Evaluating the effectiveness of explanations for recommender systems

Nava Tintarev · Judith Masthoff

Abstract

- Seven aims for explanations:
  - effectiveness, satisfaction, transparency, scrutability, trust, persuasiveness, efficiency
- Focuses on tradeoff between satisfaction and effectiveness
- Shows that personalization is detrimental to effectiveness

Introduction

- Recommender systems are traditionally evaluated in recommendation accuracy
- Increased interest in user-centered evaluation, like satisfaction
- When recommended a movie - it’s natural to wonder why.

Introduction - the 7 aims

Table 1 Explanatory aims

Aim	Definition
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness	Help users make good decisions
Persuasiveness	Convince users to try or buy
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of use or enjoyment

Introduction

- Feature based explanations
  - strong persuasive effect

- “We will investigate how personalization of item features can affect explanation effectiveness and user satisfaction.”

The Domain

What features do users use to decide :

- Which movie to watch
- Which digital camera to buy

Effectiveness

- How do you measure effectiveness?
  - Bilgic and Mooney (2005)
    - (Rating1) The user rates the item on the basis of the explanation
    - The user tries the item
    - (Rating2) The user re-rates the item
- High effectiveness: *minimize(Rating1 - Rating2)*

Over- and Underestimation

- Difference of 2 is same as -2 (in the math)
  - overestimation less helpful than underestimation
- Perceived negative effect of rating discrepancy is greater in high cost domains
- Gaps in the negative end of the scale have higher negative impact on perceived effectiveness

Experimental Setup

Explanations:

- Baseline: neither personalized nor item describing
- Non-personalized feature based
- Personalized feature based

Experimental Setup

4 experiments

- 1. movies, simulated (already done)
- 2. movies, simulated
- 3. cameras, simulated
- 4. movies, real

The Procedure

1. Participants provided background information
2. Participants rated importance of features and selected favourite actors, directors etc.
3. Participants were shown items and rated
    a. How much they would like this item
    b. How good the explanation was
4. Participants tried the item
5. Participants re-rated the item

The Hypothesis

- **H1:** Personalized feature-based explanations will be more effective than non-personalized feature-based and baseline explanations.
- **H2:** Users will be more satisfied with personalized feature-based explanations compared to non-personalized feature-based and baseline explanations.

Experiment 1 and 2 - results

Condition	Movie Before	Movie After	Effectiveness (absolute)	Effectiveness (signed)
Exp. 1				
Baseline	3.45 (1.26)	4.11 (1.85)	1.38 (1.20)	−0.69 (1.69)
Non-personalized	3.85 (1.87)	4.43 (2.02)	1.14 (1.30)	−0.57 (1.64)
Personalized	3.61 (1.65)	4.37 (1.93)	1.40 (1.20)	−0.77 (1.68)
Exp. 2				
Non-personalized	3.84 (1.95)	3.93 (1.95)	0.96 (0.81)	−0.09 (1.25)
Personalized	3.75 (2.05)	4.00 (1.87)	1.33 (1.27)	−0.25 (1.85)

Experiment 1 and 2 - results

Condition	Explanation Before	Explanation After
Experiment 1		
Baseline	2.38 (1.54)	2.85 (1.85)
Non-personalized	2.50 (1.62)	2.66 (1.89)
Personalized	3.09 (1.70)	3.15 (1.99)
Experiment 2		
Non-personalized	2.72 (1.68)	2.83 (1.74)
Personalized	3.31 (1.55)	2.97 (1.33)

Experiment 3

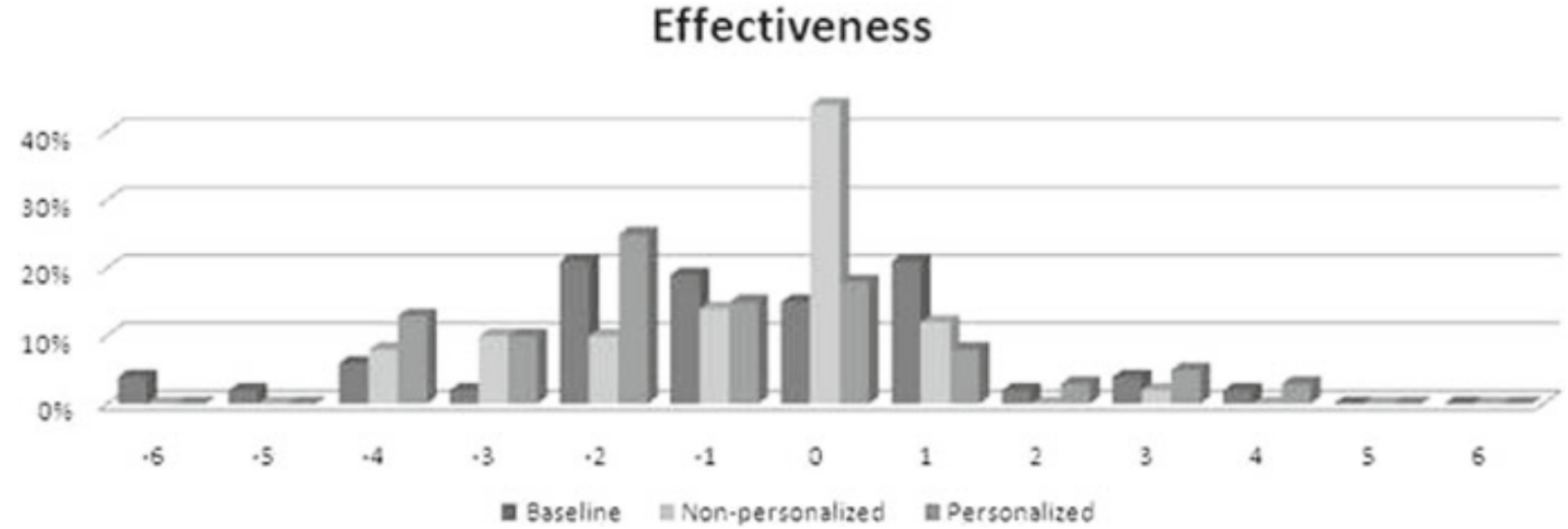
The camera domain:

- serves as a control
- explanations accompanied with generic camera image

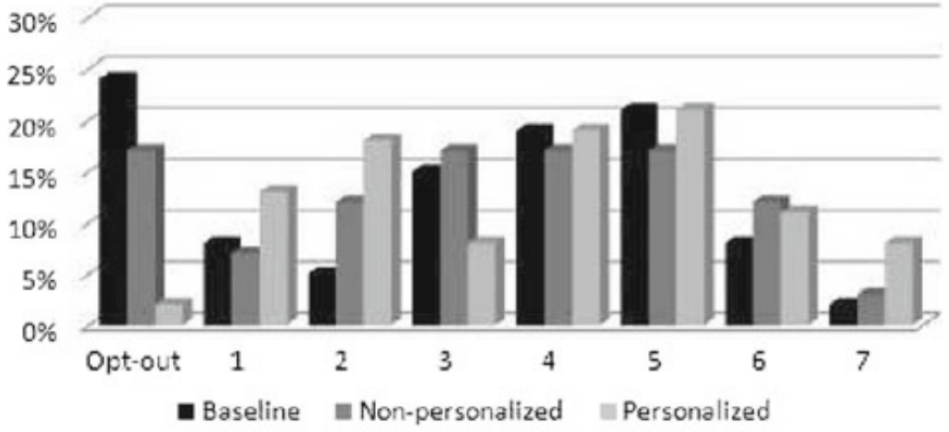
Experiment 3 - results

Condition	Camera Before	Camera After	Effectiveness (absolute value)	Effectiveness (signed value)
Baseline	3.94 (1.47)	4.75 (1.73)	1.77 (1.50)	−0.77 (2.20)
Non-personalized	3.88 (1.62)	4.78 (1.75)	1.14 (1.32)	−0.78 (1.57)
Personalized	3.83 (1.86)	4.95 (1.77)	1.88 (1.34)	−1.08 (2.05)

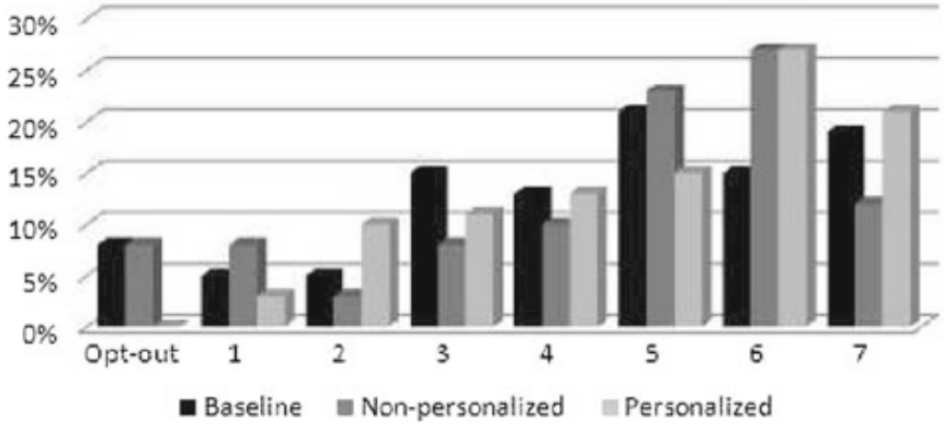
Experiment 3 - results



Camera Before



Camera After



Experiment 3 - results

Were users more satisfied?

Table 12 Experiment 3: means (SD) of the two explanation ratings, excluding opt-outs, (on a scale from 1 to 7, 1 = really bad, 7 = really good) per condition. "Before" and "After" denote the two explanation ratings before and after viewing Amazon reviews

Condition	Explanation Before	Explanation After
Baseline	2.83 (1.44)	3.80 (1.87)
Non-personalized	2.38 (1.64)	2.87 (1.94)
Personalized	3.27 (1.27)	2.67 (1.56)

Experiment 4

- Experiment 1, 2 and 3 were simulated
- Used short movies

The procedure

1. User rates the item based on the title
2. User rates the item based on explanation
3. User actually tries the item
4. User rates the item again

Experiment 4 - results

Movie ratings

Condition	Movie Title	Movie Before	Movie After
Baseline	4.36 (0.95)	4.28 (0.81)	4.76 (1.67)
Non-personalized	4.12 (1.67)	4.45 (1.53)	4.58 (1.88)
Personalized	3.86 (1.23)	4.31 (1.26)	4.93 (1.86)

Effectiveness

Condition	Effectiveness (absolute value)	Effectiveness (signed value)
Baseline	1.09 (1.00)	−0.41 (1.43)
Non-personalized	1.78 (1.37)	−0.08 (2.26)
Personalized	1.69 (1.08)	−0.41 (1.98)

Experiment 4 - results

Table 17 Experiment 4: means (SD) and percentage of opt-outs of the two explanation ratings, (on a scale from 1 to 7, 1 = really bad, 7 = really good) per condition. "Before" and "After" denote the two explanation ratings before and after viewing the movie. Means exclude opt-outs

Condition	Explanation Before		Explanation After	
	Mean (SD)	Opt-outs (%)	Mean (SD)	Opt-outs (%)
Baseline	2.55 (1.43)	14.6	2.89 (1.60)	2.1
Non-personalized	3.51 (1.61)	4.5	3.53 (2.00)	0.0
Personalized	3.21 (1.46)	4.3	3.16 (1.83)	2.2

Experiment 4 - summary

- Baseline most effective, but lowest satisfaction
- People opt-out less when given an explanation
- Weaker results for personalized explanations is probably due to restricted choice of materials

Conclusions

Table 19 Summary: means (SD) of initial satisfaction with explanations, excluding opt-outs, (on a scale from 1 to 7, 1 = really bad, 7 = really good) per condition per experiment

Condition	Movies I	Movies II	Cameras	Final evaluation
Baseline	2.38 (1.54)	–	2.83 (1.44)	2.55 (1.43)
Non-personalized	2.50 (1.62)	2.72 (1.68)	2.38 (1.64)	3.51 (1.61)
Personalized	3.09 (1.70)	3.31 (1.55)	3.27 (1.27)	3.21 (1.46)

Conclusion

Table 20 Summary: means (SD) of absolute effectiveness (excluding opt-outs), per condition per experiment

Condition	Movies I	Movies II	Cameras	Final evaluation
Baseline	1.38 (1.20)	–	1.77 (1.50)	1.09 (1.00)
Non-personalized	1.14 (1.30)	0.96 (0.81)	1.14 (1.32)	1.78 (1.37)
Personalized	1.40 (1.20)	1.33 (1.27)	1.88 (1.34)	1.69 (1.08)

Conclusion

Table 22 Overview of the results related to effectiveness and satisfaction across all experiments

Experiment	Effectiveness	Satisfaction
Movies I	Trend: non-personalized best	Significant: personalized highest
Movies II	Trend: non-personalized best	Trend: personalized highest
	Significant: non-personalized had least opt-outs	
Cameras	Significant: non-personalized best, personalized had least opt-outs	Significant: personalized higher than non-personalized
		Trend: personalized higher than baseline
Final Eval.	Significant: baseline best, but had most opt-outs	Significant: non-personalized higher than baseline
		Trend: personalized higher than baseline

Conclusions

- Personalization decreases effectiveness
- Users are likely to be more satisfied with feature-base explanations than baseline.
- User satisfaction is reflected in the proportion of opt-outs

Lessons learned

- Effectiveness does not consider opt-outs
- Mid scale ratings could mean the user don’t know what to do or just feel neutral.
  - also likely to lead to better effectiveness