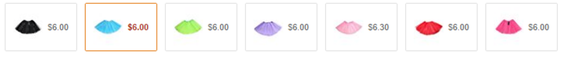
Redundancies in Data and their Effect on the Evaluation of Recommendation Systems: A case study on the Amazon Reviews Dataset

The Amazon Reviews Datasets Collection

- Crawled from Amazon website
  - *Huge* (35 M reviews), *temporal* (spans over 18 years, up to March 2013), *diverse* (#30 product categories), *heterogeneous* (review text, ratings)
- Contain duplicates for items that are variants of the same product
  - Different size, color
  - dvd and blue-ray versions



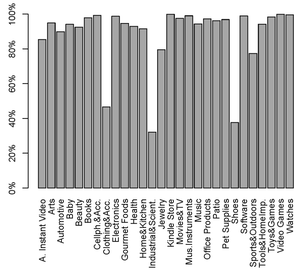
★★★★★ tutu and sparkly with stars not glitter which is a good thing since you don't have to worry about glitter ...
By Charlette K on November 29, 2016
Color: Red | Verified Purchase
Really cute, tutu and sparkly with stars not glitter which is a good thing since you don't have to worry about glitter falling off and getting all over the place. That was my concern when buying it. Other than that, it fit my 3 yr old perfectly.

This review goes to all different colour variants

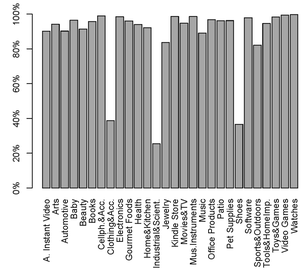
Elimination of Duplicates

- We kept only 1 of the product variants and its corresponding reviews and ratings

User	Product	Time	Review	Rating
user 1	item 1	t_1	"As a professional plumber..."	3
~~user 1~~	~~item 2~~	~~t_7~~	~~"As a professional plumber..."~~	~~3~~
~~user 1~~	~~item 3~~	~~t_7~~	~~"As a professional plumber..."~~	~~3~~
user 3	item 8	t_2	"This is a great product..."	5
user 3	item 7	t_3	"This is a great product..."	5



% of retained items per category

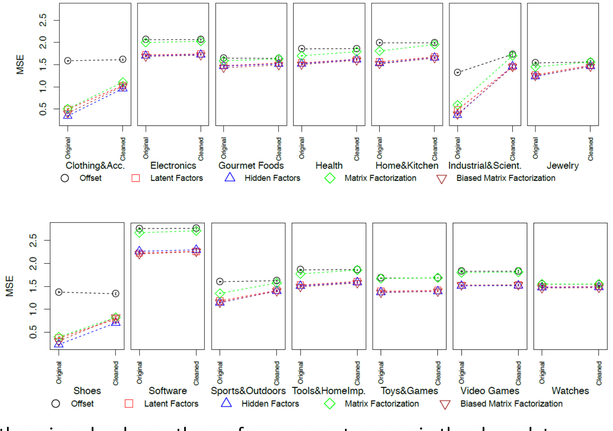


% of retained reviews per category

Evaluating Redundancy Effect on Recommenders

- Original (with redundancies) vs cleaned dataset
- Experimentation procedure
  - Randomly split each Amazon dataset D into training (80%) D_{tr}, testing (10%) D_t and validation (10%) D_v sets
  - Due to redundancies in the original dataset D, D_{tr}, D_t and D_v sets are *not disjoint*. They are *disjoint* in the cleaned dataset.
- Compared methods
  - *Offset* (only numerical ratings)
  - *Latent Factors* (only numerical ratings)
  - *Hidden Factors* (numerical ratings & review texts)
  - *Matrix Factorization* (only numerical ratings)
  - *Biased Matrix Factorization* (only numerical ratings)

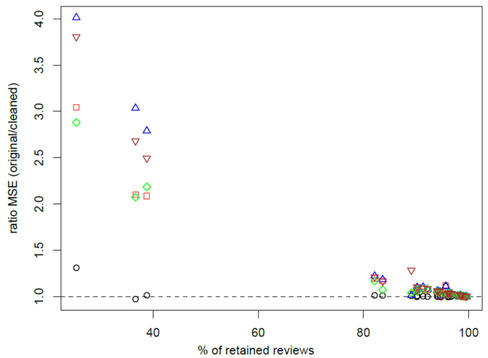
Effects on the Quality of Recommendations



- If there is redundancy, the performance gets worse in the clean data.
- The performance differences are not as large as they appear to be on the original data

Overall picture redundancy vs MSE

$$\text{relative increase (MSE)} = \frac{MSE_{\text{cleaned}}}{MSE_{\text{orig}}}$$



- The performance degrades the stronger (higher y-values), the more redundancy was contained in the original dataset (lower x-values).
- Different methods differ in their sensitivity to redundancy (higher potential to overfit).
  - *No sensitivity for simpler methods (offset), high sensitivity for complex methods (hidden factors)*

Conclusions and Outlook

- We evaluated the impact of redundancy on the evaluation of recommendation models
  - Conclusions may change when tested on cleaned data
  - This effect is stronger for more complex methods
  - Our results suggest that the collection should be used only after duplicate elimination
  - Our results do not suggest that the recent methods do not imply real methodological improvements.
- So what?
  - The quality of the data is critical for learning
  - Can we automatically control datasets for potential problems?
  - The community would benefit from having established benchmarks for different tasks (common practice in other communities like IR, NLP)
  - Evaluation and competitive comparison of existing methods is equally important to the development of new methods.

Daniel Basaran[1], Eirini Ntoutsi[2], Arthur Zimek[3]

[1]Ludwig-Maximilians-Universität, Germany [2]Leibniz Universität Hannover, Germany [3]University of Southern Denmark

2017 SIAM International Conference on DATA MINING
April 27 - 29, 2017
The Westin Galleria Houston Houston, Texas, USA