SISTEM PEREKOMENDASI DENGAN SINGULAR VALUE DECOMPOSITION DAN TEKNIK SIMILARITAS PEARSON CORRELATION

Rimbun Siringoringo[1], Jamaluddin[2], Gortap Lumbantoruan[3]

[1,2,3] Manajemen Informatika, Universitas Methodist Indonesia

[1]rimbun.ringo@gmail.com, [2]jac.satuno@gmail.com, [3]lumbatoruan.gortap@gmail.com

ABSTRACT

The growth of e-commerce has resulted in massive product information and huge volumes of data. This results in data overload problems. In the case of e-commerce, consumers or users spend a lot of time choosing the goods they need. The urgent question to be answered at this time is how to provide solutions related to intelligent information restrictions so that the existing information is truly information that is by preferences and needs. This research performs information filtering by applying the singular value decomposition method and the Pearson similarity technique to the book recommendation system. The data used is the Book-Crossing Dataset which is the reference dataset for many research recommendation systems. The resulting recommendations are then compared with e-commerce recommendations such as amazom.com. Based on the results of the study obtained data that the results of the recommendations in this study are very good and accurate.

Keywords: Singular Value Decomposition, Teknik Similaritas, Pearson Correlation

I. PENDAHULUAN

Data lembaga riset pasar *e-marketer*, jumlah pengguna internet di Indonesia diproyeksikan mencapai 175 juta orang pada tahun 2019, atau sekitar 65,3% dari total penduduk 268 juta jiwa. Data dari *Dimensional Research*, pertumbuhan *e-commerce* ritel di Indonesia akan tumbuh 133,5% menjadi US$ 16,5 miliar atau sekitar Rp 219 triliun pada 2022 (Mansur et al., 2019). Masalah utama yang diakibatkan oleh pertumbuhan e-*commerce* adalah terjadinya lonjakan arus informasi yang masif serta volume data yang sangat besar (Al-Ghuribi & Mohd Noah, 2019) sehingga kita diperhadapkan pada masalah *overload* data (Nilashi et al., 2016). Kondisi ini meningkatkan kompleksitas pengambilan keputusan, dimana keputusan yang diambil tidak akurat dan tidak efektif (Ocón Palma et al., 2020). Pada kasus *e-commerce*, konsumen atau pengguna menghabiskan waktu yang tidak sedikit dalam memilih barang kebutuhannya. Pertanyaan yang *urgen* untuk dijawab saat ini adalah bagaimana memberi solusi terkait pembatasan informasi secara cerdas (*intelligent*) supaya informasi yang ada benar-benar merupakan informasi yang sesuai dengan preferensi dan kebutuhan. Sistem perekomendasi atau *recommender system* merupakan solusi populer melakukan filterisasi informasi (Zhu et al., 2017), (Liu et al., 2019). Ada dua kategori sistem rekomendasi, sistem klasik dan multi-kriteria. Sistem klasik seperti *content based*, *collaboratif filtering*, *knowledge based* dan *hybrid* (Chen et al., 2015) memiliki kelemahan utama yaitu keterbatasan rekomendasi hanya berdasarkan satu kriteria saja atau *overall rating*. Kelemahan kedua adalah sistem klasik memberi rekomendasi tergantung pada frekuensi interaksi pengguna. Hal ini menjadi buruk untuk produk-produk baru atau *cold-start* yang belum banyak dilirik (Batmaz et al., 2019). Sistem reokmendasi klasik memiliki keterbatasan dalam hal skalabilitas. Ketika bekerja pada data yang sangat besar seperti data *MovieLens* dan *Book-Crossing Dataset* (Anwar et al., 2021). Sistem rekomenasi tradisional kurang menunjukkan akurasi yang baik. Salah satu solusi untuk mengatasi skalabilitas tersebut adalah faktorisasi matrik (Yuan et al., 2019). Faktorisasi matrik atau *matrix factorization* adalah penguraian suatu matriks menjadi beberapa buah matrik yang berukuran lebih kecil. Tujuan faktorisasi ini adalah untuk meningkatkan interpretasi data. Pada penelitian ini, faktorisasi matrik dengan metode *Singular Value Decomposition* atau yang lebih dikenal sebagai SVD diterapkan pada sistem perekomendasi.

II. ULASAN LITERATUR

2.1 Faktorisasi Matrik

Proses faktorisasi matrik akan memfaktorkan sebuah matriks menjadi lebih dari satu matriks yang lebih kecil (Wang et al., 2017). *Singular Value Decomposition* atau yang lebih dikenal sebagai SVD, adalah salah satu teknik dekomposisi berkaitan dengan nilai singular (*singular value*) suatu matriks. Sebuah matrik A m x n . Sebuah

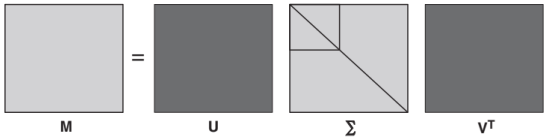
matrik $A = U\Sigma V^T$ adalah sebuah *Singular Value Decomposition* untuk A dengan *U* merupakan matriks orthogonal *m* x *m*, *V* matriks orthogonal *n* x *n* dan Σ matriks diagonal *m* x *n* bernilai riil tak negatif yang disebut nilai-nilai singular. Dengan kata lain Σ = diag (σ1, σ2, … , σn) terurut sehingga σ1 ≥ σ2 ≥ … ≥ σn. Jika $U = (u_1, u_2, \ldots, u_n)$ dan $V = (v_1, v_2, \ldots, v_n)$ maka :

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T \quad (1)$$

Teorema tersebut juga menyatakan bahwa matriks *Amxn* dapat dinyatakan sebagai dekomposisi matriks yaitu matriks *U*, ∑ dan *V* . Matriks ∑ merupakan matriks diagonal dengan elemen diagonalnya berupa nilai-nilai singular matriks *A*, sedangkan matriks *U* dan *V* merupakan matriks-matriks yang kolom-kolomnya berupa vektor singular kiri dan vektor singular kanan dari matriks *A* untuk nilai singular yang bersesuaian.

2.2 SVD pada sistem rekomendasi

Penerapan faktorisasi matrik dapat diterapkan pada sistem perekomendasi. Matrik M atau yang sering dikenal dengan user-item matrix (M) dapat di dekomposisi menjadi tiga buah matrik lain, yaitu matrik User-feature (U), matrik sigma (∑), dan matrik item-feature matrix (V^T)(Falk, 2019)

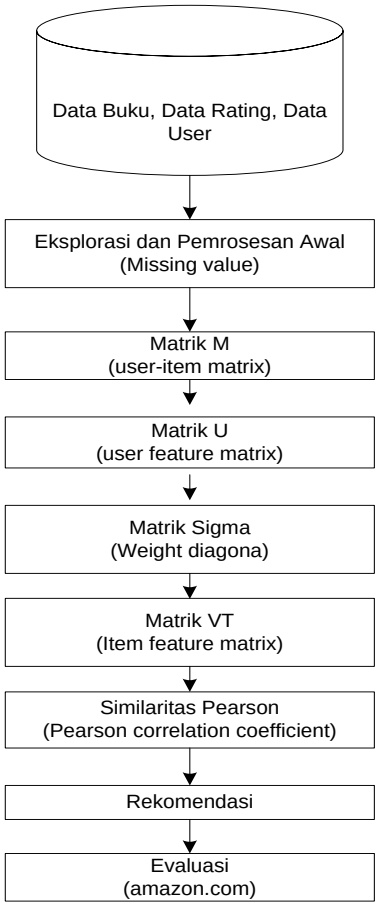


Gambar 1. Faktorisasi Matrik pada Sistem Rekomendasi

III. METODOLOGI

3.1 Alur penelitian

Alur penelitian ditampilkan pada gambar 1 berikut ini.



Gambar 1. Alur Penelitian

3.1 Data

Data yang digunakan pada penelitian ini adalah **Book-Crossing Dataset** yang bersumber dari http://www2.informatik.uni-freiburg.de/~cziegler/BX/. Terdapat tiga jenis data yang digunakan untuk membangun sistem perekomendasi yaitu buku, data rating, dan data user. Data rating adalah tabel yang berisi userID, ISBN, dan jumah rating buku. Data user berisi userID, lokas, dan usia pembeli. Data buku terdiri ISBN, bookTitle, bookAuthor, yearOfPublication, publisher, imageUrlS, imageUrlM, dan imageUrlL lihat Tabel 1 dan Tabel 2

Tabel 1. Data Rating

userID	ISBN	Book wRating
276725	034545104X	0
276726	0155061224	5
276727	0446520802	0
276729	052165615X	3
276729	0521795028	6

Tabel 2. Data User

userID	Location	Age
1	nyc, new york, usa	nan
2	stockton, california, usa	18.000
3	moscow, yukon territory, russia	nan
4	porto, v.n.gaia, portugal	17.000
5	farnborough, hants, united kingdom	nan

IV. HASIL DAN PEMBAHASAN

4.1 Pemrosesan awal

Pemrosesa awal bertujun untuk mempersiapkan dataset su paya dapat diproses dengan mudah serta membuang infor masi yang tidak dibutuhkan. Untuk mendapatkan data yan g representatif, maka perlu dilakukan penggabungan tabel rating dan dan tabel user dengan userID sebagai *field* kun ci. Sebahagian hasil penggabungan ditampilkan pada Tab el 3 berikut.

Tabel 3. Penggabungan tabel rating dan user

userID	ISBN	Book Rating	bookTitle
276725	034545104X	0	Flesh Tones: A Novel
2313	034545104X	5	Flesh Tones: A Novel
6543	034545104X	0	Flesh Tones: A Novel
8680	034545104X	5	Flesh Tones: A Novel
10314	034545104X	9	Flesh Tones: A Novel

Selanjutnya dilakukan filter data buku yang mengandung *missing value* serta pengelompokan berdasarkan judul bu ku. Pada tabel 4, user “2313” memberi rating “5” pada bu ku “034545104X”. Total rating buku “Flesh Tones: A No vel” adalah sebesar 60 seperti pada tabel 4.

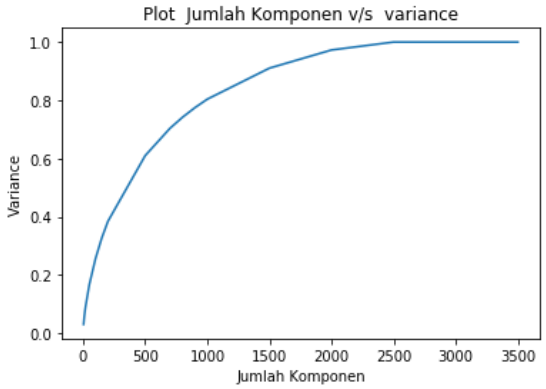
Tabel 4. Penggabungan Tabel Rating dan User

userID	ISBN	Book Rating	bookTitle	Total Rating Count
276725	034545104X	0	Flesh Tones: A Novel	60
2313	034545104X	5	Flesh Tones: A Novel	60
6543	034545104X	0	Flesh Tones: A Novel	60
8680	034545104X	5	Flesh Tones: A Novel	60
10314	034545104X	9	Flesh Tones: A Novel	60

4.2 Single Value Decomposition (SVD)

Langkah 1. Setting paremeter.

Parameter SVD yang dibutuhkan adalah jumlah komponen (*k*). Penentuan jumlah komponen berpengaruh terhadap nilai variance yang dihasilkan. Pada gambar 2 ditampilkan grafik plot antara jumlah komponen dengan nilai *variance*. Dengan Untuk *k* =12, masih diperoleh nilai variance yang kecil sebesar 0.0587



Gambar 1. Plot Jumlah Komponen vs Variance

Langkah 2. Pembentukan matrik utility (M)

Matriks *utility* disebut juga matriks *user-item*, yaitu matri k dua dimensi dengan vipot baris **userID** dan kolom adal ah **bookTitle**. Tabel berisi nilai rating setiap item buku. M atrik ini memiliki dimensi (40017, 2442). Pada matrik in i, setiap nilai yang kosong atau *missing value* diganti den gan *null*

Langkah 3. Matriks singular values

Untuk jumlah komponen =12, matrik *singular values* merupakan matriks dengan dimensi (kolom, baris) = (1, 12) seperti pada Tabel 5..

Tabel 5. Matrik *Singular Values*

Index	*Singular values*
0	329.194
1	224.117
2	210.927
3	191.241
4	188.546
5	178.070
6	167.620
7	162.655
8	160.522
9	156.632
10	152.577
11	147.536

Langkah 3. Pembentukan Matrik U

Matrik *U* atau dikenal dengan *User-feature matrix* (*U*) adalah hasil faktorisasi matrik *M* dengan jumlah baris 2441 dan fitur adalah 12 seperti pada Tabel 6.

Tabel 6. Tabel *User-feature matrix* (*U*)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
0	0.004	0.003	0.000	0.002	-0.001	0.002	0.004	0.002	-0.007	-0.004	-0.006	-0.003
1	0.010	0.000	-0.016	0.015	-0.021	0.006	-0.036	-0.008	-0.010	-0.013	-0.014	-0.004
2	0.043	0.004	0.006	-0.051	0.057	-0.008	-0.013	-0.040	-0.014	-0.059	0.013	-0.005
3	0.068	-0.037	-0.058	-0.004	-0.057	0.008	0.115	0.051	-0.026	-0.013	-0.058	-0.025
4	0.013	0.004	-0.009	-0.000	0.008	-0.004	-0.025	-0.022	0.005	-0.023	0.005	-0.000
5	0.008	-0.004	-0.007	0.014	-0.014	0.008	-0.013	0.011	-0.021	0.000	-0.019	-0.011
6	0.006	0.008	-0.008	-0.004	0.016	0.002	0.001	-0.000	0.003	-0.000	-0.003	0.000
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
2441	0.005	0.001	0.004	-0.006	0.002	0.008	-0.004	-0.001	-0.004	-0.008	-0.004	-0.005

Langah 4. Pembentukan matrik sigma (∑)

Matrik sigma adalah matrik diagonal dimana nilai diagonal berisi angka-angka tak negatif sebagaimana ditampilkan pada Tabel 7

Tabel 7. Matrik sigma

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
0	**329.194**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	**224.117**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	**210.927**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	**191.241**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	**188.546**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	**178.070**	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	**167.620**	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**162.655**	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**160.522**	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**156.632**	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**152.577**	0.000
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**147.536**

Langah 4. Pembentukan matrik V^T

Matrik V^T adalah *item-feature matrix* dengan jumlah baris *k=12* dan fitur item sebesar 40017.

Tabel 8. Matrik V^T

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**…**	**40016**
0	0.000	0.000	-0.000	0.001	0.000	0.000	0.004	0.001	0.001	0.000	**…**	0.001
1	0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.001	-0.000	-0.000	-0.000	**…**	-0.001
2	-0.000	0.000	0.000	-0.001	-0.000	-0.000	0.003	0.001	-0.002	-0.001	**…**	-0.001
3	-0.000	-0.001	-0.000	0.000	0.000	0.000	-0.006	-0.001	-0.000	-0.000	**…**	0.000
4	-0.000	0.000	0.000	0.001	-0.000	0.000	0.002	-0.000	0.001	0.001	**…**	-0.001
5	0.000	-0.000	0.000	0.001	-0.000	-0.000	-0.000	-0.001	0.000	0.000	**…**	-0.000
6	-0.000	-0.000	0.000	0.002	0.000	-0.000	-0.006	0.001	0.001	-0.001	**…**	0.002
7	0.000	-0.000	-0.000	-0.001	0.000	0.000	-0.004	-0.001	0.001	0.001	**…**	-0.000
8	-0.000	0.000	0.000	0.001	-0.000	0.000	0.002	-0.000	0.000	-0.000	**…**	-0.000
9	0.000	-0.000	0.000	-0.001	0.000	-0.000	-0.012	-0.001	0.002	0.003	**…**	0.000
10	-0.000	0.000	0.000	-0.001	0.000	-0.000	0.010	-0.000	0.002	0.001	**…**	-0.001
11	-0.000	-0.000	-0.000	-0.000	0.000	0.000	0.003	-0.001	0.000	-0.000	**…**	-0.001

Langkah 5. Pengukuran similaritas

Hasil matrik akhir adalah matrik similaritas antar item-item buku. Similaritas ditentukan berdasarkan *Pearson's R correlation coefficient.* Pada tabel 9 ditampilkan smililaritas antar item-item, dimana nilai similariats antar item yang sama adalah bernilai "1"

Tabel 9 Simililaritas Antar Item

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**…**	**2441**
0	**1.000**	0.328	0.287	0.643	0.101	0.653	0.263	0.790	0.719	0.672	0.594	**…**	0.549
1	0.328	**1.000**	0.035	0.181	0.564	0.744	0.099	0.219	0.425	0.630	0.103	**…**	0.275
2	0.287	0.035	**1.000**	0.116	0.723	-0.140	0.655	0.446	-0.004	-0.116	0.825	**…**	0.692
3	0.643	0.181	0.116	**1.000**	-0.067	0.498	0.160	0.811	0.887	0.619	0.195	**…**	0.191
4	0.101	0.564	0.723	-0.067	**1.000**	0.060	0.560	0.229	-0.042	-0.048	0.528	**…**	0.451
5	0.653	0.744	-0.140	0.498	0.060	**1.000**	-0.040	0.386	0.758	0.955	0.087	**…**	0.284
6	0.263	0.099	0.655	0.160	0.560	-0.040	**1.000**	0.423	0.207	0.022	0.348	**…**	0.351
7	0.790	0.219	0.446	0.811	0.229	0.386	0.423	**1.000**	0.730	0.491	0.548	**…**	0.565
8	0.719	0.425	-0.004	0.887	-0.042	0.758	0.207	0.730	**1.000**	0.843	0.116	**…**	0.213
9	0.672	0.630	-0.116	0.619	-0.048	0.955	0.022	0.491	0.843	**1.000**	0.050	**…**	0.316
10	0.142	0.114	0.446	0.338	0.256	0.171	-0.096	0.220	0.232	0.131	0.543	**…**	0.417
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
2441	0.549	0.275	0.692	0.191	0.451	0.284	0.351	0.565	0.213	0.316	0.822	**…**	**1.000**

Langkah 6. Penentuan rekomendasi

Berdasarkan matrik similaritas diatas, maka rekomendasi untuk buku dengan judul "*The Green Mile: Coffey's Hands (Green Mile Series)*" dengan batas korelasi **(corr <1.0) && (corr>0.9)** ditampilkan pada daftar berikut. Judul buku yang ditebalkan adalah buku dengan judul yang tepat dan sangat sesuai dengan judul "*The Green Mile: Coffey's Hands (Green Mile Series)*".

Tabel 10. Hasil Rekomendasi Buku

No	**Judul**
1	Needful Things
2	The Bachman Books: Rage, the Long Walk, Roadwork, the Running Man
3	**The Green Mile: Coffey on the Mile (Green Mile Series),**
4	**The Green Mile: Night Journey (Green Mile Series),**
5	**The Green Mile: The Bad Death of Eduard Delacroix (Green Mile Series),**
6	**The Green Mile: The Mouse on the Mile (Green Mile Series),**
7	The Shining,
8	**The Two Dead Girls (Green Mile Series)]**

Hasil rekomendasi pada tabel di atas dapat dibadingkan dengan sistem rekomendasi yang diperoleh pada situs e-commerce amazon.com. Dengan memasukkan kata kunci pencarian buku "*The Green Mile: Coffey's Hands (Green Mile Series)"* maka hasil yang diperoleh sangat mirip atau akurat seperti pada gambar 3 berikut.

Gambar 3. Hasil rekomendasi pada amazon.com

V. KESIMPULAN

Sistem perekomendasi dengan metode *singular value decomposition* atau SVD dapat mengatasi masalah sparse matrix yang sering menjadi kendala pada Sistem perekomendasi konvesional. Penelitian ini menerapkan Sistem perekomendasi dengan SVD dan teknik similaritas *Pearson*. Hasil penelitian menunjukkan bahwa penerapan teknik *Pearson* pada SVD dapat memberikan hasil rekomendasi yang baik. Hasil penelitian ini masih perlu diperbaiki dengan memperluas penerapan pada dataset yang lain supaya hasil yang diperoleh lebih teruji.

DAFTAR PUSTAKA

[1]. Al-Ghuribi, S. M., & Mohd Noah, S. A. (2019). Multi-Criteria Review-Based Recommender System-The State of the Art. *IEEE Access*, *7*, 169446–169468. https://doi.org/10.1109/ACCESS.2019.2954861

[2]. Anwar, T., Uma, V., & Srivastava, G. (2021). Rec-CFSVD++: Implementing Recommendation System Using Collaborative Filtering and Singular Value Decomposition (SVD)++. *International Journal of Information Technology & Decision Making*, 1–19. https://doi.org/10.1142/S0219622021500310

[3]. Batmaz, Z., Yurekli, A., Bilge, A., & Kaleli, C. (2019). A Review on Deep Learning for Recommender Systems: Challenges and Remedies. *Artif. Intell. Rev.*, *52*(1), 1–37. https://doi.org/10.1007/s10462-018-9654-y

[4]. Chen, L., Chen, G., & Wang, F. (2015). Recommender Systems Based on User Reviews: The State of the Art. *User Modeling and User-Adapted Interaction*, *25*(2), 99–154. https://doi.org/10.1007/s11257-015-9155-5

[5]. Falk, K. (2019). *Practical Recommender Systems*. Simon and Schuster.

[6]. Liu, X., Su, X., Ma, J., Zhu, Y., Zhu, X., & Tian, H. (2019). Information filtering based on eliminating redundant diffusion and compensating balance. *International Journal of Modern Physics B*, *33*(13), 1950129. https://doi.org/10.1142/S0217979219501297

[7]. Mansur, D., Sule, E., Kartini, D., & Oesman, Y. M. (2019). Trust and Habit As Key Success on Digital Consuming Behavior in Indonesia Mediated By Behavior Intention. *AFEBI Management and Business Review*, *3*(02), 16. https://doi.org/10.47312/ambr.v3i02.197

[8]. Nilashi, M., Dalvi-esfahani, M., Zamani, M., Ramayah, T., & Ibrahim, O. (2016). A Multi-Criteria Collaborative Filtering Recommender System Using Clustering and Regression Techniques. *Journal of Soft Computing and Decision Support Systems*, *3*(5), 24–30.

[9]. Ocón Palma, M. del C., Seeger, A.-M., & Heinzl, A. (2020). Mitigating Information Overload in e-Commerce Interactions with Conversational Agents. In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. Randolph, & T. Fischer (Eds.), *Information Systems and Neuroscience* (pp. 221–228). Springer International Publishing.

[10]. Wang, X., Zhong, Y., Zhang, L., & Xu, Y. (2017). Spatial Group Sparsity Regularized Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(11), 6287–6304. https://doi.org/10.1109/TGRS.2017.2724944

[11]. Yuan, X., Han, L., Qian, S., Xu, G., & Yan, H. (2019). Singular value decomposition based recommendation using imputed data. *Knowledge-Based Systems*, *163*, 485–494. https://doi.org/https://doi.org/10.1016/j.knosys.2018.09.011

[12]. Zhu, X., Tian, H., Chen, G., & Cai, S. (2017). Symmetrical and overloaded effect of diffusion in information filtering. *Physica A: Statistical Mechanics and Its Applications*, *483*, 9–15. https://doi.org/https://doi.org/10.1016/j.physa.2017.04.087