Effects of Human-curated Content on Diversity in PSM: ARD-M Dataset

Marcel Hauck[1,2,*], Ahtsham Manzoor[3] and Sven Pagel[4]

[1] Johannes Gutenberg University Mainz, Jakob-Welder-Weg 9, 55128 Mainz, Germany
[2] ARD Online, Isaac-Fulda-Allee 1, 55124 Mainz, Germany
[3] University of Klagenfurt, P.O. Box 1212, Klagenfurt, Austria, 43017-6221
[4] Mainz University of Applied Sciences, Lucy-Hillebrand-Str. 2, Mainz, Germany, 55128

Abstract

Public service media (PSM) providers, like *ARD*, are continuously undergoing digital disruptions due to the ubiquitous availability of the internet and technological breakthroughs. To keep up with such advancements, PSM providers face new challenges, particularly in adopting artificial intelligence technologies such as recommender systems. However, due to the heterogeneous nature of the content, the unavailability of PSM datasets, and human involvement in content curation and decision-making, it remains unclear how internal domain knowledge experts (e.g., publishers and editors) can be bridged with external technologists. This poses further challenges in creating recommendation technology for PSM providers that automatically deliver relevant, transparent, and diverse content for their consumers. To this end, real-world datasets can be pivotal in investigating and closing this information gap. In this work, therefore, we release a real-world dataset for researchers containing features like items and their metadata, user interaction feedback, and complementary features like item position, timestamps, and item categories, in the movie domain. Furthermore, through a series of analyzes, we show how different data sources (publisher, editors, TMDB) and metadata features (genres, keywords, plot) impact the diversity of items. Finally, we present an outlook of future directions to be explored by the research community using the ARD-M dataset. This is an important step forward to strengthen *relevance*, *transparency* and *diversity* in the recommendation process, as part of the PSM core values.

Keywords

Public service media providers, recommender systems, dataset, user interaction features, diversity

1. Introduction

As of today, over 90%[1] of the population in Western countries are internet users (e.g., 92% in the USA). Moreover, around 90%[2] of them are also smartphone users (e.g., 89% in the USA). Such ubiquitous advancements have led to the digital transformation of internet services such as e-commerce or streaming of video content. Additionally, advancements in technology, particularly in the media industry, have largely replaced editorial recommendations with algorithmic content recommendations [1, 2, 3]. A prominent outcome of recent transformations is an across-the-board shift towards consuming media content predominantly through online channels.

In this regard, Public Service Media (PSM) platforms, like *ARD Mediathek*, are still striving to incorporate recommendation technology and maintaining market shares in competition with Netflix, Amazon Prime, and other online streaming providers. One reason could be that PSM platforms provide a variety of content, for example, daily news, stories, movies, sports, or cultural shows. Thus, exercising recommendation technology for such mixed content delivery requires modern yet holistic solutions. Therefore, PSM providers traditionally rely on editors as domain experts for delivering the content [4]. On the one hand, editors play a key role in shaping content delivery procedures by providing feedback and implementing policies that align with legal and ethical mandates and guidelines [5]. Furthermore, algorithmic content selection and user personalization can introduce risks and societal threats (e.g., misinformation, discrimination, bias, and privacy issues), see also [6].

PSM providers are generally committed to asserting core values such as relevance, transparency, and diversity while ensuring public trust. To this end, editors' domain knowledge and role in ensuring such values while content curation and selection offer a great deal to retrospect how recommender systems are designed. For instance, digitizing media processes has created numerous opportunities to collect and analyze vast amounts of audience and consumption data. This data can be leveraged to customize services and content based on the perceived interests of individual consumers while adapting *human-in-the-loop* design practices. For this reason, the need

Workshop on Learning and Evaluating Recommendations with Impressions (LERI) @ RecSys 2023, September 18-22 2023, Singapore

[*] Corresponding author.

mhauck@uni-mainz.de (M. Hauck); ahtsham.manzoor@aau.at (A. Manzoor); sven.pagel@hs-mainz.de (S. Pagel)
0000-0003-1105-1519 (M. Hauck); 0000-0001-9418-7539 (A. Manzoor)

[1] https://datareportal.com/reports/digital-2023-july-global-statshot
[2] https://www.gsma.com/mobileeconomy/

Table 1
Descriptive statistics of dataset

Unique lists (categories)	Unique items	Unique users	Unique sessions	Total item views	Total item clicks
20	340	294.001	487.766	42.849.058	818.238

for externally available real-world industry datasets is crucial for the research community to make practical use of domain experts' knowledge in recommender systems.

In addition, *relevance*, *transparency*, and *diversity* of item recommendations are highly vigilant topics in recommender systems research, see e.g., [7, 8, 9]. Relevance refers to the degree to which recommended items align with the user's preferences, needs, or interests, see also [10, 11]. Research has shown that by offering relevant recommendations, these systems can enhance user satisfaction and engagement and ultimately drive conversion rates [12]. Moreover, diversity plays an equally important role in recommendation algorithms. It refers to the variety and heterogeneity of items suggested to users preventing over-specialization and filter bubbles, where users are only exposed to a limited set of items [9]. Also, diversity helps to address the recommendation bias issue [13]. Therefore, achieving an optimal balance between relevance and diversity is crucial for delivering high-quality recommendations in PSM platforms that cater to users' diverse interests while providing exposure to multifarious and potentially serendipitous content.

However, given the lack of real-world, human-curated PSM datasets, it is unclear how internal editorial domain knowledge and editors' decisions can be bridged with recommender systems from external technologists serving users' needs while achieving market share and re-growth.

In this work, therefore, we first collected a real-world dataset in the movie domain containing items, their metadata, and users' interaction feedback records, i.e., item impressions and item clicks. Subsequently, using TMDB[3] database, we enriched the dataset with additional metadata features like movie plots and genres. Second, with the help of a series of analyzes, we show how item diversity and similarity vary in intra-list navigational interfaces over the period of one week. Finally, we present an outlook of future directions for the recommender systems research community and release the ARD-M dataset online under CC BY-NC license.

2. Dataset Collection

ARD is a PSM provider in Germany, which offers a variety of online streaming content, including movies[4]. On the movie page, items are ordered within horizontal lists, each titled with a category such as "*Newly available films*" or "*Movies for the whole family*". So far, mainly human editors are responsible for making decisions regarding the specific position and listing of an item in a specific list. Specifically, given a large pool of content provided by publishers in regional broadcasters (so-called "Landesrundfunkanstalten"), editors select and integrate a specific item in a particular list along with curated mixed-quality metadata features like keywords or genres. With the goal of building a bridge between algorithmic and human curation of content in PSM platforms, we collected data for the first *whole* week of January 2023.

Apart from items and their metadata (genres, plots, and keywords), we provide two user interaction features, i.e., item views and item clicks. For simplicity, we represent item impressions as item views, whereas it remains an open question whether such an assumption is plausible in the context of an online streaming platform. In addition, we provide complementary features like *timestamp*, *item position* in a particular *list.* The descriptive statistics of our dataset are shown in Table 1. The numbers of distinct values for each item metadata feature are in Table 2. For both user interaction features, we show the data characteristics like sparsity and density in Table 3. We believe a rich user feedback dataset will open new directions to research recommendation technology for PSM providers by external technologists.

To protect user privacy, we utilized random numbers as visitor and session identifiers to ensure decoupling from production data. To ensure data protection for rare events (e.g., usage in the middle of the night), we rounded timestamps to the closest minutes. Finally, to enable cross-source exploratory analysis, we enriched the data with a summary of the plots, keywords, and genres from the community-built TMDB[5] movies database. With a variety of metadata collected from multiple sources, i.e., publishers, editors, and TMDB, we believe new algorithmic dimensions of content or collaborative filtering-based approaches to recommendations can be explored. Moreover, with a humans-in-the-loop approach, domain experts' knowledge can be leveraged in various ways, for example, by investigating content categorization, user behavior, and item diversity and similarity in lists. We release our dataset publicly at https://www.kaggle.com/datasets/marcelhauck/ardmovies/ with kind permission of ARD Online and its Data Protection Officer.

[3] https://www.themoviedb.org/
[4] https://www.ardmediathek.de/filme
[5] https://www.themoviedb.org/

Table 2
Number of distinct values for item metadata features in dataset

Plot		Genres			Keywords	
Publisher	TMDB	Publisher	Editorial	TMDB	Publisher	Editorial
340	272	112	19	120	258	311

Table 3
Sparsity and density of interaction features in dataset

Views		Clicks	
Sparsity	Density	Sparsity	Density
57.1%	42.9%	99.2%	0.8%

3. Comparison with Existing Datasets

In this section, we briefly discuss how our ARD-M dataset compares to existing predominant datasets available in the literature on recommender systems. Overall, we observe a number of datasets available in the literature to carry out research on recommender systems. For example, the predominant tendency is to *design* recommender algorithms and *evaluate* them using one or more datasets. Technically, a recommender system can serve multiple user objectives such as providing relevant item recommendations, endorsing transparency via explanations, or ensuring diversity in the recommendation process. Recommender system designers thereby leverage datasets like ARD-M to attain and assess such objectives by implementing specific computational tasks like Next item-recommendation, Next session prediction, Top-N recommendations, or Explanation generation. Overall, as explained in Section 2, our ARD-M dataset includes various features like items, users, metadata (genres, keywords, plots), user interaction feedback (impressions and clicks), and complementary features (list IDs, item positions, timestamps, and item duration).

To this end, our ARD-M dataset resembles various datasets in terms of available features and therefore can serve multiple objectives in the recommendation process. For example, e-commerce datasets like YOOCHOOSE, used in [14], DIGINETICA [15], and Retailrocket [16] offer similar user interaction features, and are extensively used, specifically, for collaborative filtering or session-based recommendations. In the movies domain, MovieLens 10M [17, 18], Douban [19], and Netflix [20, 21] datasets share common features like categories, keywords, and timestamps, yet include additional user interaction feedback, i.e., user ratings. Similarly, in the movies domain, unlike ARD-M, datasets like Yahoo! Movies [22], EachMovie [19] and FilmTrust [23] include complementary user features like gender, age, and trust network, which can be further used to design context-aware recommendations. However, a social network of users sharing common preferences or interests can be created with the ARD-M dataset as well, e.g., by including consumed items, and their metadata features like keywords, or genres.

There are additional datasets that with impressions on different domains. The “Microsoft News Dataset (MIND)” contains data from the Microsoft News website with interactions (impressions, clicks) and metadata features like title, body, and category [24]. The “ContentWise Impressions” dataset includes interactions of movies and TV series that are collected from a recommender system on an Over-The-Top media service [25]. The “FINN.no slate dataset” contains data from the real estate marketplace FINN.no with sequential interaction features on recommendations and search results in lists [26].

In addition, our ARD-M dataset offers a variety of features, which can be further used to implement specific computational tasks to achieve certain recommender system objectives. For example, in the context of horizontal list interfaces like in the case of commercial online streaming platforms such as Netflix, or Amazon Prime Video, we provide specific item positions in a specific list. Such fine-grained details can be modeled in the recommender algorithm for investigating aspects, for example, CTR optimization with respect to item positions, user behavior analysis, and design optimization of different user interfaces. Furthermore, our dataset is exceptionally dense (42.9%) in terms of user impressions, presenting various ways to aid personalization, user segmentation, or comparing different recommendation strategies. Finally, unlike other datasets, we integrate additional metadata obtained through multiple sources, thereby providing opportunities to conduct metadata-based content modeling.

To the best of our knowledge, there is no dataset curated from a PSM platform, where decisions regarding diverse content selection are made by domain experts. Thus, with a humans-in-the-loop approach, we believe ARD-M extends substantial opportunities to investigate novel recommender system use cases.

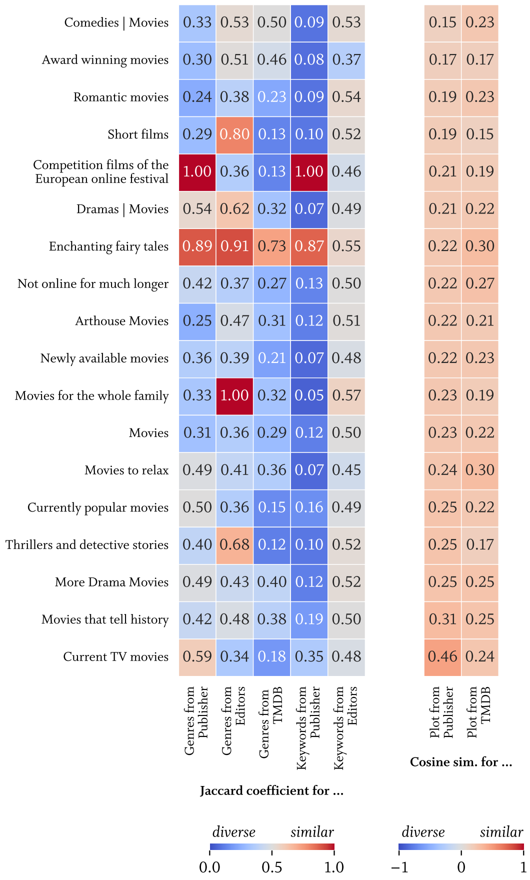

Figure 1: ILS for all lists by their metadata features

4. Intra-list Similarity (ILS) Analysis

With a goal of understanding the diversity and similarity of items curated by human editors, inspired by work in [27], we compute the *intra-list similarity* (ILS) score. Diversity can be seen as the opposite of similarity. Specifically, we compute the ILS for each pair of items in a list with the available metadata features, e.g., plots and genres. We separately use the *Jaccard Coefficient* [28] for similarity calculation of genres and keywords from all data sources (i.e., publishers, editors, and TMDB). For the Jaccard Coefficient, a value of 0 shows low similarity (= high diversity), and 1 as high similarity (= low diversity). For plots, we first produced embeddings using a Sentence Transformer [29] from Hugging Face[6] and computed cosine similarity score for plots between each pair of items in a list. For the cosine similarity score, a value of -1 shows low similarity (= high diversity), and 1 as high similarity (= low diversity).

Figure 1 shows the ILS for every list in the dataset based on Jaccard coefficient (left part) and cosine similarity (right part). Editors frequently removed or added items to those lists, leading to a changed ILS. Therefore, the values in the figure are the average in the complete timespan. As can be seen, the individual metadata has higher (e.g., Keywords from Editors) or lower consistency (e.g., Keywords from Publishers). Editors select content for lists so that they have a high semantic similarity. Therefore, this is also reflected in the consistent similarity based on their curated keywords. The situation is different for the numerous regional broadcasters with human publishers, who generate and upload keywords in different ways (e.g., use of default values or structured keyword hierarchies). The average ILS scores for all metadata features are shown in Table 4. The results suggest that editorial features based on Jaccard coefficient (genres, keywords) with a value range from 0 (diverse) to 1 (similar) are reasonably better represented (genres $= 0.52$, keywords $= 0.50$) for the semantic connection between items in a list than genres (0.30) from TMDB and keywords from publishers (0.21). As described before, the ILS scores based on the movie plot are calculated using the pairwise cosine similarities with a value range from -1 (diverse) to 1 (similar). Therefore, the results (avg. ≈ 0.22) reveal a slightly similar content representation.

Next, we conduct a correlation analysis to investigate the relationships of content diversity between all metadata sources (publishers, editors, TMDB); see results in Figure 2. Specifically, first, we compute the average ILS score for each item list and for all metadata features. Subsequently, the average ILS scores are used for computing the correlation. Results indicate that publishers' metadata features (keywords and genres) correlate strongly (0.91). A similar case is found in TMDB features (genres and plot $= 0.53$) and editorial features (keywords and genres $= 0.49$), even when their average ILS scores are different, see also Table 4. This implies that editors can rely on additional data curation sources like publishers with substantial domain knowledge and expertise while selecting the content for a particular list. Also, different metadata sources can be uniquely combined in a recommendation model to assert values like the diversity of items with a humans-in-the-loop design approach. We present additional analyzes, e.g., the variation of the ILS scores for all categories in time, in our Kaggle data repository at https://www.kaggle.com/code/marcelhauck/data-analyses.

[6] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

Table 4
Average ILS for all lists per feature
Note: Value range for plot is from −1 to 1 (Cosine similarity) and for genres and keywords from 0 to 1 (Jaccard coefficient)

Plot		Genres			Keywords	
Publisher	TMDB	Publisher	Editorial	TMDB	Publisher	Editorial
0.23	0.22	0.45	0.52	0.30	0.21	0.50

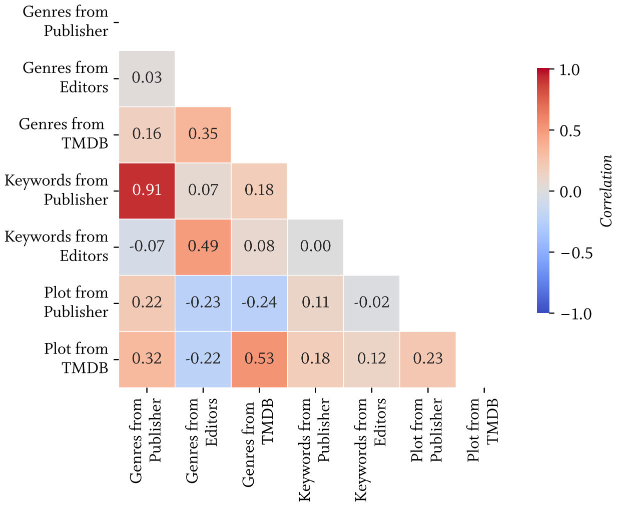

Figure 2: Average ILS correlation for metadata features

5. Conclusion and Future Research Directions

Public Service Media (PSM) providers have been continuously adapting to technological developments and mostly provide digital services for their consumers. However, the integration of external recommendation technology for heterogeneous content presents further challenges in incorporating internal domain experts' knowledge to assert core values such as content diversity while projecting business and consumer value. To this end, real-world PSM datasets can be pivotal to fostering research in this domain. Therefore, in this work, we release a real-world ARD-M dataset created from *ARD Mediathek* and analyze the content regarding diversity and similarity aspects.

Finally, we highlight a few directions for the research community to build on the ideas. First, datasets like ours can be used to tailor services and content through automated recommendations to the perceived interests of individual consumers. Also, understanding users' behavior and their experiences with consumption data can further help in meaningfully devising ways for publishers to create content that interests their audience. Second, PSM providers are looking for ways to create common yet holistic evaluation metrics that illustrate algorithmic performance, the relevance of content, and business value for both providers and consumers. We believe our released dataset can pave the way for researching and devising such novel metrics in the future. Finally, in PSM providers, editors possess domain expertise relating to content curation and selection while adhering to PSM responsibilities like educative and unbiased content delivery. Therefore, it is largely unclear without popularity bias content features how public-service values can be protected and infused in recommendation algorithms development. We believe this work opens new avenues of innovations and collaborations between PSM actors like editors, researchers, and recommender system designers.

References

[1] M. V. Álvarez, J. M. T. López, M. J. U. Ruíz, What are you offering?: An overview of vods and recommender systems in european public service media, Information Technology and Systems: Proceedings of ICITS 2020 (2020) 725–732.

[2] J. Hildén, The public service approach to recommender systems: Filtering to cultivate, Television & New Media 23 (2022) 777–796.

[3] J. K. Sørensen, J. Hutchinson, Algorithms and public service media, in: Public Service Media in the Networked Society: RIPE@ 2017, Nordicom, 2018, pp. 91–106.

[4] N. Herm-Stapelberg, F. Rothlauf, The crowd against the few: Measuring the impact of expert recommendations, Decision Support Systems 138 (2020) 113345.

[5] A. Grün, X. Neufeld, Challenges experienced in public service media recommendation systems, in: Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 541–544.

[6] C. Trattner, D. Jannach, E. Motta, I. Costera Meijer, N. Diakopoulos, M. Elahi, A. L. Opdahl, B. Tessem, N. Borch, M. Fjeld, et al., Responsible media technology and ai: challenges and research directions, AI and Ethics 2 (2022) 585–594.

[7] M. Kunaver, T. Požrl, Diversity in recommender systems–a survey, Knowledge-based systems 123 (2017) 154–162.

[8] S. Vargas, L. Baltrunas, A. Karatzoglou, P. Castells, Coverage, redundancy and size-awareness in genre diversity for recommender systems, in: Proceedings of the 8th ACM Conference on Recommender systems, 2014, pp. 209–216.

[9] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, 2011, pp. 109–116.

[10] D. Kotkov, J. Veijalainen, S. Wang, Challenges of serendipity in recommender systems, in: International conference on web information systems and technologies, SCITEPRESS, 2016.

[11] D. Kotkov, S. Wang, J. Veijalainen, A survey of serendipity in recommender systems, Knowledge-Based Systems 111 (2016) 180–192.

[12] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma, How good your recommender system is? a survey on evaluations in recommendation, International Journal of Machine Learning and Cybernetics 10 (2019) 813–831.

[13] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, ACM Transactions on Information Systems 41 (2023) 1–39.

[14] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks (2015). arXiv:1511.06939.

[15] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, J. Ma, Neural attentive session-based recommendation, in: CIKM '17, 2017, pp. 1419–1428.

[16] C. Gao, K. Xu, K. Zhou, L. Li, X. Wang, B. Yuan, P. Zhao, Value penalized q-learning for recommender systems, in: SIGIR '22, 2022, pp. 2008–2012.

[17] A. Goyal, L. V. Lakshmanan, RecMax: Exploiting recommender systems for fun and profit, in: KDD '12, 2012, pp. 1294–1302.

[18] F. Zhuang, D. Luo, N. J. Yuan, X. Xie, Q. He, Representation learning with pair-wise constraints for collaborative ranking, in: WSDM '17, 2017, pp. 567–575.

[19] R. v. d. Berg, T. N. Kipf, M. Welling, Graph convolutional matrix completion (2017). arXiv:1706.02263.

[20] S. Kabbur, X. Ning, G. Karypis, Fism: factored item similarity models for top-n recommender systems, in: KDD '13, 2013, pp. 659–667.

[21] F. Khawar, L. Poon, N. L. Zhang, Learning the structure of auto-encoding recommenders, in: WWW '20, 2020, pp. 519–529.

[22] L. Zheng, F. Zhu, S. Huang, J. Xie, Context neighbor recommender: Integrating contexts via neighbors for recommendations, Information Sciences 414 (2017) 1–18.

[23] J. Bobadilla, S. Alonso, A. Hernando, Deep learning architecture for collaborative filtering recommender systems, Applied Sciences 10 (2020) 2441.

[24] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, et al., Mind: A large-scale dataset for news recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3597–3606.

[25] F. B. Pérez Maurera, M. Ferrari Dacrema, L. Saule, M. Scriminaci, P. Cremonesi, Contentwise impressions: An industrial dataset with impressions included, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3093–3100.

[26] S. Eide, D. S. Leslie, A. Frigessi, J. Rishaug, H. Jenssen, S. Verrewaere, Finn. no slates dataset: A new sequential dataset logging interactions, all viewed items and click responses/no-click for recommender systems research, in: Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 556–558.

[27] M. Jesse, C. Bauer, D. Jannach, Intra-list similarity and human diversity perceptions of recommendations: the details matter, User Modeling and User-Adapted Interaction (2022) 1–34.

[28] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of jaccard coefficient for keywords similarity, in: Proceedings of the international multiconference of engineers and computer scientists, volume 1, 2013, pp. 380–384.

[29] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.