Exploring The Semantic Gap for Movie Recommendation

Conference Paper · September 2017

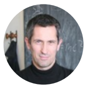
DOI: 10.1145/3109859.3109908

Mehdi Elahi

Libera Università di Bozen-Bolzano

Yashar Deldjoo

Politecnico di Milano



Paolo Cremonesi

Politecnico di Milano

Exploring The Semantic Gap for Movie Recommendations

Mehdi Elahi
Free University of Bozen - Bolzano
Piazza Domenicani 3
Bozen - Bolzano 39100, Italy
meelahi@unibz.it

Yashar Deldjoo
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
yashar.deldjoo@polimi.it

Farshad Bakhshandegan Moghaddam
Karlsruhe Institute of Technology
Karlsruhe 76131, Germany
ol5379@kit.edu

Leonardo Cella
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
leonardo.cella@mail.polimi.it

Stefano Cereda
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
stefano.Cereda@mail.polimi.it

Paolo Cremonesi
Politecnico di Milano
Via Ponzio 34/5
Milan 20133, Italy
paolo.cremonesi@polimi.it

ABSTRACT

In the last years, there has been much attention given to the *semantic gap* problem in multimedia retrieval systems. Much effort has been devoted to bridge this gap by building tools for the extraction of high-level, semantics-based features from multimedia content, as low-level features are not considered useful because they deal primarily with representing the perceived content rather than the semantics of it.

In this paper, we explore a different point of view by leveraging the gap between low-level and high-level features. We experiment with a recent approach for movie recommendation that extract low-level Mise-en-Scène features from multimedia content and combine it with high-level features provided by the wisdom of the crowd.

To this end, we first performed an offline performance assessment by implementing a pure content-based recommender system with three different versions of the same algorithm, respectively based on (i) conventional movie attributes, (ii) mise-en-scène features, and (iii) a hybrid method that interleaves recommendations based on movie attributes and mise-en-scène features. In a second study, we designed an empirical study involving 100 subjects and collected data regarding the quality perceived by the users. Results from both studies show that the introduction of mise-en-scène features in conjunction with traditional movie attributes improves both offline and online quality of recommendations.

CCS CONCEPTS

•Information systems → Recommender systems;

KEYWORDS

content-based video recommendation, semantic gap, online user study, offline simulation, mise-en-scene, low-level visual features

1 INTRODUCTION AND CONTEXT

In the last years, we have seen much attention given to the semantic gap problem in multimedia retrieval systems [20, 24]. The semantic gap is the gap between the high-level concepts that users expect when searching for interesting multimedia content (*e.g.*, genre, plot, actors) and the low-level features that it is possible to automatically extract from the same content (*e.g.*, brightness, contrast, *etc.*). The users' ability to comfortably and efficiently use multimedia systems is impacted as the result of the sharp discontinuity that exists between the primitive low-level features and the richness of user queries encountered in multimedia search. As a result, the multimedia information retrieval community has long struggled to *bridge* this semantic gap as several studies have confirmed the difficulty of addressing users' information needs with such low-level features, because they deal primarily with representing the perceived content rather than the semantics of it [14, 20].

Because of this gap, most movie recommender systems rely on the assumption that user preferences are mainly influenced by the high-level *semantic* features of movies (*e.g.*, plot, genre, director, actors) and, to a less extent, by the *stylistic* properties [10] (*e.g.*, color, motion, lighting key). Recommendations are automatically computed using implicit or explicit preferences of users on these attributes [1, 8]. However, recent works on RSs suggest that users' preferences when choosing a product are influenced more by the *visual* aspect of items and less by their semantic or syntactic properties [11, 19, 26, 29].

In this paper, we make a different assumption. We still believe that the semantic gap exists and limits the adoption of low-level features in information retrieval applications, where the goal is to provide a way of indexing multimedia content so that users can *explicitly* (*i.e.*, manually) query that content at the semantic level. However, we wish to investigate if this assumption holds also for recommender systems, where the goal is for the system to *automatically* find content that the user likes, without the user querying the system.

Our research hypothesis is that the semantic gap is not a problem but an opportunity. Our goal is to leverage the low-level features automatically extracted from multimedia content and complement them with the high-level features. For this purpose, we explore and evaluate a recent approach for movie recommendations that integrates traditional high-level semantic attributes, such as genre,

RecSys'17,
978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

director and cast, with low-level mise-en-scène features, *i.e.*, the design aspects of movie making that influence aesthetic and style. Examples of mise-en-scène characteristics are lighting, colors, background, and movements.

We believe that mise-en-scène features can be used to make video recommender systems more effective and useful as they help in strengthening two weak spots when working with semantic attributes: lack of diversity and novelty in video recommendations [16]. Past works show evidence that the lack of diversity and novelty in traditional recommender systems happens because recommender algorithms are designed to recommend videos similar to the ones users liked in the past [3, 16, 25]. Recommendations lacking novelty and diversity have negative consequences on user satisfaction, even if recommendations perfectly match users' tastes [4, 31]. This paper builds on previous results identifying mise-en-scène features that potentially influence accuracy of recommendations from system-centric perspective.

In this work we seek to answer the following research questions: **RQ 1.** Can the introduction of low-level mise-en-scène features in video recommender systems, combined with high-level semantic features, improve offline quality of recommendations? **RQ 2.** Can the introduction of low-level mise-en-scène features in video recommender systems, combined with high-level semantic features, induce measurable effects on the perceived utility of recommendations?

To this end, we carried out two wide and articulated empirical studies: (a) a *system-centric* evaluation to measure the offline quality of recommendations in terms of precision, novelty, diversity and coverage. (b) a *user-centric* online experiment involving 100 users, measuring different subjective metrics (*i.e.*, relevance, novelty, diversity, and satisfaction). Both studies explore the effects of recommendations under three different experimental conditions defined by one manipulated variable: the type of movie attribute used for recommendations. The studies consider (i) a content-based algorithm based on semantic movie attributes, (ii) the same algorithm based on mise-en-scène features, and (iii) a hybrid algorithm that interleaves recommendations based on attributes and mise-en-scène features.

Results show that (1) the adoption of mise-en-scène features can strongly affect precision/relevance, diversity and novelty of recommendations, determining an increased utility of recommendations and influencing the user's decision to actually watch a movie, and (2) the introduction of mise-en-scène features in conjunction with traditional attributes, tends to diversify recommendations and suggest users with less obvious choices.

2 EXPERIMENT SETUP

Table 1 summarizes the experimental conditions used in two studies.

2.1 Recommendation Algorithm

In order to generate recommendations, in both offline and online experiments, we tested a widely used pure content-based filtering (CBF) algorithm based on k-nearest neighbors, and considered 20 neighbors, cosine similarity, and log-quantile normalization of the mise-en-scène features [11].

Table 1: Experimental conditions used in two studies

Study	Feature Types	Quality Metrics	#Users
A: Offline	(i) semantic (ii) mise-en-scène (iii) combination of semantic and mise-en-scène	Precision Diversity Novelty Coverage	113,682
B: Online		Relevance Diversity Novelty Satisfaction	100

2.2 Movie Features

In both studies, the quality of recommendations have been evaluated under three different experimental conditions defined by one manipulatable variable: the type of movie features.

2.2.1 Movie attributes. Content-based movie recommender systems traditionally base their recommendations on high-level attributes such as genre, director, and actors (movie attributes). Such metadata are human-generated, either editorially (*e.g.* title, cast) or by leveraging the wisdom of the crowd (*e.g.*, tags).

2.2.2 Mise-en-Scène features. Recently, low-level stylistic features – such as color, motion and lighting – have been proposed for content-based video recommendations (mise-en-scène features) [11]. Mise-en-Scène features can be extracted automatically by processing the video files. Different types of visual features have been explored in the community of multimedia recommender systems [19, 21, 29].

The general procedure to extract this five categories of features comprise four steps: (a) videos are segmented into shots and for each shot, a representative key frame is extracted (this is not necessary for the shot duration); (c) for each shot, the values of the features are computed; (d) the feature values are averaged over all shots.

In our experiments, we have selected five categories of mise-en-scène low-level features described in [11, 28] as they are explainable, easy to extract from video files, and show promising results in offline experiments: *shot duration*, *color variance*, *lighting key*, *average motion*, and *motion variation.* The use of keyframe-based feature representation has been motivated in [9][1].

2.3 Study A: Offline Experiment

In the offline experiment, we evaluated Top-N recommendation quality of each experimental condition under a pure CBF algorithm by running a hold-out (80% train - 20% test) of a subset of the MovieLens (ML-20M) dataset. The dataset contains 113,682 users who provided 775,090 preferences (*i.e.*, ratings) to 12,573 movies. Each movie is provided with a set of 127 attributes concerning, among the others: title, genre, cast, tags. On average, each movie is labeled with six attributes. The mise-en-scène dataset contains five feature values extracted according to the procedure described in Section

[1] The work is partially covered by a US patent application under the title of "Enhanced content based multimedia recommendation method", with application number 15/277490 [5]

2.2.2. In the offline evaluation we computed recommendation quality with respect to: *precision*, *diversity*, *novelty* and *coverage* [18].

We estimate precision by computing the percentage of movies in the recommendation list that were relevant to the user. To estimate diversity, we measure the intra-list similarity between items with respect to movie genres, *i.e.*, we compute pairwise cosine similarity between movies in the recommendation list and calculate the average similarity S. Diversity is later computed as the complement of the intra-list similarity $(1 - S)$. We compute novelty by computing the mean popularity rank of the items recommended to the user. Finally, coverage is computed by calculating the percentage of pairs of $< user, movie >$ for which we can predict a rating [2, 15, 17]. All metrics range from 0% to 100%, where 100% is the best. In our experiment, we have chosen $N = 4$ for the length of the list as it is desired to have a similar value for N in both offline and online experiments (see next section). The detailed procedure used to measure different recommendation quality metrics is similar to the one used in previous works [7, 11] maintaining compatibility with results published in other research papers.

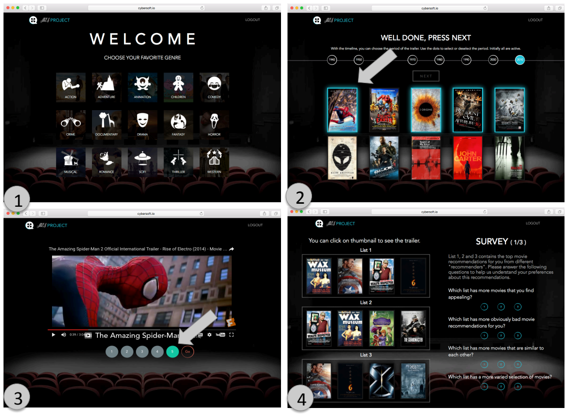
2.4 Study B: Online Evaluation

2.4.1 Perceived Quality Metrics. The goal of the second study is to measure the user's perceived quality of the recommender system. Perceived quality is the degree to which the users judge recommendations positively and appreciate the overall experience with the recommender system. To better scope our research, we operationalize this notion in terms of fours metrics: *perceived accuracy*, *novelty*, *diversity*, and overall *user satisfaction* as defined in [15, 22, 27]: *Perceived accuracy* (also called *Relevance*) measures how much the recommendations match the users' interests, preferences and tastes; *Diversity* measures how much users perceive recommendations as different from each other, *e.g.* movies from different genres; *Novelty* measures the extent to which users receive new recommended movies; *Overall Users' Satisfaction* measures the global users' feeling of the experience with the recommender system.

2.4.2 Procedure. For the purpose of our study, we have developed *MISRec*[2], a web-centric testing framework for the movie domain, which can be easily configured to facilitate the execution of controlled empirical studies. MISRec is powered by the same pure CBF algorithm described in Section 2.1 and supports users with a wide range of functionalities that are common in online video-streaming services such as Netflix and Lovefilm (Figure 1). MISRec contains the same catalog of movies used in the first study.

Our main research audience is represented by users aged between 20 and 50 who have some familiarity with the use of the web and had never used MISRec before the study (to control for the potentially confounding factor of biases or misconceptions derived from previous uses of the system). The total number of recruited subjects who completed the task was 100 (63% men, 37% women, average age: 27.8, std 3.83).

The interaction begins with a sign-up process, where each participant (user) is initially asked to provide some basic data. Afterward, users are invited to browse the movie catalog, and then, asked to freely select five movies and rate them using a 5-level Likert scale.



With the timeline, you can choose the period of the trailer. Use the dots to select or deselect the period. Initially all are active.

The Amazing Spider-Man 2 Official International Trailer - Rise of Electro (2014) - Movie ...

You can click on thumbnail to see the trailer.

List 1, 2 and 3 contains the top movie recommendations for you from different "recommenders". Please answer the following questions to help us understand your preferences about this recommendations.

Which list has more movies that you find appealing?

Which list has more obviously bad movie recommendations for you?

Which list has more movies that are similar to each other?

Which list has a more varied selection of movies?

Figure 1: Example screenshots of the online application

On the basis of these ratings, three recommendation lists are generated - one for each experimental conditions - each list containing 4 recommended movies. The user is asked to watch the trailers of the recommended movies within each of the 3 lists and reply to a set of questions related to the quality of the recommendations. Users are asked to indicate their answers to each of the questions by selecting one of the three lists. To avoid possible biases, the positions of the recommendation lists were randomized for each user. It is worth noting that since watching trailers is a time-consuming process, we decided to include 4 recommendations within each of the 3 lists.

A subset of the questionnaire as reported in [15, 22] was used to measure the perceived quality of recommendations: **Acc.** Which list better understand your taste in movies? **Div.** Which list has movies that match a wider variety of moods? **Novelty.** Which list has more movies that are familiar to you? **Satisfaction.** Which list would you be more likely to recommend to your friends?

We also introduced a number of equivalent questions but formulated in a different way, in order to check for inconsistency of answers.

3 EXPERIMENTS AND RESULTS

3.1 Experiment A: Offline Evaluation

Table 2 presents the offline quality of the recommendations for each of the three experimental conditions: *movie attribute*: conventional high-level semantic attributes of movies; *mise-en-scène*: low-level visual features; *hybrid*: a combination of mise-en-scène features and movie attributes. As for the last approach, the hybridization can be done at feature-level (*e.g.* in [12]) or ensemble-level. We chose the second approach for its ease of implementation. The ensemble-level hybridization in our system was implemented by interleaving the recommendation results based on movie attributes and mise-en-scène features.

The quality has been measured with the offline methodology as described in Section 2.3. Analysis of variance suggests that the three experimental conditions (traditional movie attributes, mise-en-scène features, and hybrid) have a significant impact ($p < 0.05$) on the four variables: precision, diversity, novelty, coverage.

[2] short for Mise-en-Scène Movie Recommender

The *precision* metric suggests that recommendations based solely on mise-en-scène features are less accurate, while both the hybrid approach and the traditional approach are the most accurate.

If we look at *diversity* and *coverage* of recommendations, both approaches based on mise-en-scène features (either alone or in combination with traditional attributes) provide the best recommendations, *i.e.*, the most diverse recommendations, able to span almost all of the items in the catalog.

In terms of *novelty*, the approach based solely on mise-en-scène features provides the best recommendations, with the hybrid approach being marginally better than the approach based on traditional attributes.

Table 2: Results of experiment A: Offline evaluation. Results in bold are significantly different ($p < 0.05$)

Research Variables	Movie attributes	Mise-en-Scène features	Hybrid
Precision	**16.9%**	6.3%	**16.1%**
Diversity	20.8%	**21.8%**	**21.9%**
Novelty	94.9%	**96.7%**	95.7%
Coverage	75.5%	**92.7%**	**92.7%**

3.2 Experiment B: Real User Study

We first polished the collected data by removing the ones referring to subjects who showed apparent evidences of gaming with the testing system. We removed the participants who interacted with the system for less than 2 minutes, or left some questions unanswered. We also introduced a number of equivalent questions formulated in a different way, in order to check for inconsistency of answers. In the final questionnaire participants were asked to choose the preferred recommendation list. We ran multiple pair-wise Cochran Q tests on the responses from the users, as it well fits to the characteristics of the collected data [30]. All tests were run considering significance level $\alpha = 0.05$.

Table 3: Results of the experiment B: Real User Study. Results in bold are significantly different ($p < 0.05$).

Research Variables	Movie attributes	Mise-en-Scène features	Hybrid
Relevance	25%	15%	**60%**
Diversity	25%	22%	**53%**
Novelty	21%	19%	**60%**
Satisfaction	21%	22%	**57%**

The final results, presented in Table 3, show that the adoption of mise-en-scène features alone decreases almost all of the perceived quality metrics with respect to traditional semantic attributes, although not significantly (for accuracy, diversity, and novelty). This result is partially in contrast with the previous study, in which novelty and diversity with mise-en-scène recommendations are significantly better than with traditional attributes. This could be explained by previous works suggesting that offline evaluations metrics are not always good predictors of the perceived quality of recommender systems [7, 23]. However, when hybridizing recommendations based on low-level mise-en-scène features and high-level semantic attributes, they are perceived as better along all metrics. ($p < 0.05$).

4 DISCUSSION

The internal validity of our study is supported by the accuracy of our research design and by the quality of study execution. We have carefully implemented a number of mechanisms to control the accuracy of the tasks' execution. In terms of external validity, the results of our study are limited to those participants and conditions used in our study. Moreover, most services available in the market provide a user experience very similar to the one used in our study, in terms of filtering criteria and information/navigation structures [6], and it is likely that replications of our study on other systems may lead to results consistent with our findings. Finally, the high overall number of testers (100) allows us to generalize our results to a wider population of users aged 20-50.

A finer grained analysis of the statistically relevant relationships among all the different variables offers a much more articulated picture of the results, which show apparently contrasting results. More specifically, the two studies (offline and online experiment) illustrate different pictures with respect to the impact of mise-en-scène features on novelty and diversity. Our explanation is that the low accuracy of visual-only recommendations negatively affects the user opinion on the other metrics. A possible interpretation of this result is to consider that previous studies confirmed a mismatch between offline and online quality of recommendations [7, 23].

5 CONCLUSIONS

This work represents a contribution to the research and practice in the design of recommender systems, for the specific domain of movie recommendations and, from a more general perspective, for video recommendations. Our research differs from previous work in this domain for a number of aspects:

- We designed an online movie recommender system which incorporate mise-en-scène features, to be used for evaluation of recommendations with real users. In contrast, previous works on mise-en-scène features are based on only offline experiment.
- We compare three different recommendation scenarios, one of which combines recommendations based on mise-en-scène features and semantic attributes. Previous works limit their analysis to mise-en-scène features alone.
- Our results on the online evaluation of recommender systems based on mise-en-scène features (either alone or combined with semantic attributes) are totally new for the movie domain.

The results of our work can be applied to other domains where the system recommends multimedia products, such as music (e.g., Spotify, and Pandora) and images (e.g., Instagram, and Facebooks). In future, we plan to extend the range of features extracted to include both visual and audio features, using features based on the MPEG-7 standard and DNNs [13].

ACKNOWLEDGMENTS

This work has been partially funded by Telecom Italia S.p.A., Services Innovation Department, Joint Open Lab S-Cube, Milano. The work has been also partially supported by the Amazon AWS Cloud Credits for Research program.

REFERENCES

[1] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender Systems*. Springer, 139–166.

[2] Charu C Aggarwal. 2016. *Recommender Systems*. Springer. 225–254 pages.

[3] Oscar Celma. 2010. Music recommendation. In *Music Recommendation and Discovery*. Springer, 43–85.

[4] Òscar Celma and Pedro Cano. 2008. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*. ACM, 5.

[5] Paolo Cremonesi, Mehdi Elahi, and Yashar Deldjoo. 2016. Enhanced content based multimedia recommendation method. (09 2016). http://www.polimi.it/index.php?id=6247&sel_brevetto=5093 US Patent 15/277490.

[6] Paolo Cremonesi, Mehdi Elahi, and Franca Garzotto. 2016. User interface patterns in recommendation-empowered content intensive multimedia applications. *Springer Multimedia Tools and Applications* (2016), 1–35.

[7] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. 2013. User-centric vs. system-centric evaluation of recommender systems. In *IFIP Conference on Human-Computer Interaction*. Springer, 334–351.

[8] Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2015. Semantics-Aware Content-Based Recommender Systems. In *Recommender Systems Handbook*. Springer, 119–159.

[9] Yashar Deldjoo, Paolo Cremonesi, Markus Schedl, and Massimo Quadrana. 2017. The effect of different video summarization models on the quality of video recommendation based on low-level visual. In *Content-Based Multimedia Indexing (CBMI), 2017 15th International Workshop on*. ACM.

[10] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, and Pietro Piazzolla. 2016. Recommending movies based on mise-en-scene design. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1540–1547.

[11] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-based Video Recommendation System based on Stylistic Visual Features. *Journal on Data Semantics* Special Issue on Recommender Systems (2016).

[12] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Farshad Bakhshandegan Moghaddam, and Andrea Luigi Edoardo Caielli. 2016. How to Combine Visual Features with Tags to Improve Movie Recommendation Accuracy?. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 34–45.

[13] Yashar Deldjoo, Massimo Quadrana, Mehdi Elahi, and Paolo Cremonesi. 2017. Using Mise-En-Scene Visual Features based on MPEG7 and Deep Learning for Movie Recommendation. *arXiv preprint arXiv:1704.06109* (2017).

[14] Chitra Dorai and Svetha Venkatesh. 2003. Bridging the semantic gap with computational media aesthetics. *IEEE multimedia* 10, 2 (2003), 15–17.

[15] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 161–168. DOI:http://dx.doi.org/10.1145/2645710.2645737

[16] Daniel Fleder and Kartik Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science* 55, 5 (2009), 697–712.

[17] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 257–260.

[18] Asela Gunawardana and Guy Shani. 2015. Evaluating recommender systems. In *Recommender Systems Handbook*. Springer, 265–308.

[19] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. Vista: A Visually, Socially, and Temporally-aware Model for Artistic Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 309–316. DOI:http://dx.doi.org/10.1145/2959100.2959152

[20] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. 2015. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 27–34.

[21] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. 2013. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 105–112.

[22] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[23] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123.

[24] Hao Ma, Jianke Zhu, Michael Rung-Tsong Lyu, and Irwin King. 2010. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia* 12, 5 (2010), 462–473.

[25] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 1097–1101.

[26] Pablo Messina, Vicente Dominquez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2017. Exploring Content-based Artwork Recommendation with Metadata and Visual Features. (2017).

[27] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 157–164.

[28] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. 2005. On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on* 15, 1 (2005), 52–64.

[29] Sujoy Roy and Sharath Chandra Guntuku. 2016. Latent Factor Representations for Cold-Start Video Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 99–106. DOI:http://dx.doi.org/10.1145/2959100.2959172

[30] David J Sheskin. 2003. *Handbook of parametric and nonparametric statistical procedures*. crc Press.

[31] Mi Zhang and Neil Hurly. 2009. Evaluating the diversity of top-n recommendations. In *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on*. IEEE, 457–460.