Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

PERTEMUAN KE: 11

Recommender System

KULIAH

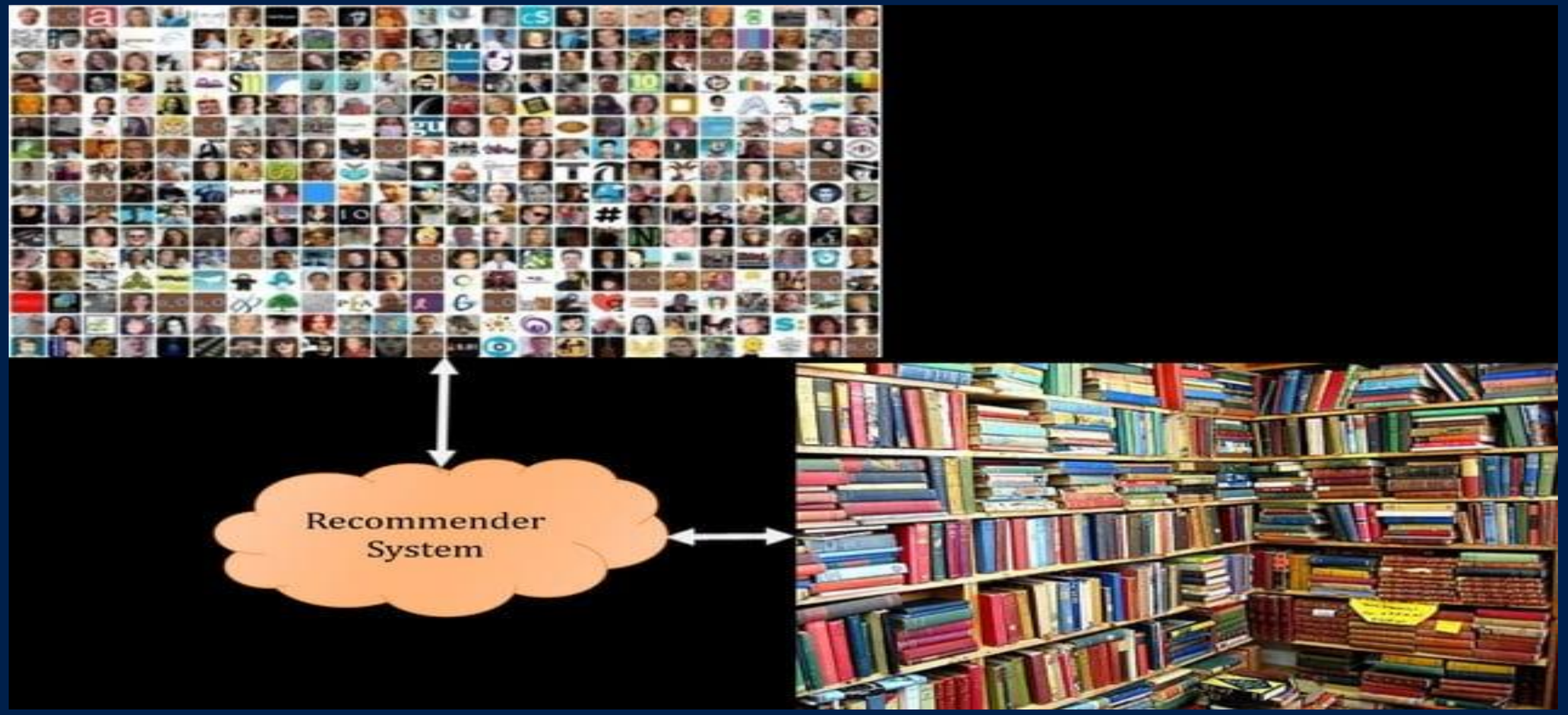
OLEH: NURJOKO

[12-09-2023] | [KODE MK: SSD23201 | MATAKULIAH: PENGANTAR SAINS DATA | SKS: 2 | VERSI: 01

Learning Objectives

- Mampu menjelaskan apa itu sistem rekomendasi, termasuk jenis-jenisnya seperti collaborative filtering, content-based filtering, dan hybrid systems.
- Memahami algoritma-algoritma yang digunakan dalam sistem rekomendasi, seperti collaborative filtering algorithms, content-based algorithms, dan matrix factorization.
- Memahami bagaimana data pengguna diolah dan dianalisis untuk menghasilkan rekomendasi yang personal dan relevan.
- Memahami aplikasi sistem rekomendasi dalam berbagai industri seperti e-commerce, streaming services, media sosial, dan lainnya.
- Learning Objective 5

Recommendation System



Pengenalan Sistem Rekomendasi

Machine Learning

Memahami keterhubungan (*relationships*) dan ketergantungan (*depedencies*) dalam suatu koleksi data adalah suatu aspek yang sangat penting dalam menganalisa koleksi data tersebut. Ketika tidak ada pendekatan pemodelan (*modelling approaches*) yang mudah untuk melakukan hal tersebut, maka pendekatan data (*data-driven approaches*) melalui metode-metode cerdas ***machine learning*** menjadi solusi alternatif.

Why Recommendation System Important?

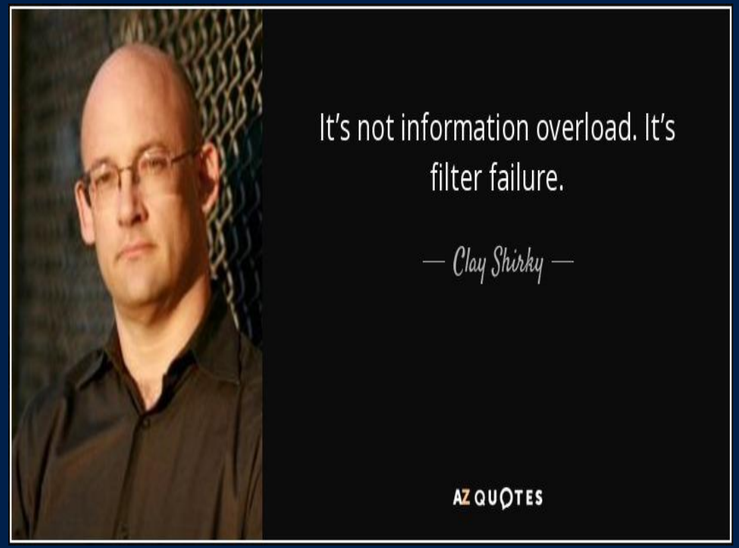
We have now entered a data explosion era.

It is hard for a person to find useful information from vast amounts of data.

In many cases, users don't fully realize their own needs, or their demand is difficult to express in words.

Definition of Recommendation System

Recommendation system is an application of information filtering. It studies the user‘s interests and preferences, and imply some rules to find the user’s personalized needs and actively and efficiently recommend information and content to users.



Applications of Recommendation System

Tons of websites:

Youtube, Spotify, Netflix, Google, Ebay...

Recommendation Engine – Examples

Facebook–"People You May Know"

Netflix–"Other Movies You May Enjoy"

LinkedIn–"Jobs You May Be Interested In"

Amazon–"Customer who bought this item also bought …"

YouTube–"Recommended Videos"

Google–"Search results adjusted"

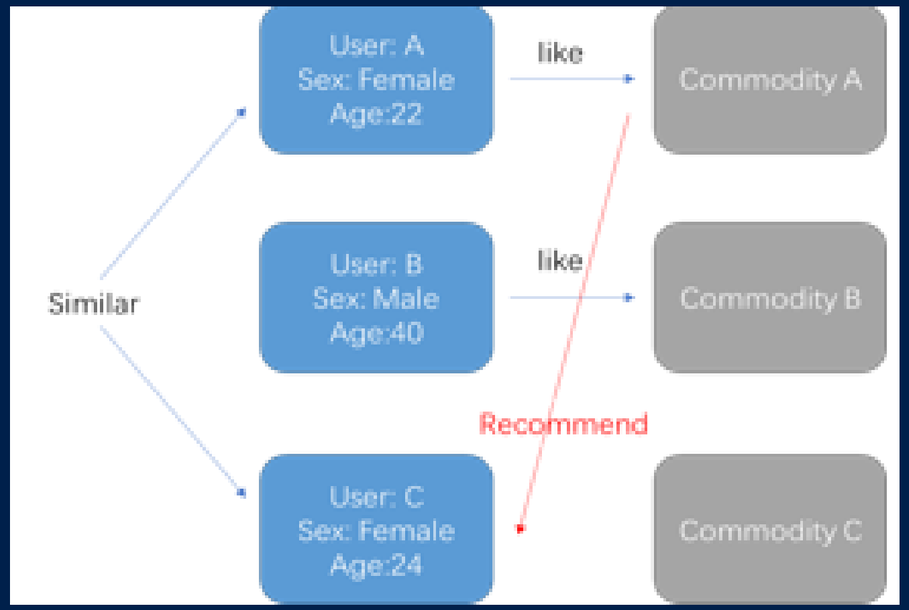
Pinterest–"Recommended Images"

Principle behind Recommendation System

- There are many different theories and algorithms behind recommendation system at current stage.
- Theories and algorithms are not independent.
- They are cooperating with each other to achieve the best results.

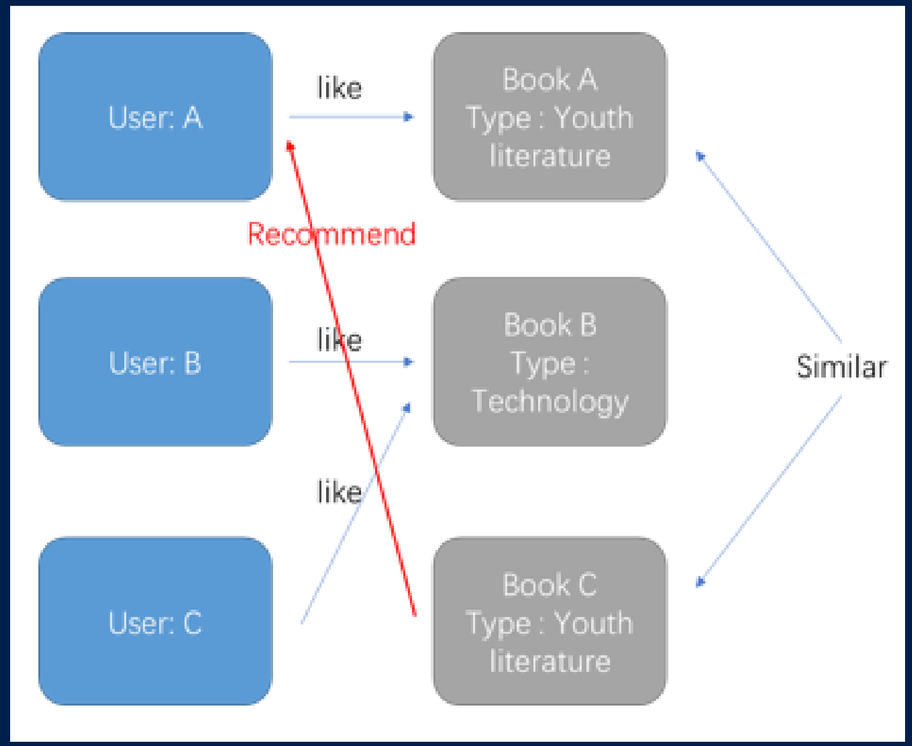
Basic Idea:
UserStatistic-based Recommendation

This is the simplest recommendation algorithm, it is simply based on the basic information of users in the system to find the relevance of the user, and then recommend other items which similar users like to the current user.



Basic Idea:
Content-based Recommendation

- The most widely used recommendation mechanism in early recommendation engine.
- Its core idea is based on the metadata of recommended items or content, discover the relevance of an article or content, and recommend similar item to user



Basic Idea:
Association Rule-based Recommendation

Associated product is not necessarily complementary product:

The findings were that men between 30- 40 years in age, shopping between 5pm and 7pm on Fridays, who purchased diapers on behalf of their wives were most likely to also have beer in their carts. This motivated the grocery store to move the beer isle closer to the diaper isle and wiz-boom-bang, instant 35% increase in sales of both.

Techniques : Data Acquisition

1. Explicit Data
   - Customer Ratings
   - Feedback
   - Demographics
   - Physiographics
   - Ephemeral Needs

2. Implicit Data
   - Purchase History
   - Click or Browse History

3. Product Information
   - Product Taxonomy
   - Product Attributes
   - Product Descriptions

Tipe Recommendation

- Secara umum, suatu sistem rekomendasi menggunakan salah satu dari dua strategi berikut ini, yaitu *content filtering* atau *collaboartive filtering*.
- *Content filtering* akan mendefiniskan profil untuk masing-masing pengguna atau *item*. Contoh profil film adalah genre, aktor, popularitas, dll; Profil pengguna adalah informasi geografis, isian kuisioner, dll.
  - Profil-profil ini memungkinkan sistem untuk mengasosiasikan pengguna dengan *items* yang sesuai.
  - Strategi berbasis konten ini membutuhkan informasi eksternal yang boleh jadi susah untuk diperoleh
- *Collaborative filtering* bergantung hanya pada aktifitas pengguna sebelumnya, misal transaksi sebelumnya atau pemberian *rating*, tanpa perlu mendefinisikan profil secara eksplisit.

Tipe Recommendation

- Dua metode utama yang sering digunakan pada *collaborative filtering*, yaitu: *nearest neighbors method* dan *latent variabel models*.
- Ada dua cara untuk memprediksi rating pada metode *nearest neighbors,* yaitu dengan pendekatan berorientasi *item* (*item-oriented approach*) atau pendekatan berorientasi pengguna (*user-oriented approach*).
- Pada metode berorientasi *item*, prediksi *rating* yang akan diberikan oleh seorang pengguna kepada suatu *item* adalah berdasarkan rating yang diberikan oleh pengguna tersebut kepada *items* tetangga terdekat dari *item* tersebut.
- Contoh: untuk film *Saving Privat Ryan*, maka film-film tetangganya yang mungkin adalah film pertempuran, film Spielberg, film Tom Hanks, dll.

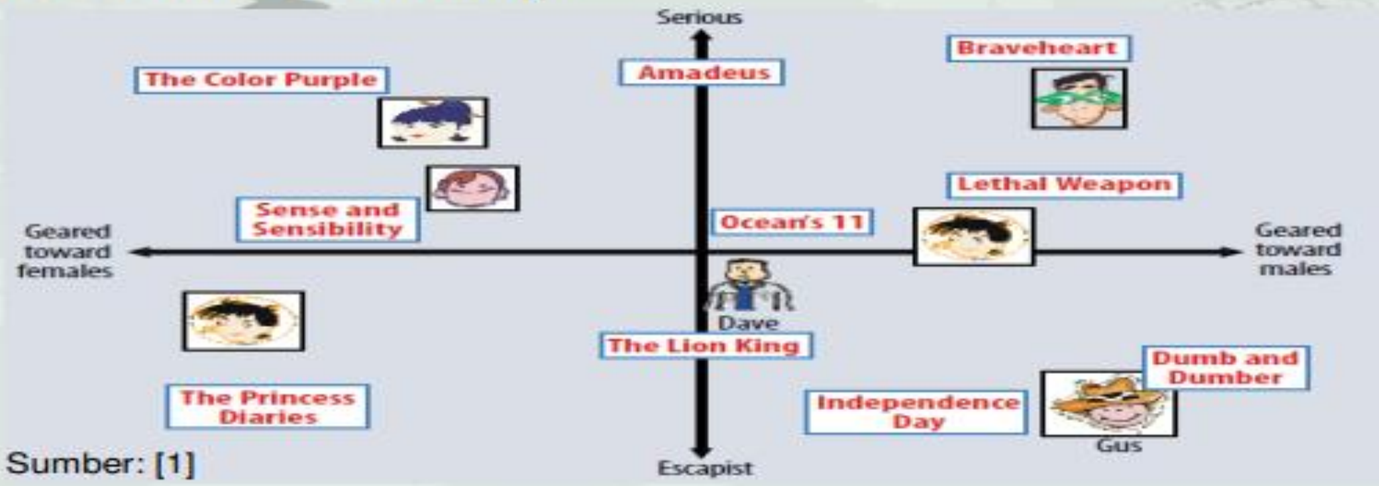
Tipe Recommendation

- Pada metode berorientasi pengguna, prediksi *rating* yang akan diberikan oleh seorang pengguna kepada suatu *item* adalah berdasarkan rating yang diberikan oleh pengguna tetangga terdekat kepada *item* tersebut.
- Contoh: Joe menyukai tiga film. Untuk membuat rekomendasi bagi Joe, sistem akan mencari pengguna yang mirip dengan Joe yang juga menyukai ketiga film tersebut, dan kemudian menentukan film lain yang mereka senangi. Pada contoh ini, ketiganya senang dengan film Saving Private Ryan, sehingga dijadikan rekomendasi pertama, dst.

Sumber: [1]

Tipe Recommendation

- Latent variabel models akan memprediksi rating berdasarkan posisi relatif dari pengguna dan item pada beberapa variabel tersembunyi yang diekstrak dari pola rating. Untuk film, variabel tersembunyi tersebut boleh jadi genre, misal komedi, drama, aksi, anak-anak, dll.
- Contoh: Misal terdapat dua variabel tersembunyi sebagai sumbu x dan y, serta posisi relatif pengguna dan film pada koordinat tsb. Dari gambar, misal, kita dapat memprediksi bahwa Gus akan menyukai film *Dumb and Dumber*, tapi tidak suka film *The Color Purple*.



Sumber: [1]

Matrix Factorization

- Beberapa realisasi paling sukses untuk *latent variabel models* adalah berdasarkan *matrix factorization*. Metode ini menjadi populer karena memberikan skalabilitas yang baik dengan akurasi yang prediktif.
- Pada metode ini, data biasanya direpresentasi dalam bentuk matrik, dimana satu dimensi menunjukkan *items*, sementara dimensi lainnya menunjukan pengguna, yaitu matrik R_{mxn}, dimana r_{ij} menunjukkan *rating* dari *item i* oleh pengguna *j*.

	P-1	P-2	...	P-n
I-1	r_{11}	r_{12}	...	r_{1n}
I-2	r_{21}	r_{22}	...	r_{2n}
:	:	:		:
I-m	r_{m1}	r_{m2}	...	r_{mn}

Matriks Faktorisasi

Representasi Sparse

- Biasanya matrik rating tersebut adalah jarang (*sparse*), karena seorang pengguna umumnya hanya memberikan rating kepada sebagian kecil *items* saja.

	p-1	p-2	...	p-n
o-1	r_{11}	-	...	-
o-2	-	r_{22}	...	r_{2n}
:	:	:		:
o-m	r_{m1}	r_{m2}	...	-

- Sehingga penggunaan *singular value decomposition* (SVD) untuk mengekstrak variabel tersembunyi seperti pada *latent semantic analysis* (LSA) hanya bisa dilakukan dengan sejumlah trik, misal mengganti entri yang tidak diketahui tersebut dengan nol (*imputation*) [2], akan tetapi menyebabkan peningkatan jumlah data yang jika tidak dilakukan secara benar akan merusak data.

Matriks Faktorisasi

Formulasi Umum

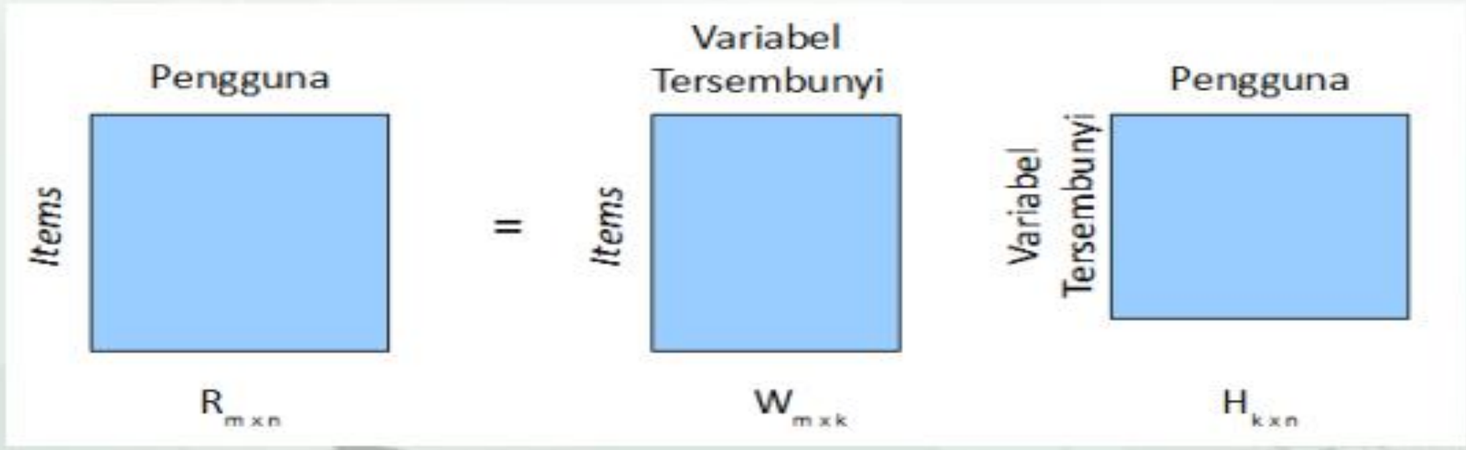
- Pendekatan alternatif adalah melakukan pemodelan secara langsung pada data *rating* yang diketahui saja (data training). Cara ini dapat diperoleh dengan melakukan formulasi masalah dengan bentuk umum sbb:

Diberikan suatu matrik R_{mxn}, maka masalah faktorisasi matrik adalah mencari matrik W_{mxk} dan H_{kxn} sedemikian sehingga $R \approx WH$, atau:

$$\min_{W,H} \; f(W,H) = \frac{1}{2}\|R - WH\|^2$$

Matriks Faktorisasi

Formulasi Umum: Interpretasi



- Untuk film, variabel tersembunyi tersebut boleh jadi berupa *genre*, misal komedi, drama, aksi, anak-anak, dll.
- Vektor baris dari matrik W adalah representasi *items* relatif terhadap variabel tersembunyi yang terekstraksi
- Vektor kolom dari matrik H adalah representasi pengguna relatif terhadap variabel tersembunyi yang terekstraksi

Matriks Faktorisasi

Formulasi Parsial

- Misal r_{ip} adalah rating yang berikan oleh pengguna *p* untuk item *i*, $\boldsymbol{w}_i$ adalah vektor baris dari *W* yang menunjukan vektor item *i*, $\boldsymbol{h}_p$ adalah vektor kolom dari *H* yang menunjukkan vektor pengguna *p*, maka:

$$\min_{\boldsymbol{h},\boldsymbol{w}} \; f(\boldsymbol{w},\boldsymbol{h}) = \frac{1}{2} \sum_{(i,p)\in T} (r_{ip} - \boldsymbol{w}_i^T \boldsymbol{h}_p)^2$$

dimana *T* adalah himpunan pasangan *(i,p)* dimana r_{ip} diketahui (data training)

- Untuk menghindari *overfitting*, bentuk teregularisasi sering juga digunakan, yaitu:

$$\min_{\boldsymbol{h},\boldsymbol{w}} \; f(\boldsymbol{w},\boldsymbol{h}) = \frac{1}{2} \sum_{(i,p)\in T} (r_{ip} - \boldsymbol{w}_i^T \boldsymbol{h}_p)^2 + \frac{\lambda}{2}(\|\boldsymbol{w}_i^T\|^2 + \|\boldsymbol{h}_p\|^2)$$

Matriks Faktorisasi

Algoritma Gradient Descent

- Algoritma dasar yang digunakan untuk memecahkan masalah *matrix factorization* adalah metode *gradient descent* [3]. Algoritma ini memodifikasi parameter dengan suatu besaran α berlawanan arah gradien, yaitu:

Algoritma *Gradient Descent*

1. $W = rand(m,k)$
2. $H = rand(k,n)$
3. **while (not** *stoping criteria***) do**
4. $H = H - \alpha_H \, \partial f(W,H) / \partial H$
5. $W = W - \alpha_W \, \partial f(W,H) / \partial W$
6. **end while**

Matriks Faktorisasi

Algoritma Gradient Descent

- Algoritma gradient descent tersebut memungkinkan kita untuk melakukan operasi hanya pada nilai rating yang diketahui saja, yaitu:

$$w_i \leftarrow w_i + \alpha_w (r_{ip} - w_i^T h_p) h_p$$
$$h_p \leftarrow h_p + \alpha_h (r_{ip} - w_i^T h_p) w_p$$

- Untuk bentuk teregularisasi akan menjadi:

$$w_i \leftarrow w_i + \alpha_w [(r_{ip} - w_i^T h_p) h_p - \lambda w_i]$$
$$h_p \leftarrow h_p + \alpha_h [(r_{ip} - w_i^T h_p) w_p - \lambda h_p]$$

dimana r_{ip} diketahui (data training)

Matriks Faktorisasi

Algoritma Lain

- *Alternating least squares* [4]. Karena baik W maupun H tidak diketahui maka formulasi *matrix factorization* tidak *convex*. Akan tetapi, jika kita membuat tetap (konstan/tidak bebas) salah satu parameter tersebut, maka bentuknya menjadi *convex* dan dapat diselesaikan dengan metode *least squares* secara bergantian (*alternating*) sbb:

Algoritma *Alternating Least Squares*
1. W = rand(m,k)
2. **while (not** *stoping criteria***) do**
3. Pecahkan SPL $W^T WH = W^T R$ untuk mendapatkan H baru
4. Pecahkan SPL $HH^T W = HR^T$ untuk mendapatkan W baru
5. **end while**

- *Probabilistic matrix factorization* [5]
- *Maximum-margin matrix factorization* [6]

Matriks Faktorisasi

Contoh

- Salah satu contoh aplikasi adalah sistem rekomendasi film pada *movielens.com* dimana salah satu data eksperiment yang dibangun dari sistem tersebut dan dibuat terbuka untuk publik memiliki karakteristik sbb:

1	Jumlah pengguna	943
2	Jumlah *items*	1682
3	Jumlah *rating*	100.000
4	*Sparsity*	93.7%
5	Jumlah data training	80.000
6	Jumlah data testing	20.000
7	*cross-validation*	5-fold

Matriks Faktorisasi

CONCLUSION

Fill in .....

REFERENCES

1) Y. Koren, R. Bell, C. Volinsky. *Matrix Factorization Techniques for Recommender Systems*, The IEEE Computer Society, 2009
2) B. M. Sarwar et al., *Application of Dimensionality Reduction in Recommender System – A Case Study*. Proceeding of KDD Workshop on Web Mining for eCommerce: Challenges and Oppurtunities, 2000
3) S. Funk. *Netflix Update: Try This at Home*. Dec. 2006 (http://sifter.org/~simon/journal/20061211.html)
4) R. Bell, Y. Koren. *Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights*. Proceeding IEEE International Conference on Data Mining, 2007
5) R. Salakhutdinov, A. Minih. *Probabilistic Matrix Factorization*. Advances in Neural Information Processing System 20, 2008
6) N. Srebreo, J. D. M. Rennie, T. S. Jaakkola. *Maximum-Margin Matrix Factorization*. Advances in Neural Information Processing System 17, 2005

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”

[12-09-2023] | [KODE MK: SSD23201 | MATAKULIAH: PENGANTAR SAINS DATA | SKS: 2 | VERSI: 01