске студије на Факултету техничких наука у Новом Саду, студијски програм рачунарство и аутоматика.
Контакт:
leopoldina.djanic01@gmail.com