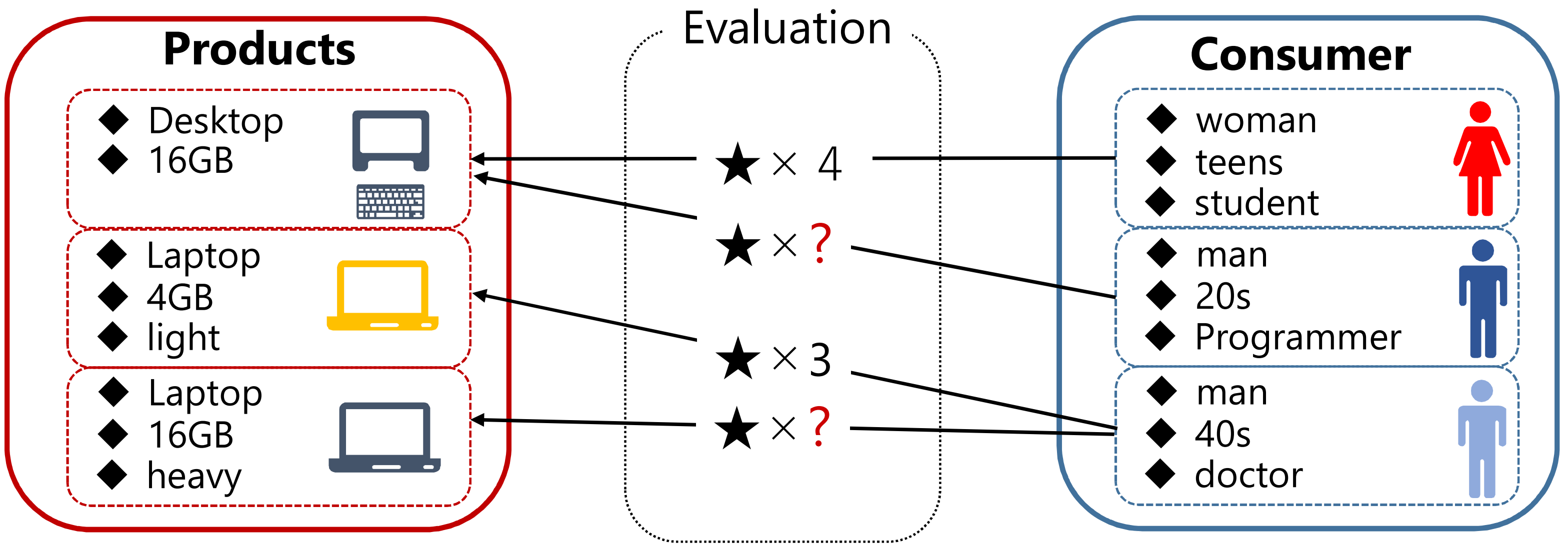
Recommendation via user's personality and extracted product features from text

Graduate School of Economics and Management, Tohoku University
Yosuke Sakai (M1)

Introduction

➢ Recommendation systems predict ★×? and recommend items which users seem to highly evaluate



➢ In the marketing, it is desirable that recommendation systems have interpretable results or insights

My research question

- ☐ What attributes consumers prefer products with what features ?

Literature review

- Many researches in the marketing fields make recommendation with consumer attributes and product features given explicitly “Genre” as explanation variable of probabilistic models.

	Method	Used information	Research features
Ansari et al. (2000)	Hierarchical regression model	･Genre of products ･Demographics	They capture unobserved features of products by introducing ‘Product heterogeneity’
Ying et al. (2006)	Hierarchical ordinal probit model	･Genre of products ･Demographics	In addition to Ansari et al.(2000), they modeled selecting action of users.
Chung and Rao (2012)	Hierarchical ordinal probit model	･Genre of products ･Demographics	They capture unobserved features of products by residuals of other users.

- They partially answer my question as ‘Interaction effects ’ between product features and user attributes.

Ex) The older users seem to prefer ‘Classic’, ‘Thriller’, ‘Drama’ movie.

➡ These are marketing insights, which can help recommend new user

Literature review

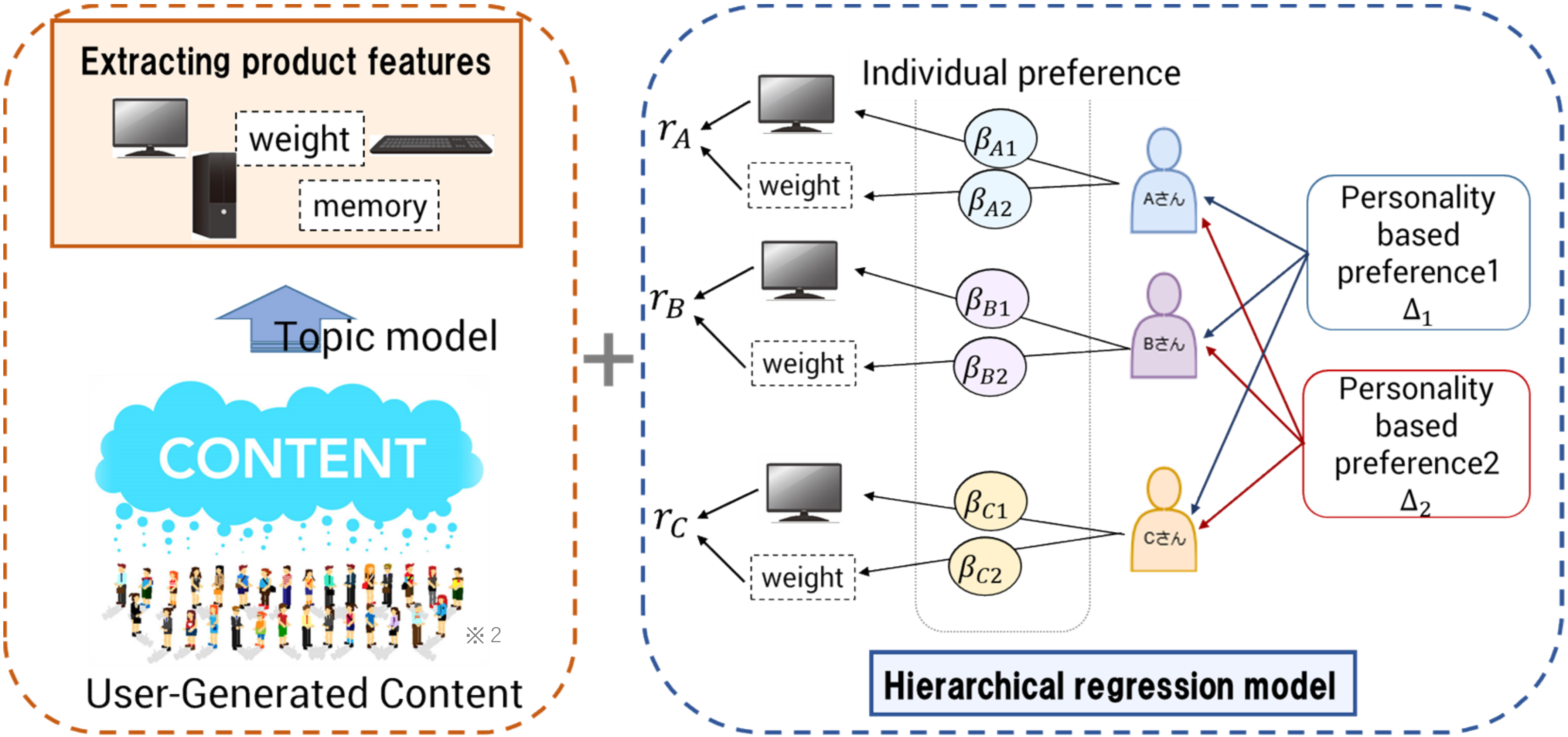
Problem of previous research

- They used product features given explicitly, ‘Genres’
  - These may be different from actual impression which users have
  - Insights given by their analysis are limited to about these features
- They used “Demographics” as user attributes
  - Demographics may not relate to preference very much

My research

- I used Topic model against ‘User-generated content’, and extracted more suite to user’s impression and more flexible features automatically
- I used user’s personality which is said to strongly related to preference of products especially “entertainment products” (Toubia et al.2019, Rentfrow et al. 2011)

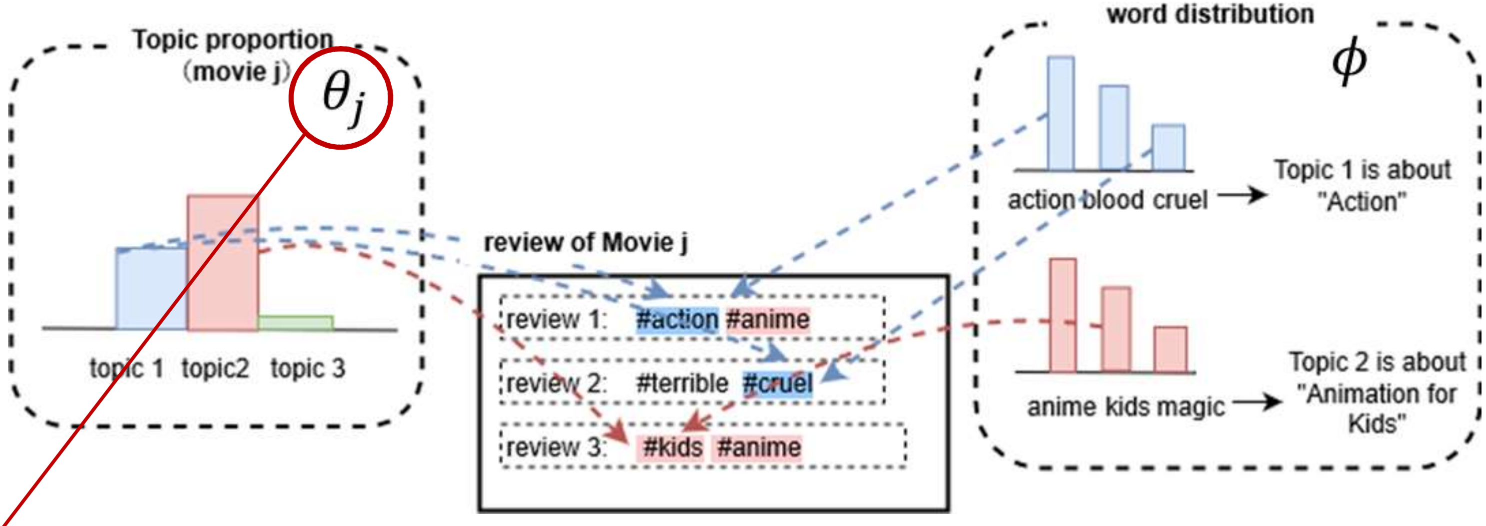
Outline figure of empirical study



※2　Image from https://medium.com/digital-reflections/the-rise-of-user-generated-content-24e85428f13e

What’s “topic model”?

- “Latent Dirichlet Allocation (Blei et al. 2003)” and its extension models are called “Topic model”
- It is used for dimension reduction or summary of documents.



Intuitive image of “LDA”

- <u>θ_j is used as features of product j</u>

Empirical study

- <u>Outline of empirical study</u>
  - Empirical study is divided into two stage
  - Data are reviews of movies. So, in this study, items are “movie”.

Stage 1

- Extraction of movie’s features by “Latent Dirichlet Allocation (Blei et al. 2003)” from UGC text
  - Interpreting means of topics through word distribution $\boldsymbol{\phi_k}$.
  - Topic proportion $\boldsymbol{\theta_j}$ are used as feature of movie j

Stage 2

- Modeling user’s rating by hierarchical regression model of Ansari et al.(2000)

$$r_{ij} = \boldsymbol{\theta}_j' \boldsymbol{\beta}_i + \boldsymbol{X}_i{}' \boldsymbol{\tau}_j + \epsilon_{ij} \quad \epsilon_{ij} \sim N\left(0, \sigma^2\right)$$
$$\boldsymbol{\beta}_i = \boldsymbol{\Delta}' \boldsymbol{X}_i + \boldsymbol{\lambda}_i \quad \boldsymbol{\lambda}_i \sim N_{K+1}(0, \boldsymbol{\Lambda})$$
$$\boldsymbol{\tau}_j = \boldsymbol{\gamma}_j \quad \boldsymbol{\gamma}_j \sim N_{p+1}(0, \boldsymbol{\Gamma})$$

r_{ij} : rating of movie j by user i

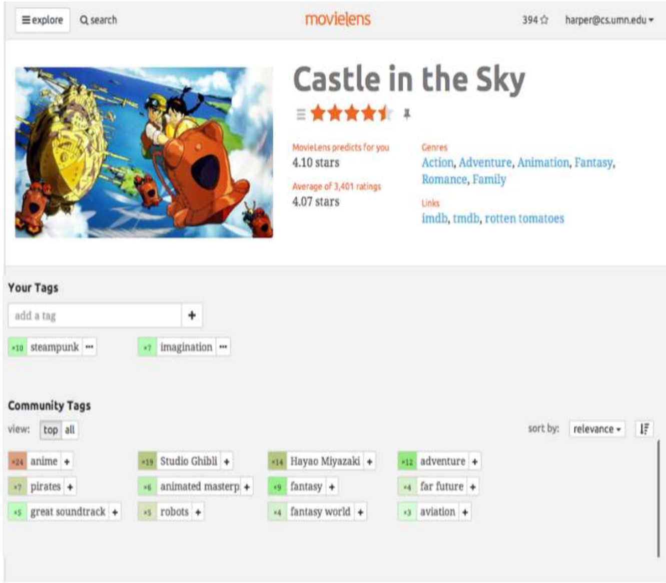
$\boldsymbol{\theta_j}$: Topic proportion of movie j $\left(\boldsymbol{\theta_j} \in \mathbb{R}^{\boldsymbol{K+1}}\right)$

$\boldsymbol{X_i}$: information of user i $\left(\boldsymbol{X_i} \in \mathbb{R}^{\boldsymbol{p+1}}\right)$

Data for empirical study

Stage 1

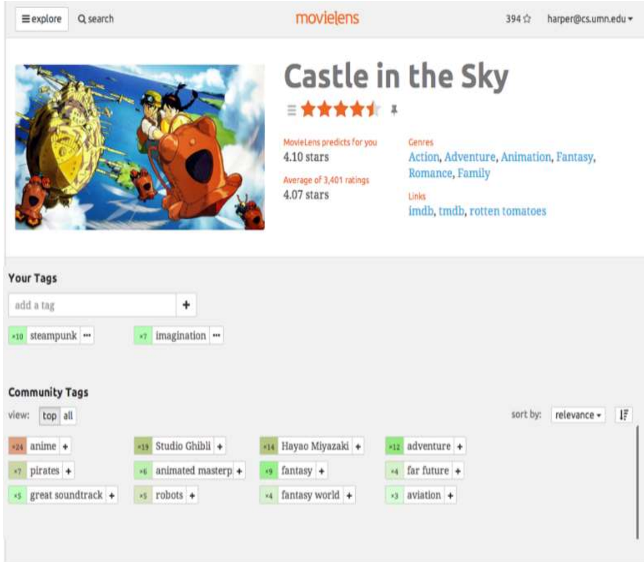
- UGC tag data of movies (“Movielens 10M”)
  - 7,038 movies with 71,060 word tags which given by users.
  - Each movies have several following 18 genres.
    - *“Action“, "Adventure“, "Animation“, "Children's“, "Comedy" ,"Crime" , "Documentary“, "Drama“, "Fantasy“, "Film-Noir“, "Horror“, "Musical“, "Mystery“, "Romance“, "Sci-Fi“, "Thriller“, "War“, "Western"*



Stage 2

- User’s rating about movies and personality (“Personality 2018” provided by grouplens)
  - Ratings (★×0.5 ~★×5 in 10 revels) about 7,038 movies given by users
  - User’s personalities are given as score (1~7 by 0.5 point) by following viewpoints
    - *“Openness”, “Agreeableness”, ”Emotional stability”, “Conscientiousness”, “Extraversion”* →called “The Big five personality”

Data for empirical study



Stage 1

- Review data of movies (“Movielens 10M”)
  - 7,038 movies with 71,060 word tags which given by users.
  - Each movies have several following 18 genres.
    - *“Action“, "Adventure“, "Animation“, "Children's“, "Comedy“, "Crime", "Documentary“, "Drama“, "Fantasy“, "Film-Noir“, "Horror“, "Musical“, "Mystery“, "Romance“, "Sci-Fi“, "Thriller“, "War“, "Western“*

Common

Different

Stage 2

- Review data of movies (“Personality 2018” provided by grouplens)
  - Ratings (★×0.5 ~★×5 in 10 revels) about 7,038 movies given by users
  - User’s personalities are given as score (1~7 by 0.5 point) by following viewpoints
    - *“Openness”, “Agreeableness”, ”Emotional stability”, “Conscientiousness”, “Extraversion”* →called “The Big five personality”

Data for empirical study

- What's "The Big five personality" ?
  - 5 viewpoints for measuring one's personality which are often used in psychology.

	Relating Feature	Opposite Feature
Openness	Open to new experience, complex	Conventional, Uncreative
Agreeableness	Sympathetic, warm	Critical, Quarrelsome
Emotional stability	Anxious, easy to upset	Calm, Emotional stable
Conscientiousness	Dependable, self-disciplined	Disorganized, Careless
Extraversion	Extraverted, enthusiastic	Quiet

- Higher score means higher tendency about the viewpoints

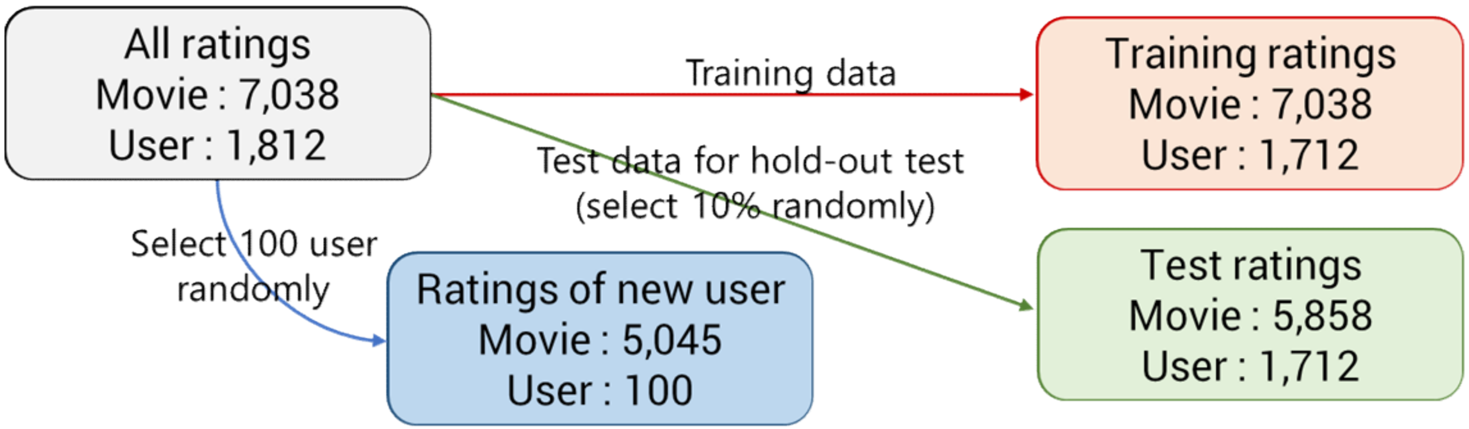
Data for empirical study

Stage 1

- Deleted tag which occurs less than 2.
- Deleted tag about name of actor or director.

Number of tag	Vocabulary
71,060	4,042

Stage 2



Settings for empirical study

Stage 1

- Setting genre's names as seed words. Ex) Word "Action" → always go to "topic 1"
- Deciding total number of topics as 35 because of interpretability ($K = 35$)
- Sampling 1000 times and using mean of parameters from 900 to 1000.
- Setting hyper parameters as $\alpha = \frac{1}{K}, \beta = 0.1$

Stage 2

- Sampling 10000 times and using mean of parameters from 5000 to 10000.

Result of empirical study

Stage 1

- <u>Interpretation of topics</u>

↓ Top 40 words having high ϕ_k

↓ Top 10 movies having high θ_k

Bigger words means they have larger ϕ_k

Topic k
“meaning of topic”

Result of empirical study

Stage 1

➢ Interpretation of topics

Henry: Portrait of a Serial Killer (1986)
Untouchables, The (1987)
Ocean's Eleven (2001)
Boondock Saints, The (2000)
Red Dragon (2002)
Analyze This (1999)
Falling Down (1993)
Casino (1995)
Bronx Tale, A (1993)
Copycat (1995)

Topic 6
“crime”

Result of empirical study

Stage 1

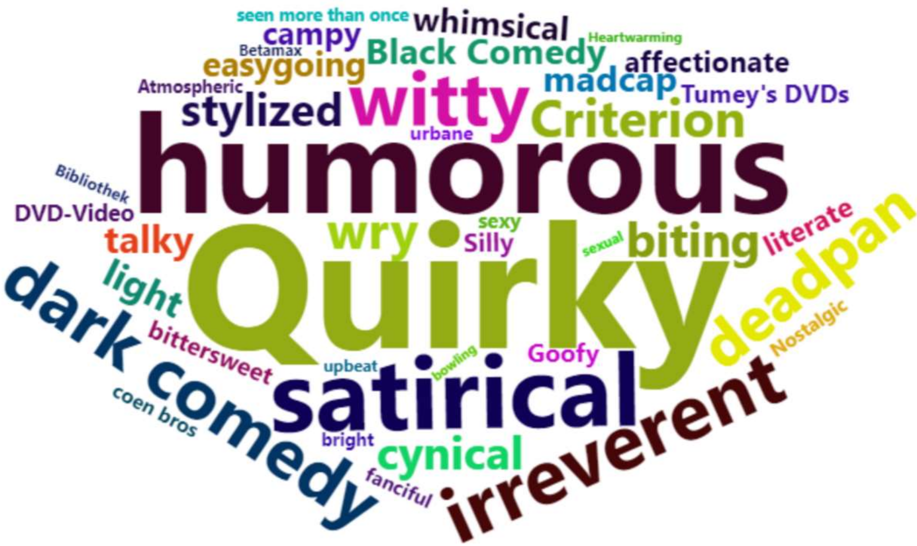
➢ Interpretation of topics

Topic 21
“Award, classic”

Result of empirical study

Stage 1

➢ Interpretation of topics



Wet Hot American Summer (2001)
Going Places (Les Valseuses) (1974)
Drowning by Numbers (1988)
Mr. Hulot's Holiday (Les Vacances de M. Hulot) (1953)
Day for Night (Nuit Amテウricaine, La) (1973)
CQ (2001)
Irma Vep (1996)
Woman Is a Woman, A (Une femme est une femme) (1961)
Stolen Kisses (Baisers volテウs) (1968)
Clean Slate (Coup de Torchon) (1981)

Topic 27
“With message (Satirically)”

Result of empirical study

Stage 2

- Hold-out test for existing user
  - Model 1: model of this research
  - Model 2 : $r_{ij} = \boldsymbol{G_j}'\boldsymbol{\beta_i} + \boldsymbol{X_i}'\boldsymbol{\tau_j} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$

$$\boldsymbol{\beta_i} = \boldsymbol{\Delta}'\boldsymbol{X_i} + \boldsymbol{\lambda_i} \quad \boldsymbol{\lambda_i} \sim N_{18+}(0, \boldsymbol{\Lambda})$$

$$\boldsymbol{\tau_j} = \boldsymbol{\gamma_j} \quad \boldsymbol{\gamma_j} \sim N_{p+1}(0, \boldsymbol{\Gamma})$$

$\boldsymbol{G_j}$: Genres of product $\boldsymbol{j}$ $\left(\boldsymbol{G_j} \in \mathbb{R}^{18+1}\right)$

These are dummy variable

- Genres are following 18 genres.
- *"Action", "Adventure", "Animation", "Children's", "Comedy", "Crime", "Documentary", "Drama", "Fantasy", "Film-Noir", "Horror", "Musical", "Mystery", "Romance", "Sci-Fi", "Thriller", "War", "Western"*

- Model 3 : $r_{ij} = \boldsymbol{\theta_j}'\boldsymbol{\beta_i} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$

$$\boldsymbol{\beta_i} = \boldsymbol{\Delta}'\boldsymbol{X_i} + \boldsymbol{\lambda_i} \quad \boldsymbol{\lambda_i} \sim N_{K+1}(0, \boldsymbol{\Lambda})$$

- Model 4 : $r_{ij} = \boldsymbol{G_j}'\boldsymbol{\beta_i} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$

$$\boldsymbol{\beta_i} = \boldsymbol{\Delta}'\boldsymbol{X_i} + \boldsymbol{\lambda_i} \quad \boldsymbol{\lambda_i} \sim N_{K+1}(0, \boldsymbol{\Lambda})$$

Not consider product heterogeneity

Result of empirical study

Stage 2

- Hold-out test and WAIC for existing user

	RMSE	WAIC
Model 1	**0.7843**	**1.202**
Model 2	0.8216	1.241
Model 3	0.8572	1.275
Model 4	0.9172	1.335

- WAIC (Watanabe 2010) is information criterion to evaluate generation error of models
- A model which has lower WAIC has better predicted performance

- Model 1 shows the highest accuracy.
- Even when I didn't consider "Product heterogeneity", Model 3 shows better predicted performance than Model 4.
- So, I think topics describe more enough or suite to user's impression products than "Genre"

Result of empirical study

Stage 2

- Hold-out test for new user
  - Model 1: model of this research
  - Model 5 : $r_{ij} = \boldsymbol{\theta_j}'\boldsymbol{\beta_i} + \boldsymbol{X_i}'\boldsymbol{\beta_j} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$

$$\boldsymbol{\beta_i} = \overline{\boldsymbol{\beta}} + \boldsymbol{\lambda_i} \quad \boldsymbol{\lambda_i} \sim N_K(0, \boldsymbol{\Lambda})$$

$$\boldsymbol{\tau_j} = \overline{\boldsymbol{\tau}} + \boldsymbol{\gamma_j} \quad \boldsymbol{\gamma_j} \sim N_{p+1}(0, \boldsymbol{\Gamma})$$

➡ Not consider interaction effect

	RMSE	WAIC
Model 1	**0.827**	**1.320**
Model 2	0.846	1.432

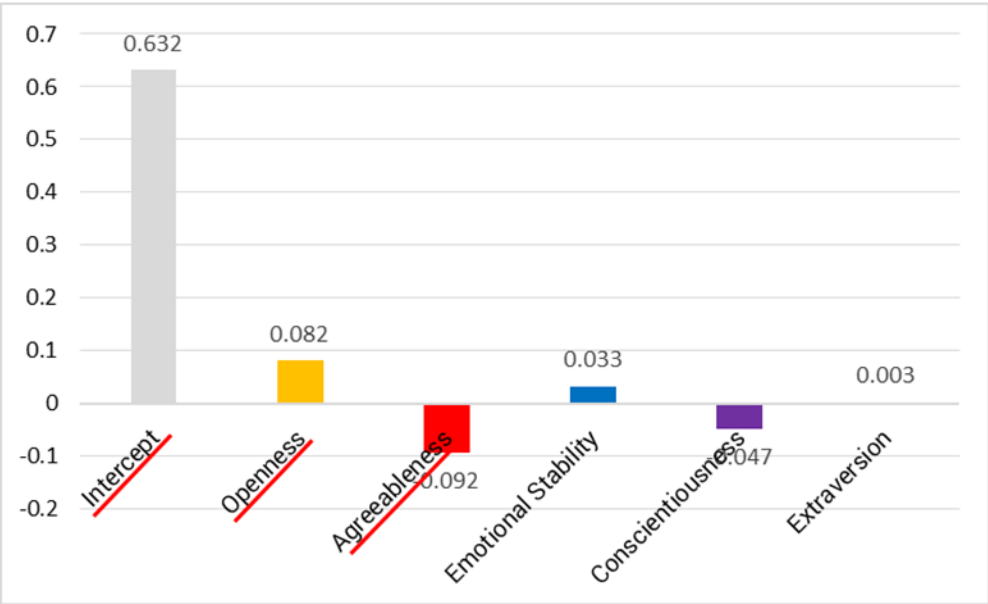
- Considering interaction effect between products features and user personality improved recommendation accuracy for new user

Result of empirical study

Stage 2

➢ Check interaction effects from Δ

Topic 6 "Crime"



※red underline means significant at 95% level

Openness
... Users who have high "Open to new experience, complex " personality prefer Crime movie

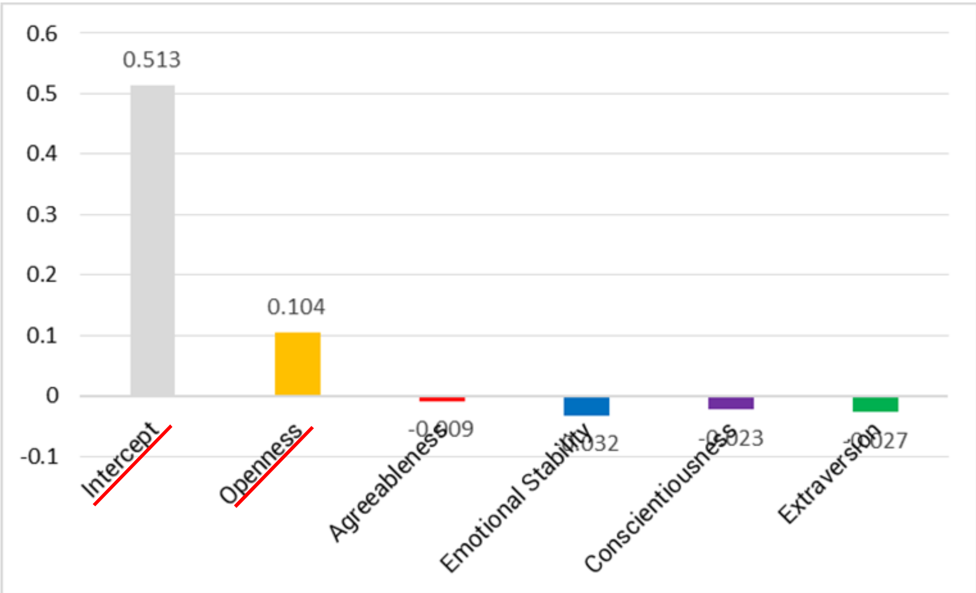
Agreeableness
... Users who have high "Sympathetic, Warm" personality don't prefer Crime movie.

Result of empirical study

Stage 2

➢ Check interaction effects from Δ

Topic 21 "Award, classic"



※red underline means significant at 95% level

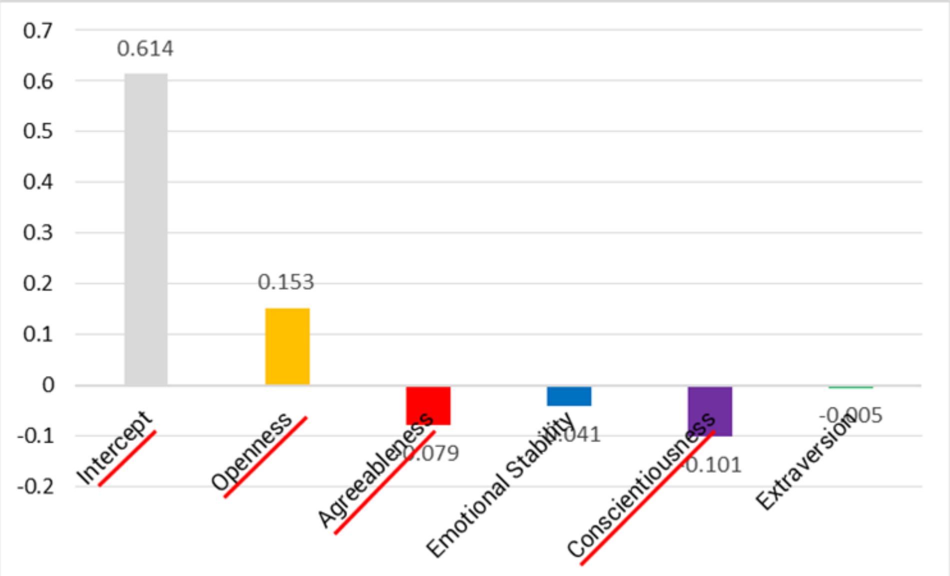
Openness
… Users who have high "Open to new experience, complex " personality prefer classic movie with award.

Result of empirical study

Stage 2

➢ <u>Check interaction effects from Δ</u>

Topic 27 “With message (Satirically)”



※red underline means significant at 95% level

Openness
… Users who have high “Open to new experience, complex ” personality prefer movies with message (Satirically).

Agreeableness
… Users who have high “Sympathetic, Warm” personality don’t prefer movie with message (Satirically).

Conscientious
… Users who have high “Dependable, self-disciplined” personality don’t prefer movies with message (Satirically).

Conclusion and challenge, future work

- Through extracting feature of products from UGC text by topic model
  and modeling rating by hierarchical regression model, I could improve recommendation accuracy.
- Existence of interaction effect between extracted feature and user's personality
  was revealed, and it could improve recommendation accuracy for new user.

【Challenge】

- Primitive topic model → more appropriate model.
- Automating decision of number of topics K using information criterion.

【Future work】

- Simultaneous estimation of topics and parameters in regression model.
- Hierarchical regression model → Hierarchical ordinal probit model.
  Ex) Ansari and Zhang (2018)

Reference

- Asim Ansari, Skander Essegaier, Rajeev Kohli (2000), “Internet Recommendation Systems” *Journal of marketing research Vol.37 pp. 363-375*
- David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003) “Latent Dirichlet Allocation” *Journal of Machine Learning Research*
- Jaihak Chung, Vithala R.Rao (2012) “A General Consumer Preference Model for Experience Products: Application to Internet Recommendation Services” *Journal of marketing research Vol 49 pp. 289-305*
- Oliver Toubia, Garud Iyengar, Renee Bunnell, and Alain Lemaire. (2019) “Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption” *Journal of Marketing Research*
- Peter J. Rentfrow, Lewis R. Goldberg, Ran Zilca (2011) “Listening, Watching, and Reading: The Structure and Correlates of Entertainment Preferences” *Journal of Personality Vol 79, pp.223-258*
- Sumio Watanabe (2010) ” Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory” *Journal of Machine Learning Research 11*
- YUANPINGYING, FRED FEINBERG, and MICHEL WEDEL(2006),“Leveraging Missing Ratings to Improve Online Recommendation Systems“ *Marketing Science*