# State of the art on recommender systems evaluation

Chiara Catizone - Camilla Neri

February 2022

**Abstract**

The purpose of this document is to list and discuss previous case studies belonging to the field of multimedia recommending systems to investigate an appropriate evaluating approach for our application. In this study we concentrate on multiple characteristics of recommending systems and investigate how they are measured and compared at the state-of-the-art.

# Contents

| Recommendation-centric | User-centric |
|---|---|
| Correctness | Trustworthiness |
| Coverage | Novelty |
| Diversity | Serendipity |
| Recommender confidence | Utility |
| | Risk |
| **System-centric** | **Delivery-centric** |
| Robustness | Usability |
| Learning rate | User preference |
| Scalability | |
| Stability | |
| Privacy | |

Figure 1: Avazpour's categorization of dimensions

# 1 Introduction

The object of our evaluation is a multimedia recommender system based on an algorithm using low-level features similarity and human interaction for recommending new items. We are interested in evaluating users' experience with the system in terms of quality of results and general satisfaction. We are going to develop an application in order to gather behavioural data and understand user's preferences. We will measure the system both qualitatively and quantitatively to get a better understanding of the objective characteristics of the system and the way the user experience it.

As Avazpour et al. state in their 2014 study [4] , the multi-faceted characteristics of recommendation systems lead us to consider multiple dimensions for recommender evaluation. Just one dimension and metric for evaluating the wide variety of recommendation systems and application domains is far too simplistic to obtain a nuanced evaluation of the approach as applied to the domain.[1]In their guidelines they propose a set of metrics corresponding to sixteen dimensions divided into 4 macro-categories (fig. 1).

[1]cf. Avazpour et al. "Dimensions and Metrics for Evaluating Recommendation Systems", p. 3, 2014

The reference literature we have chosen for our study is listed and discussed below in the following chapters. This research consisted in two main phases: 1) a phase in which we looked for guidelines and/or frameworks applied on evaluation of multimedia recommending systems with a special eye on the arts domain, 2) a phase in which we retrieved real case studies applying the measurements described in the frameworks.

# 2 Frameworks

In our research we first looked at frameworks developed by other scholars for evaluating multimedia recommending systems. We found two frameworks [9] and [12] suitable for evaluating the application given the goals set in the introduction of this paper.

Both frameworks were based on behavioural theories such as **TRA** (Theory of Reasoned Action)—that claims that the attitudinal and normative factors influence behavioral intentions of the users, which in turn predicts actual behavior.—and **UTAUT** (Acceptance and Use of Technology)—that retains experience-related evaluative concepts (performance expectancy, effort expectancy, social influence, and facilitating conditions) as influencing factors of users behavioural intentions; alongside with models like **TAM** (Technology Acceptance Model)—explaining the user's attitude towards using a technology by the perceived usefulness and perceived ease of use of the system—and **SUMI** (Software Usability Measurement Inventory)—a psychometric evaluation model for measuring the system's quality from the user point of view.[2]

## 2.1 ResQue

**ResQue** (Recommender systems' Quality of user experience) framework —developed in 2010 by Pu and Chen [12]—measures the user's perceptions of a recommending system qualities and tries to predict users' behavioral intentions.[3] The psychometric questionnaire derived from the framework consists of 13 constructs and a total of 60 question items. Ideally, it is administered to evaluate an online system using a 5-point Likert scale —from "strongly disagree" (1) to "strongly agree" (5)—to characterize users' responses.

[2]cit. P. Pu, L. Chen, "A User-Centric Evaluation Framework of Recommender Systems", p.15-16, 2010

[3]cf. Ibid., p.15-16

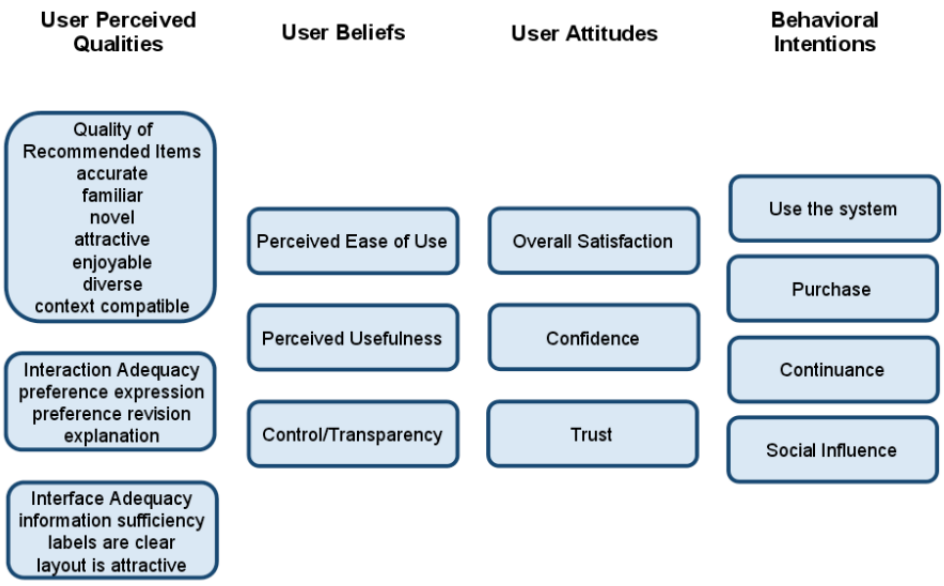


Figure 2: ResQue Framework

**ResQue essential constructs:**

1. User perceived qualities of the system
2. User beliefs as a result of these qualities in terms of ease of use, usefulness and control
3. User subjective attitudes
4. User behavioral intentions

### 2.1.1 Quality evaluation

To measure perceived quality of a recommending system Pu et al. [12] recognize the importance of running an holistic study on user's perceptions and impressions while interacting with a system. For this purpose they have created a taxonomy consisting in the following concepts to be measured by means of questionnaires submitted to users.

1. **Perceived accuracy**: is the degree to which users feel the recommendations match their interests and preferences.
2. **Familiarity**: describes whether or not users have previous knowledge of, or experience with, the items recommended to them. Reinforces **trust** and must be balanced with novelty.

3. **Novelty**: is the extent to which users receive new and interesting recommendations. It is often confused with serendipity by users, so it is good practice to merge the two metrics together in a single survey question.

4. **Attractiveness**: refers to whether or not recommended items are capable of stimulating users' imagination and evoking a positive emotion of interest or desire.

5. **Enjoyability**: refers to whether users have enjoyed experiencing the items suggested to them.

6. **Diversity**: measures how diverse are items in the recommendation list

7. **Context compatibility**: evaluates whether or not the recommendations consider general or personal context requirements.

#### 2.1.2 User Beliefs

User beliefs are based on user perceptions and their following elaboration on application's characteristics after completing an assigned task. This kind of insights can also be gathered through a survey and are useful for analysing which system's improvements needs to be implemented.

1. **Perceived ease of use** measures users' ability to accomplish tasks with ease and without frustration. This measure is useful especially if a system is entertaining and educational, and its interface and content is very appealing.

2. **Perceived usefulness** requests users' opinion on —whether or not—the system was useful to them in terms of task accomplishment.

    - **Decision support** measures the extent to which users felt assisted by the recommended system.
    - **Decision quality** can be assessed by confidence criterion, which is the level of a user's certainty in believing that he/she has made a correct choice with the assistance of a recommender.

3. **Control/transparency**

    - **User control** includes the system's ability to allow users to revise their preferences, to customize received recommendations, and to request a new set of recommendations.
    - **Transparency** determines whether or not a system allows users to understand its inner logic; e.g., a recommending system can convey its inner logic to the user through an explanation interface.

### 2.1.3 User Attitudes

By the term "attitude" Pu et al. [12] refer to the user's overall feeling towards a recommender system(derived from a previous interaction with it). This kind of feeling is generally believed to be more long-lasting than a belief, as it is highly influential on their subsequent behavioral intentions. Many researchers attribute positive attitudes, including users' satisfaction and trust of a recommender, as important factors.

1. **Overall Satisfaction** determines what users think and feel while using a recommender system.
2. **Confidence** is the recommender's ability to inspire confidence in users.
3. **Trust** indicates whether or not users find the whole system trustworthy (e.g., associated with their intentions to purchase and return to the website). Trust is influenced by the system's reputation and ability to predict good and well explained recommendations.

### 2.1.4 Behavioral Intentions

Behavioral intentions towards a system is related to whether or not the system is able to influence users' decision to use the system and purchase some of the recommended results.

1. **Continuance/Frequency** user agreement to use the system.
2. **Purchase**user acceptance of the recommended items (resulting in a purchase).
3. **Social Influence** user retention and intention to introduce this system to her/his friends.

## 2.2 Knijnenburg

**Kinjensburg et al.** [9] notice in previous literature a gap between the objective qualities of a system and the subjective evaluation of the system by the user. In order to fill this gap they expand ResQue's framework to context characteristics by means of personal characteristics of a user and the situational ones, in which the system is used.[4] This framework uses specific terms as *experience* and *interaction*, to distinguish between attitude and behaviour.[5]
The theory supporting Knjiniensburg's work is based on the concept of mediation between the objective qualities of the system (OSA) and the observable behaviour of the user (INT) by means of the subjective element of the perception of the user(SSA) and evaluation of the system(EXP).
OSA can be used in an experimental setting as in independent variable to study system's effects on the other aspects—EXP, SSA, INT.

1. **(INT) interaction:** the observable behavior of the user.
2. **(OSA) objective system aspects:** for example the algorithm used by the system, the visual and interaction design of the system, the way it presents the recommendation.
3. **(SSA) subjective system aspects** are expected to mediate the influence of the objective system aspects on the user experience they are usually measured with questionnaires and includes both pragmatic characteristics (usability and quality) and hedonic characteristics.
4. **(EXP) user experience** (users' evaluations of their interaction with the system) measured with questionnaires, includes their evaluation of the decision process (process-EXP) and the evaluation of the final decisions made (outcome-EXP)
5. **(PC and SC)** context of the interaction in terms of personal and situational characteristics (dependent on the context of the interaction:different points in time, different choice goals, trust and privacy concerns, and familiarity with the system)

In the original paper, Knijnensburg differentiates between two phases —i.e., Pre-trial and Trial —going beyond the purposes of our research. Hence, we decided to summarize and keep in this State-of-the-Art only the Trials dimensions' list, as it introduces two key-point elements to their inquiry: effort and fun.

[4] cf. Knijnensburg et al., "A User-Centric Evaluation Framework of Recommender Systems", p. 452, 2012
[5] cf. Ibid., p. 447

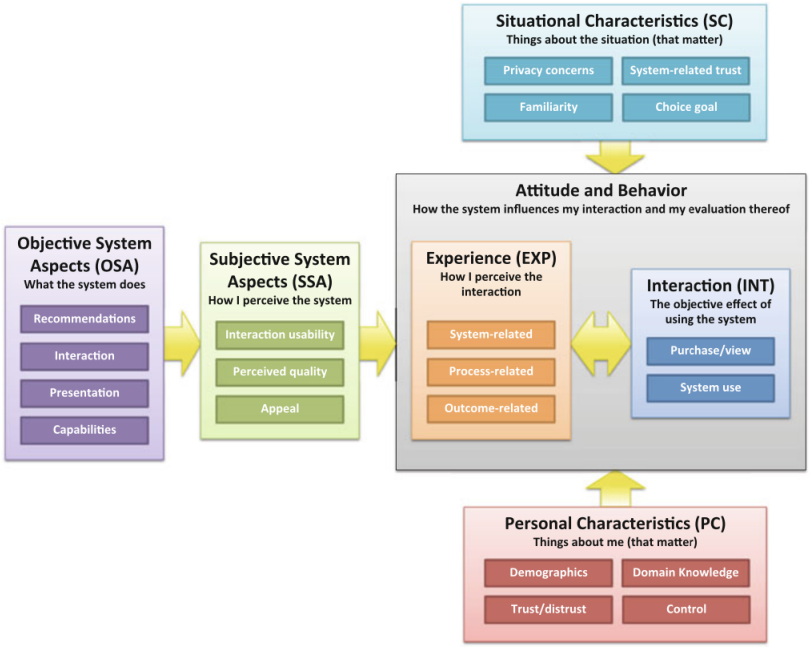


Figure 3: Knijnenburg's Framework

### Dimensions

1. Perceived recommendation quality
2. Perceived recommendation variety
3. Effort to use the system
4. Perceived system effectiveness and fun
5. Choice difficulty
6. Choice satisfaction
7. Intention to provide feedback
8. General trust in technology
9. System-specific privacy concern

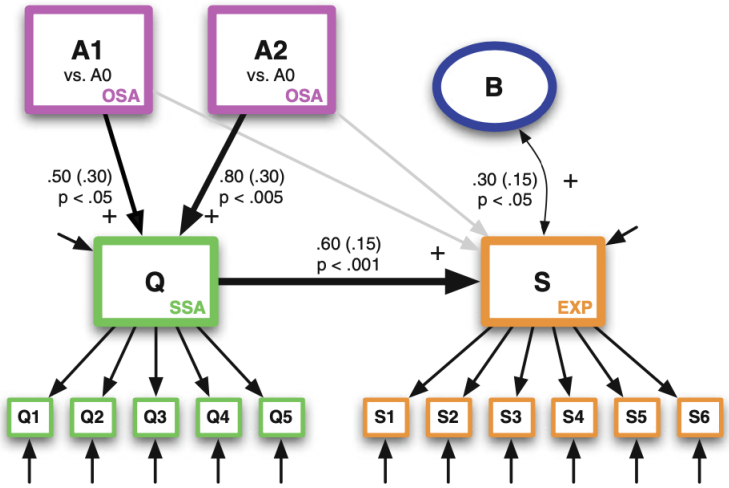


Figure 4: Representation of the structural equation modeling example. The manipulations (A1 and A2) influence the perceived recommendation quality (Q), which in turn influences the satisfaction with the system (S). The satisfaction (S) is in turn correlated with the number of clips watched from beginning to end (B). Q is measured by (Q1 . . . Q5) and S is measured by (S1 . . . S6). In the models in the main text the questionnaire items (Q1 . . . S6 ) are hidden in order to get a less cluttered representation.

### 2.2.1 Triangulation of data

Logged behavioral data can be triangulated with the user experience concepts to link the subjective experience (EXP) to objective user behavior (INT). This creates a causal chain of effects from manipulated OSAs, via subjectively measured SSAs and EXPs, to objectively measured INTs.

The framework also comes with a methodology for validating the latent concepts and testing the correlations between manipulations, latent concepts and behavioural measurements by means of the software Mplus:

1. **Exploratory Factor Analysis (EFA):** this analysis confirms whether the selected items measure the predicted latent concepts. It consists in extracting common variance between measured items and distributes this variance over a number of latent factors (fig.4) .[6]

2. **Structural Equation Modelling (SEM):** is used to test relations between manipulated features (A), latent concepts—e.g., quality (Q) and satisfaction(S)—and the behavior measurement (B).[7]

[6] cf.Ibid., p. 489
[7] cf. Ibid., p. 490

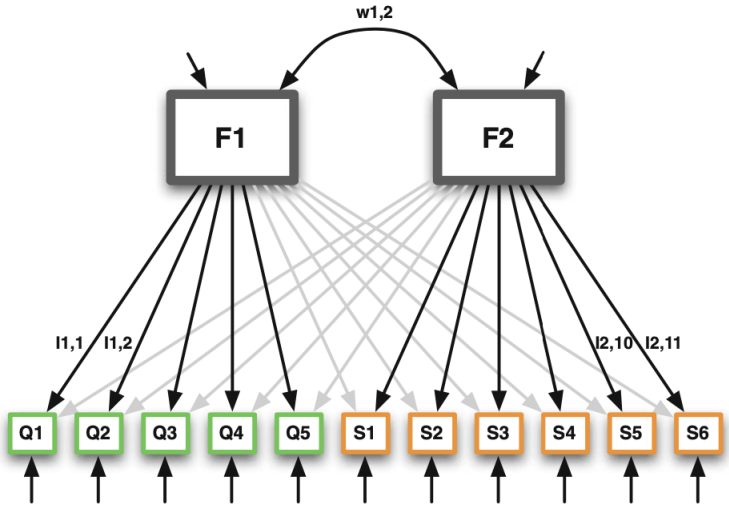


Figure 5: Representation of the exploratory factor analysis example. Two factors, F1 and F2, are extracted from the questionnaire data (items Q1. . .Q5 and S1 . . . S6). The arrows from the factors to the questions ( I1,1 . . . I2,11 ) represent the factor loadings. The arrows at the bottom represent the portion of the variance not accounted for by the analysis. Factors have a certain reliability (represented by the arrows at the top), and may be correlated (represented by w1,2). If the grey arrows are close to zero, F1 effectively measures recommendation quality (Q) and F2 effectively measures satisfaction with the system (S)

# 3 Case studies

**Chatterjee et. al.** [6] study lays in the neuroscience's domain and focuses on art attributes assessment. It comes useful to us for deciding how to group the future volunteers of our evaluation. We believe that more accurate results on the perception of our recommending system would be accomplished by dividing our users pool into two main categories: expert users and naive users. The above mentioned paper uses a pre-study survey questionnaire for assessing user's expertise level on art by means of 12 questions that are analogue to PC in Knjiniensburg's framework.

| **AAA pre-study survey question** |
|---|
| **0-7 points responses** |
| How many studio art classes have you taken at the high school level or above? |
| How many art history classes have you taken at the high school level or above? |
| How many art theory or aesthetics classes have you taken at the high school level or above? |
| In the average week how many hours do you spend making visual art? |
| In the average week how many hours do you spend reading a publication that is related to visual art? |
| In the average week how many hours do you spend each week looking at visual art? |
| **0-6 points responses** |
| On average, you visit art museums about once every |
| On average, you visit art galleries about once every |
| **others** |
| What is your gender? |
| What is the highest level of education that you have completed? |
| Do you have any visual impairments?? |

After their study Chatterjee et. al. [6] compare the results of the two groups with each other with the aim to find differences in art's perception on different levels of domain's expertise.
They trained their volunteers on some art attributes (e.g., saturation, stroke, realism, etc.) and administered them a questionnaire in which they had to evaluate 24 paintings with respect to the above mentioned attributes. The correlations between judgements of the naive and expert users groups were measured computing standard error with Spearman Rho, rank order correlation between the two groups, differences between groups and differences between group correlations (fig. 6)

Albanese et al. [1] in 2013, Bartolini et al. [5] in 2016 and Amato et al. [2] in 2017 employ for evaluation purposes the TLX factor analysis. TLX is a multi-dimensional rating procedure that provides an overall score based on

| Attribute | Spearman Rho with Std. Error (naïve participants) | Spearman Rho with Std. Error (experienced participants) | Rank order correlations between the two groups, all at $p < 0.001$ significance | $T$ statistic ($df = 88$) testing for differences between groups | $Z$-scores for test for differences between group correlations using Fisher transformations |
|---|---|---|---|---|---|
| Balance | 0.49 (±0.03) | 0.59 (±0.04) | 0.92 | −1.99 ($p = 0.05$) | −0.31 |
| Color Saturation | 0.61 (±0.04) | 0.74 (±0.03) | 0.98 | −1.98 ($p = 0.05$) | −0.51 |
| Color Temp. | 0.63 (±0.02) | 0.72 (±0.03) | 0.95 | −3.11 ($p = 0.003$) | −0.36 |
| Depth | 0.76 (±0.02) | 0.81 (±0.02) | 0.97 | −1.96 ($p = 0.05$) | −0.32 |
| Complexity | 0.61 (±0.03) | 0.64 (±0.03) | 0.97 | −0.8 (NS) | −0.11 |
| Stroke | 0.69 (±0.05) | 0.78 (±0.06) | 0.98 | −1.12 (NS) | −0.41 |
| Abstract | 0.81 (±0.02) | 0.85 (±0.02) | 0.98 | −1.79 (NS) | −0.30 |
| Animacy | 0.71 (±0.02) | 0.725 (±0.02) | 0.83 | −0.38 (NS) | −0.60 |
| Emotion | 0.67 (±0.02) | 0.64 (±0.03) | 0.78 | 1.09 (NS) | 0.13 |
| Realism | 0.80 (±0.01) | 0.74 (±0.03) | 0.98 | 1.43 (NS) | 0.30 |
| Objective Accuracy | 0.74 (±0.02) | 0.75 (±0.02) | 0.95 | −0.34 (NS) | −0.05 |
| Symbolism | 0.63 (±0.03) | 0.71 (±0.03) | 0.95 | −1.90 (NS) | −0.34 |

Figure 6: Table representing AAA results showing correlations of judgment on each attribute and Comparisons of naïve and experienced participants

a weighted average of ratings provided by users by means of proper questionnaires on six sub-scales: mental demand, physical demand, temporal demand, own performance, effort and frustration. The lower TLX scores (ranging in the 0–100 interval), the better they are.[8]

At user perception level we believe it can be worth to investigate the effects of explaining or adding more information to recommendations. This idea comes after the reading of **Friedrich and Zanker** [8] and **Dominiguez et al.** [7] works

The taxonomy introduced by Friedrich and Zanker [8] uses three dimensions to characterize explanations:(i) the recommendation paradigm (collaborative filtering, content-based filtering, knowledge-based.), (ii) reasoning model (white-box or black-box explanation), and (iii) the exploited information categories (user model, recommended item, alternatives).[9]

The study by Dominiguez et al. focuses on evaluating the effects of explanations on user's perception of recommending systems and how independent variables—such as algorithm, explainable interface and domain knowledge—interact with each other in terms of user perception. The study is based on the evaluation framework made by Knjiniesburg et. al.: for example regarding PC in the pre-survey they collected demographic data, using the questionnaire by Chatterjee et. al.
The explanations were based on Friedrich and Zanker's content-based paradigm,

[8]cit. Bartolini et al., "Recommending multimedia visiting paths in cultural heritage applications", p. 3833, 2016

[9]White-box disclose the underlying conceptual model of the recommendation engine to to the user and exploit it to produce explanations, while black-box explanations do not disclose the functioning of the system (either because they are too complex, confusing for the user, etc.)

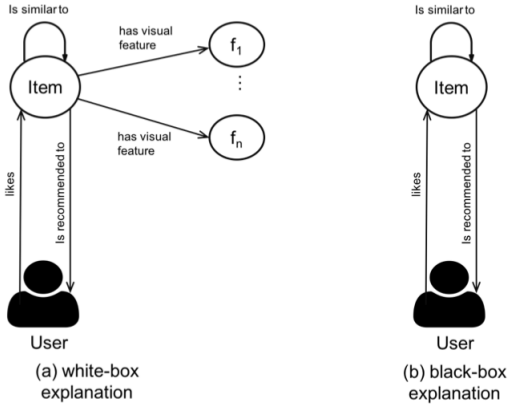


Figure 7: In (a) we have a recommendation based on transparent visual features, while in (b) recommendation is based on item similarity, without details of the features used.

in which the attractiveness of visual features (such as brightness, colorfulness, or luminance) are used to explain the recommendations in a white-box fashion (fig.7).[10]

The experiment consisted in a user study and the employed data was provided by UGallery. Users had to perform a preference elicitation task, i.e. they had to "like" at least ten paintings from the interface. After the task, the participants were asked to next answer a post-algorithm survey, in the form of statements where the user had to indicate their level of agreement in a 0 (totally disagree) to 100 (totally agree) scale:

- **Explainable:** I understood why the art images were recommended to me.
- **Relevance:** The art images recommended matched my interests.
- **Diverse:** The art images recommended were diverse.
- **Interface Satisfaction:** Overall, I am satisfied with the recommender interface.
- **Use Again:** I would use this recommender system again for finding art images in the future.
- **Trust:** I trusted the recommendations made.

**Confirmatory factor analysis** is used to test whether the selected factors are consistent with the hypothesized model.[11] They use *effort* and *satisfaction* as

[10] cf. Friedrich and Zanker,"A Taxonomy for Generating Explanations in Recommender Systems", p. 92 2011

[11] Dominiguez et al. "The Effect of Explanations and Algorithmic Accuracy on Visual Recommender Systems of Artistic Images", p. 7, 2019

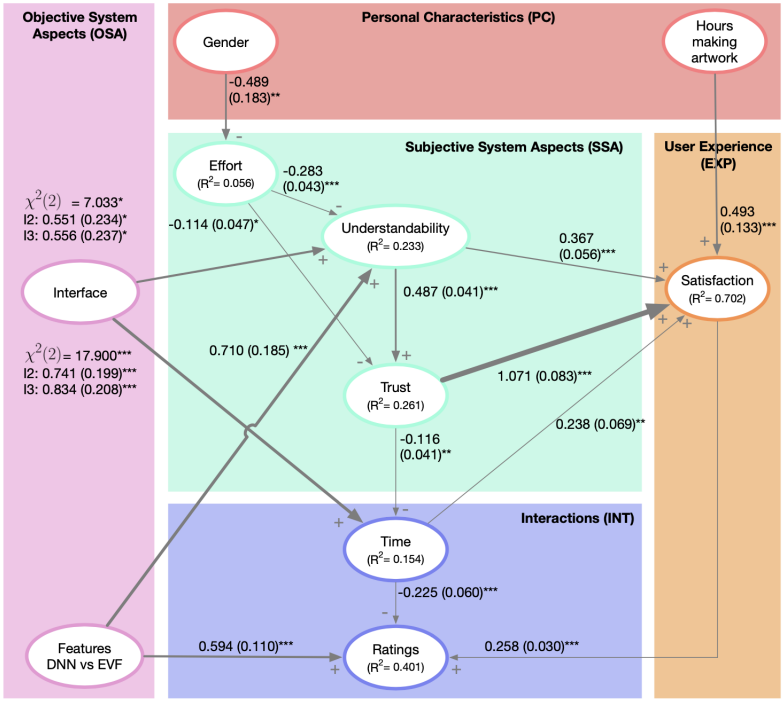


Figure 8: The structural equation model for the data of the experiment using Knijnenburg's evaluation framework for recommender systems. Significance levels: * * * p ¡ .001, * * p ¡ .01, * p ¡ 0.05. R2 is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent the coefficients (and standard error) of the effect. Factors are scaled to have an SD of 1.

factors for this analysis and six items sharing least at 56.2% of their variance with their designated construct and to ensure the convergent validity of constructs the average variance extracted (AVE) of each construct was examined.[12] After clustering variables into constructs they followed Knjininesburg's instructions and conducted a SEM analysis, providing a broader understanding of the variables involved in their study and their relationships. Through this final analysis they explain: 1)the effect of algorithm; 2) the effects of interface on understandability; 3) the effects of interface on time; 4) the effect of trust in satisfaction; 5) the effect of effort (fig. 8).

[12]cf. Ibid.

**Andjelkovic et al.** in [3] recognize the importance of including **explanation**, **transparency**, **control** and **user experience** in general when evaluating their application Moodplay. They carried an user study, testing four different versions of their interface with various degrees of information on items similarity to query in terms of low-level features. The factors studied by their evaluation are: 1) **user satisfaction** in their recommending system, related in the study with recommendation accuracy, serendipity, novelty, control and transparency. [13], and 2) **trust** that is also related to interface transparency.

| **Moodplay post-survey questions** |
|---|
| **1-100 agreement scale responses** |
| I trusted recommendations from the system |
| Interaction with the interface increased my trust in the recommendations |
| The recommendations were diverse |
| The interface helped me understand and compare moods of different artists |
| The interface helped me understand how recommendations were generated |
| The interface allowed me to control the recommendations |
| The interface was confusing |
| Overall, the recommendations were accurate |
| The system was easy to use |
| The interface was slow |
| The tutorial explained the system reasonably well |
| By the end of the session I was satisfied with the recommendations |

In this study log-analysis of user behaviour is employed for measuring diversity. **Diversity** of results is used as a tool to avoid filter bubble effects. Diversity is measured here in two ways: 1) by comparing the number of unique artists rated and played per user in each condition presented in their study. And 2) collecting users' ratings for the recommendation list as a whole and for each one of the 5 items contained.
During the experiment users had to mandatory rate an initial and a final list of recommendations.[14]For measuring differences in ranking they used **nDCG** as a metric: accumulated gain is high when relevant items (rated 4 or 5) appear at the top of the list and the non-relevant elements (rated 1,2 or 3) are placed at the bottom.

$$DCG = \frac{1}{N} \sum_{u=1}^{N} \sum_{j=1}^{J} \frac{g_{ui_j}}{max(1, log_b j)}$$

$$nDCG = \frac{DCG}{DCG*}$$

[13]Andjelkovic et al., Moodplay: Interactive Music Recommendation based on Artists' Mood Similarity, p. 3, 2019
[14]They were also free to rate more lists in between.

To explore the relationships between quantitative and qualitative experimental results collected during the user study, they performed **Principal Component Analysis** (PCA), a technique for dimensionality reduction over Personal characteristics (pre-study questions), perceived accuracy, trust and control that have shown significant effects in previous studies and log analysis data such as number of interactions and duration of the userstudy.

The 2015 study by **Parra and Brusilovsky** in [11] aims at evaluating the effects of controllabiity on user engagement and experience. We find this study interesting because they explicitly set research questions, user experiment's tasks and the evaluation metrics (divided into objective and subjective metrics). The paper also provides useful guidelines as the one by O'Brien and Toms on research on user engagement in software applications and the above described frameworks by Knijnenburg et al. in [9] and Pu et al. in [12] (ResQue). [15] Parra and Brusilovsky main aim is to evaluate the effects of different degrees of controllability on user experience and engagement in an conference items recommender system. These two dimensions are interesting for us, since our systems retrieves similar cultural items, i.e. artworks that can be experienced and perceived as engaging by consumers.

As the evaluation of SetFusion is based on Knijnienburg's framework, it also gives us an insight on how survey questions should be formulated to correctly assess users' competencies and contextual characteristics depending on the investigated dimensions. The users characteristics they decided to focus on correspond to:

- User expertise in her own domain: Is the user knowledgeable in her own domain?
- Familiarity with iConference: How familiar is the user with the user community of iConference?
- User experience with the system: Has the user used Conference Navigator 3 before?
- Trusting Propensity: Does the user have an inherent propensity to trust in people or systems?
- User experience with recommendation systems: Does the user have some previous experience or knowledge about recommender systems?

[16]

[15] cf. Parra and Brusilovsky, "User-controllable penalization: A case study with SetFusion", p. 49, in "International Journal of Human-Computer Studies", vol. 78, pp. 43-67, 2015

[16] cit. Ibid., p. 50

Finally objective metrics are also listed in the study report as follows, starting from Manning's list in [10]:

- **Log-data** referring to engagement: number of talks explored; number of talks bookmarked; number of clicks; and amount of time
- **Average rating**: we compare the conditions by calculating the mean over the average rating of each user under a particular condition.
- **Precision@k**: this metric allow us to measure the accuracy of a list of k recommendations.

$$Precision = \frac{\#(relevantitemsretrieved)}{\#(retrieveditems)} = P(relevant|retrieved)$$

- **MAP**: Mean Average Precision is a metric that calculates the mean over the average precision of several lists. The average precision of one list is calculated by averaging the precision at several cut points, usually the recall points (the positions of the list where the element found is relevant).

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$$

- **MRR**: stands for Mean Reciprocal Rank. It is calculated as the inverse of the ranking position of the first relevant element to be found on a list.

$$MRR(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{Relevance_Label_Value}{rank_i}$$

# 4 Conclusion

Among the sixteen dimensions listed by Avazpour et al. [4] in 2014 we recognized as characteristics of the context of multimedia recommending systems, dealing with images and artistic/cultural items, the dimensions of: correctness in terms of relevance and accuracy; diversity and serendipity; trust and explainability; utility, usability and user preferences (in terms of interaction an overall experience with the system).

Given the state-of-the-art we believe that employing a user study would be the best way to proceed, consisting in the creation of an application resuming

the main characteristics of our system and enhancing some interface aspects (e.g., the amount of information about artworks while scrolling the list). The goal is collecting data on user's perception of factors characterizing the recommending system and related to the above mentioned dimensions.
Questionnaires, as evaluative tools, have now a relevant role in the assessment of user opinion's, since it is possible to quantify and operate on subjective data such as human perception, making it prone to statistical analysis. TLX factor analysis along with questionnaires has revealed itself to be another useful tool when measuring perception of the user while doing tasks in an experimental environment. This test is actually used to measure the effort perceived by the user in accomplishing a given task and is useful for comparing the issues of different goups of users [17].

The triangulation of manipulations, surveys' data and logged behavioural data as Knjinienburg et al. state in [9] show several trade offs between system aspects and personal and situational characteristics. This kind of features are important in the evaluation and development of systems dealing with aesthetic taste of users as stated in [3]. Reusing an existing and widely used framework as base-ground for evaluating recommending systems, could also help to solve the difficulty of reuse in different environments and scenarios of evaluating applications, a problem Amato et al. discuss in [2] and help the community reach a standard approach in this study-field.

[17] Bartolini et al., "Recommending multimedia visiting paths in cultural heritage applications", p. 3833, 2016

| Papers' dimensions alignment | | | | |
|---|---|---|---|---|
| **Avazpour et. al.** | **Pu et al.** | **Knijnensburg et al.** | **Dominiquez et. al.** | **Andjelkovic et al.** |
| Correctness | Perceived accuracy | Perceived quality | Relevance | Accuracy |
| Coverage | | | | |
| Diversity | Familiarity | Perceived variety | Diversity | Diversity |
| Trustworthiness | Trust/ explainability | General trust | Trust | Trust |
| Recommender confidence | Confidence | Perceived effectiveness | | |
| Novelty | Novelty | | | |
| Serendipity | Attractiveness | Perceived fun | | |
| Utility | Perceived Usefulness | | Use again | |
| Risk | | | | |
| Robustness | | | | |
| Learning rate | | | | |
| Usability | Perceived Ease of Use | Effort to use the system | | |
| Scalability | | | | |
| Stability | Context compatibility | | | |
| Privacy | | System-specific privacy concerns | | |
| User preference | Overall Satisfaction/ enjoyability | Choice satisfaction / Intention to provide feedback | Interface satisfaction | User experience |

# Bibliography


[1] Massimiliano Albanese et al. "A Multimedia Recommender System". In: *ACM Transactions on Internet Technology* 13 (Nov. 2013). 79 citations. ISSN: 1533-5399, 1557-6051. DOI: 10.1145/2532640. URL: https://dl.acm.org/doi/10.1145/2532640.

[2] Flora Amato et al. "Big Data Meets Digital Cultural Heritage: Design and Implementation of SCRABS, A Smart Context-awaRe Browsing Assistant for Cultural EnvironmentS". In: *Journal on Computing and Cultural Heritage* 10 (Apr. 1, 2017), pp. 6.1–6.23. DOI: 10.1145/3012286.

[3] Ivana Andjelkovic, O'Donovan John, and John O'Donovan. "Moodplay: Interactive music recommendation based on Artists' mood similarity". In: *International Journal of Human-Computer Studies* 121 (Jan. 2019). 56 citations. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2018.04.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581918301654.

[4] Iman Avazpour et al. *Dimensions and metrics*. 106 citations. 2014. DOI: 10.1007/978-3-642-45135-5_10. URL: https://link.springer.com/chapter/10.1007/978-3-642-45135-5_10.

[5] Ilaria Bartolini et al. "Recommending multimedia visiting paths in cultural heritage applications". In: *Multimedia Tools and Applications* 75.7 (Apr. 2016). 74 citations, pp. 3813–3842. ISSN: 1380-7501, 1573-7721. DOI: 10.1007/s11042-014-2062-7. URL: http://link.springer.com/10.1007/s11042-014-2062-7.

[6] Anjan Chatterjee et al. *The Assessment of Art Attributes*. 153 citations. 2010. DOI: 10.2190/EM.28.2.f. URL: https://journals.sagepub.com/doi/abs/10.2190/EM.28.2.f.

[7] Vicente Dominguez et al. "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 25 citations. ACM, Mar. 17, 2019, pp. 408–416. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302274. URL: https://dl.acm.org/doi/10.1145/3301275.3302274.

[8] Gerhard Friedrich and Markus Zanker. "A Taxonomy for Generating Explanations in Recommender Systems". In: *AI Magazine* 32.3 (June 9, 2011). 153 citations, pp. 90–98. ISSN: 2371-9621, 0738-4602. DOI: 10.1609/aimag.v32i3.2365. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/2365.

[9] Bart P. Knijnenburg et al. "Explaining the user experience of recommender systems". In: *User Modeling and User-Adapted Interaction* 22.4 (Oct. 2012). 728 citations, pp. 441–504. ISSN: 0924-1868, 1573-1391. DOI: 10.1007/s11257-011-9118-4. URL: http://link.springer.com/10.1007/s11257-011-9118-4.

[10] Christopher Manning, Prabhakar Raghavan, and Hinrich Schuetze. *Introduction to Information Retrieval*. 21982 ctations. 2009.

[11] Denis Parra and Peter Brusilovsky. "User-controllable personalization: A case study with SetFusion". In: *International Journal of Human-Computer Studies* 78 (June 2015). 98 ctations, pp. 43–67. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2015.01.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581915000208.

[12] Pearl Pu and Li Chen. "A User-Centric Evaluation Framework of Recommender Systems". In: *AI Magazine* 612 (2010). 777 citations. ISSN: 1613-0073. DOI: 10.1145/2043932.2043962. URL: https://dl.acm.org/doi/abs/10.1145/2043932.2043962.