



PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Evaluación de Recomendadores Centrada en el Usuario

Denis Parra

IIC 3633, Sistemas Recomendadores

PUC Chile

Temas

- Transparencia y Explicabilidad
- Controlabilidad
- Visualizaciones e Interactividad
- Algunos ejemplos para evaluación de la experiencia del usuario
- Frameworks para evaluación
 - Pearl Pu
 - Bart Knijnenburg

Por qué evaluación centrada en el usuario?

- Mayoría de investigación evalúa resultado de recomendaciones off-line (RMSE, P@K, etc.)
- Mejoras pequeñas de predicción en los algoritmos no siempre se traducen en una mejor percepción de los usuarios (Konstan & Riedl 2012)
- La precisión de los algoritmos es sólo uno de los factores que influyen la aceptación de las recomendaciones por parte de los usuarios

Explicabilidad

- Capítulo en “HandBook of Recommender Systems” [Tintarev & Masthoff, 2012]
- Ellas proponen algunas direcciones generales para diseñar explicaciones para SisRec
 - Considerar beneficios a obtener (propósito)
 - Evitar (o buscar) relación con funcionamiento del recomendador
 - Presentación y forma de interacción
 - Relación entre algoritmo y tipo de explicaciones

1. Criterios de Explicación

Propósito	Descripción
1.1 Transparencia	Explicar cómo funciona el sistema
1.2 Escrutabilidad	Dejar al usuario indicar que el sistema comete un error
1.3 Confianza	Incrementar confianza del usuario en el sistema
1.4 Efectividad	Ayudar al usuario a tomar buenas decisiones
1.5 Persuasión	Convencer a usuario a probar o a comprar
1.6 Eficiencia	Ayudar a usuarios a tomar decisiones más rápido
1.7 Satisfacción	Aumentar facilidad de uso o el disfrute en el sistema

1.1 Transparencia

- Ejemplo a partir de artículo del Wall Street Journal:

“If TiVo Thinks You Are Gay, Here’s How to Set It Straight”

- Un usuario sospechó que TiVo pensó que él era homosexual pues el sistema comenzó a grabar automáticamente estos programas.
- En el artículo, se explica que este es un caso en que un usuario podría requerir transparencia en el algoritmo recomendador.

Escrutabilidad

- Permitir al usuario inspeccionar o “escrutar” el resultado de la recomendación
- Si bien está relacionado con transparencia, se sugiere identificar y separarlo como ítem.

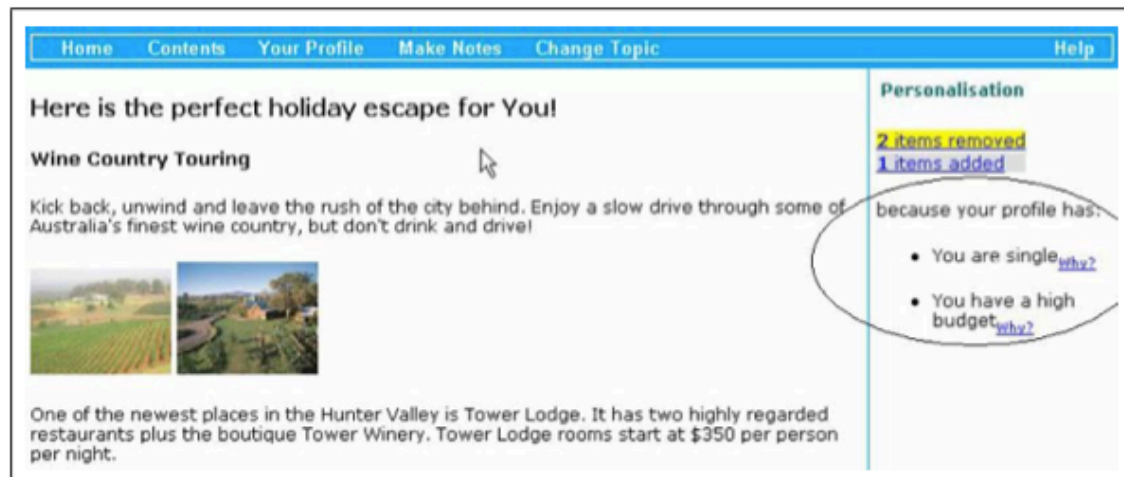


Fig. 15.1: Scrutable holiday recommender [21]. The explanation is in the circled area, and the user profile can be accessed via the “why” links.

Escrutabilidad

The screenshot displays the SetFusion recommender interface. On the left, a control panel titled "Tune weights of the recommender methods:" includes three sliders: "Most bookmarked papers" (0.4), "Similar to your favorite articles" (0.8), and "Frequently cited authors in ACM DL" (0.4). Below the sliders is an "Update Recommendation List" button. A diagram labeled (c) shows three overlapping circles representing different recommendation methods: "Most bookmarked papers" (blue), "Similar to your favorite articles" (yellow), and "A citation recommendation system" (red). A legend indicates that green circles represent "Articles in top30" and grey circles represent "Articles not in top30". A blue arrow points from the top of the article list to the diagram, and a green dashed circle highlights the top article in the list.

(b) Tune weights of the recommender methods:

- Most bookmarked papers: 0.4
- Similar to your favorite articles: 0.8
- Frequently cited authors in ACM DL: 0.4

Update Recommendation List →

* Hover over circles to explore articles
* Click on the diagram to highlight subsets

(c) Similar to your favorite articles: Articles in top30 (green), Articles not in top30 (grey)

Most bookmarked papers: A citation recommendation system

(a) 2. Can't see the forest for the trees? A citation recommendation system
by C. Lee Giles, Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra [see abstract]

3. When thumbnails are and are not enough: Factors behind users
by Dan Albertson [see abstract]

7. Gendered Artifacts and User Agency
by Andrea R. Marshall, Jennifer A. Rode [see abstract]

8. Two Paths to Motivation through Game Design Elements: Reward-Based Gamification and Meaningful Gamification
by Scott Nicholson [see abstract]

9. Automatic Identifying Search Tactic in Individual Information Seeking: A Hidden Markov Model Approach
by Zhen Yue, Shuguang Han, Daqing He [see abstract]

11. Old Maps and Open Data Networks
by Werner Robitza, Carl Lagoze, Bernhard Haslhofer, Keith Newman, Amanda Stefanik [see abstract]

14. Effects of User Identity Information On Key Answer Outcomes in Social Q&A
by Erik Choi, Craig Scott, Chirag Shah [see abstract]

15. Ebooks and cross generational perceived privacy issues
Jennifer Sue Thiele, Renee Kapusniak [see abstract]

16. Toward a mesoscopic analysis of the temporal evolution of scientific collaboration networks

SetFusion: A Controllable Hybrid Recommender

Parra, D., Brusilovsky, P., Trattner, C.

IUI 2014, Haifa, Israel

Confianza

- Mayor transparencia y posibilidad de interactuar con el recomendador está asociado en varios estudios con mayor confianza en el sistema
- Podría estar asociado directamente a la precisión de predicción de la recomendación, pero no siempre!
- Una buena métrica de confianza: Lealtad del usuario en volver a usar el sistema

Confianza

- Dos trabajos muestran que confianza/satisfacción y predicción no siempre están correlacionados

McNee et al. **Don't look stupid: avoiding pitfalls when recommending research papers.** CSCW (2006)

Cramer et al. **The effects of transparency on trust in and acceptance of a content-based art recommender.** UMUAI 18(5), 455–496 (2008).

Persuasión

- Uno de los primeros trabajos en el área de “explicabilidad” de recomendaciones intentaba explicar al usuario las recomendaciones hechas; probaron 21 métodos posibles.
- El autor del paper en algún momento llamó la atención de no considerar ese estudio como el modelo de explicabilidad, ya que hacer al usuario consciente de una decisión y persuadirlo puede tener efectos importantes

Persuasión II

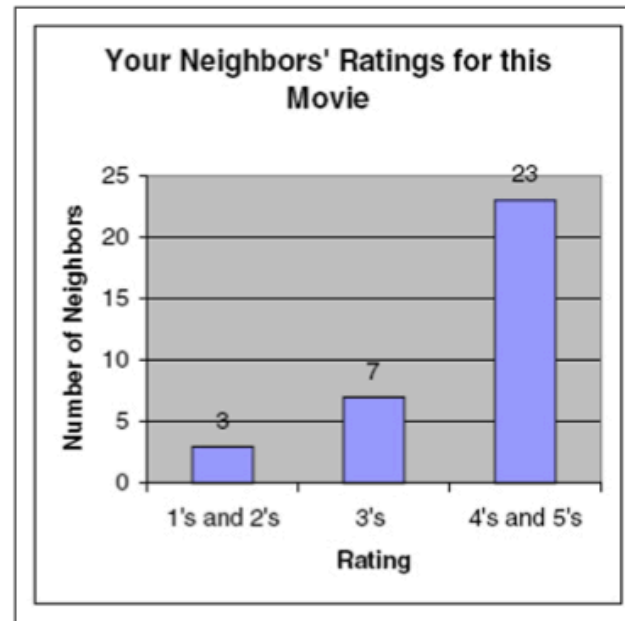


Fig. 15.2: One out of twenty-one interfaces evaluated for persuasiveness - a histogram summarizing the ratings of similar users (neighbors) for the recommended item grouped by good (5's and 4's), neutral (3's), and bad (2's and 1's), on a scale from 1 to 5 [29].

Herlocker, J.L., Konstan, J.A., Riedl, J.: **Explaining collaborative filtering recommendations.**

In: ACM conference on Computer supported cooperative work, pp. 241–250 (2000)

Efectividad

- Conectado con la definición anterior, la explicación/persuasión de una recomendación debiese estar asociada a una buena percepción del usuario
- “Vig et al. measure perceived effectiveness: “This explanation helps me determine how well I will like this movie.” [62]. ”
- ***Se podría medir como la diferencia entre la percepción del ítem al momento de elegirlo y después del consumo.***

Efectividad II

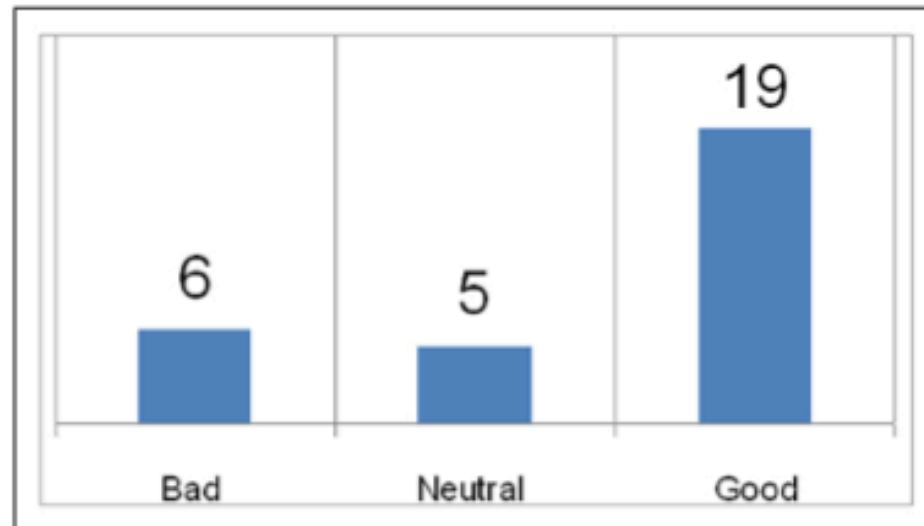
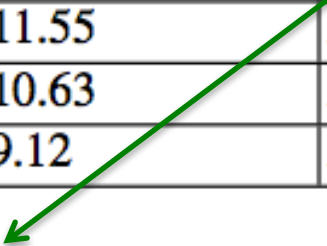


Fig. 15.3: The Neighbor Style Explanation - a histogram summarizing the ratings of similar users (neighbors) for the recommended item grouped by good (5's and 4's), neutral (3's), and bad (2's and 1's), on a scale from 1 to 5. The similarity to Figure 15.2 in this study was intentional, and was used to highlight the difference between persuasive and effective explanations [11].

Efectividad III

Table 15.3: The keyword style explanation by [11]. This recommendation is explained in terms of keywords that were used in the description of the item, and that have previously been associated with highly rated items. “Count” identifies the number of times the keyword occurs in the item’s description, and “strength” identifies how influential this keyword is for predicting liking of an item.

Word	Count	Strength	Explain
HEART	2	96.14	<u>Explain</u>
BEAUTIFUL	1	17.07	<u>Explain</u>
MOTHER	3	11.55	<u>Explain</u>
READ	14	10.63	<u>Explain</u>
STORY	16	9.12	<u>Explain</u>



Title	Author	Rating	Count
Hunchback of Notre Dame	Victor Hugo, Walter J. Cobb	10	11
Till We Have Faces: A Myth Retold	C.S. Lewis, Fritz Eichenberg	10	10
The Picture of Dorian Gray	Oscar Wilde, Isobel Murray	8	5

Eficiencia

- Bajo este parámetro, los tipos de explicaciones debieran optimizarse por dominio para elegir entre opciones que compiten. Por ejemplo, en cámaras

<<“Less Memory and Lower Resolution and Cheaper” >>

Altamente usado en “Conversational” SisRec, donde el usuario refina iterativamente sus preferencias.

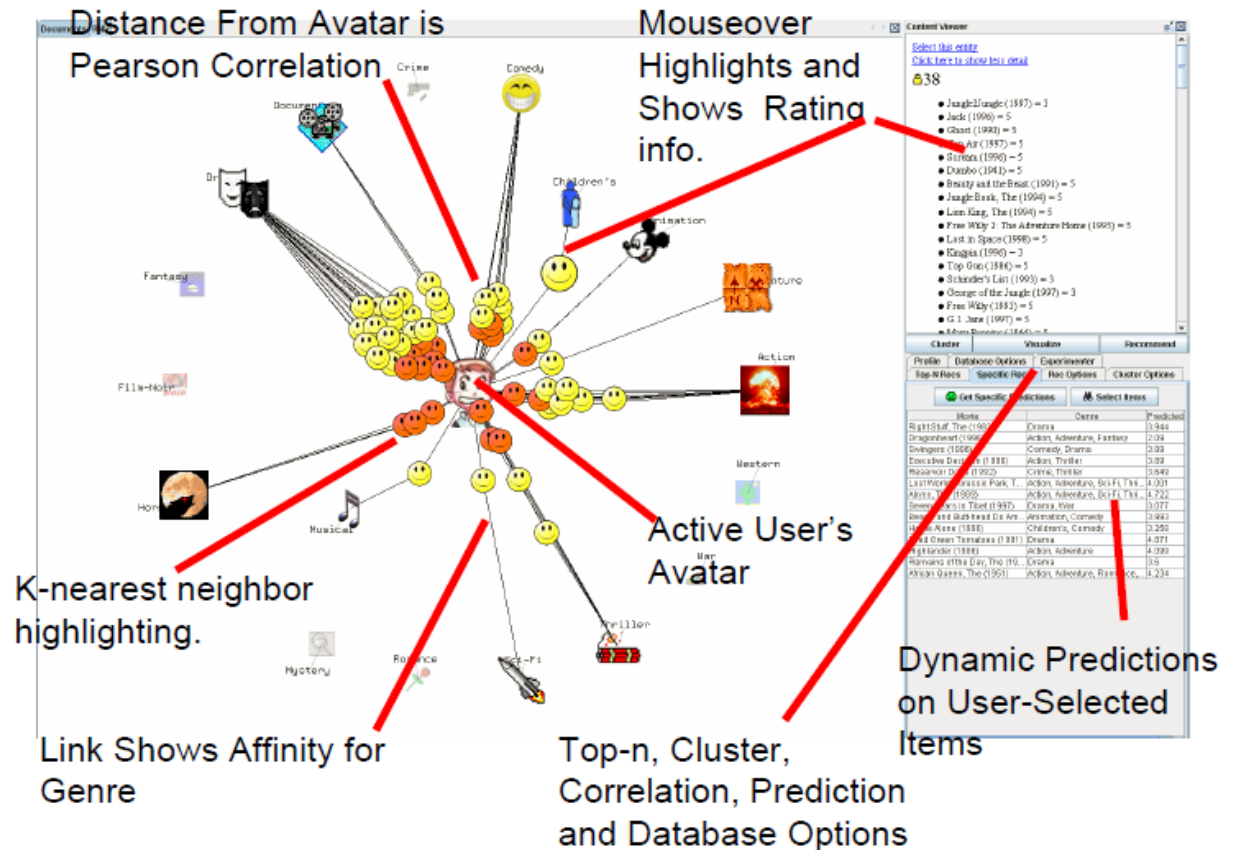
Satisfacción

- Esta es probablemente la métrica que resume de mejor forma el objetivo de un sistema recomendador
- Existen algunos instrumentos (cuestionarios con varios sets de preguntas) que intentan medir esta dimensión. Lo veremos en más detalle en User Centric Evaluation Frameworks.

Visualizaciones

Related work on Visual RS - 1

- 2008: PeerChooser (CHI 2008)
- John O'Donovan and Barry Smyth (UCD)
- Brynjar Gretarsson, Svetlin Bostandjiev, Tobias Hollerer (UCSB)

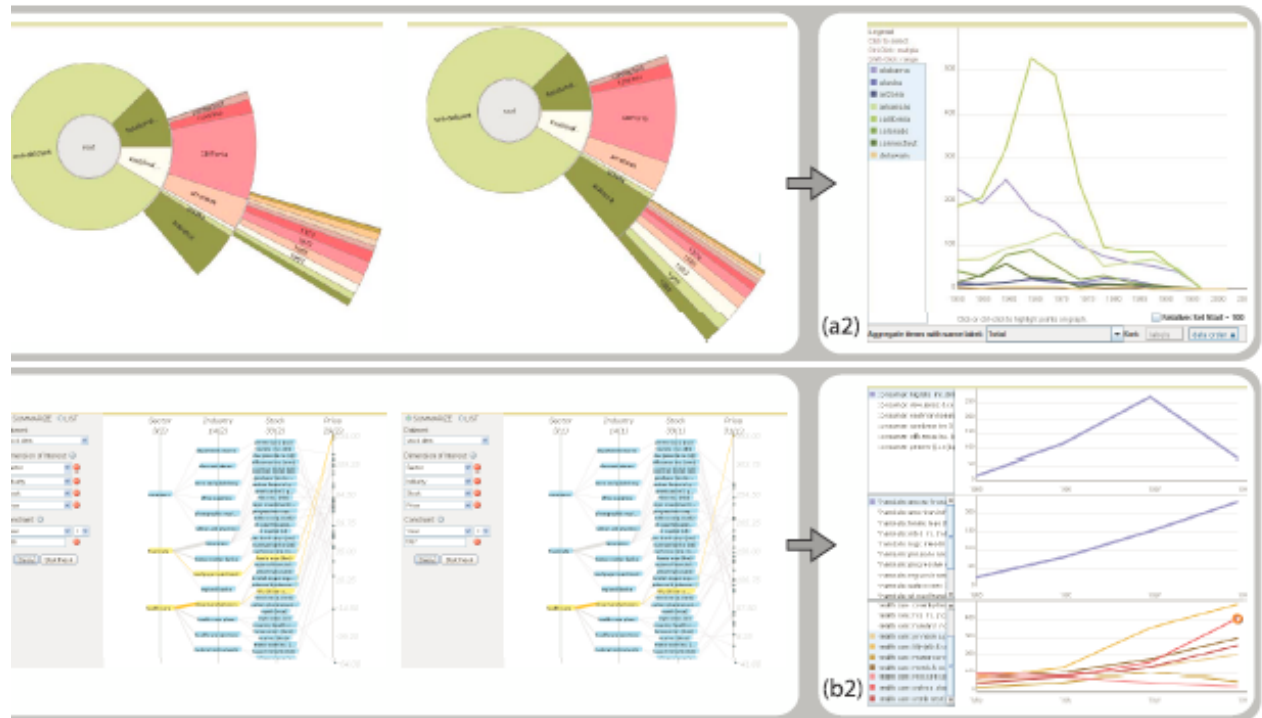


2: Annotated Screenshot of PeerChooser's Interactive Interface.

Related work on Visual RS - 2

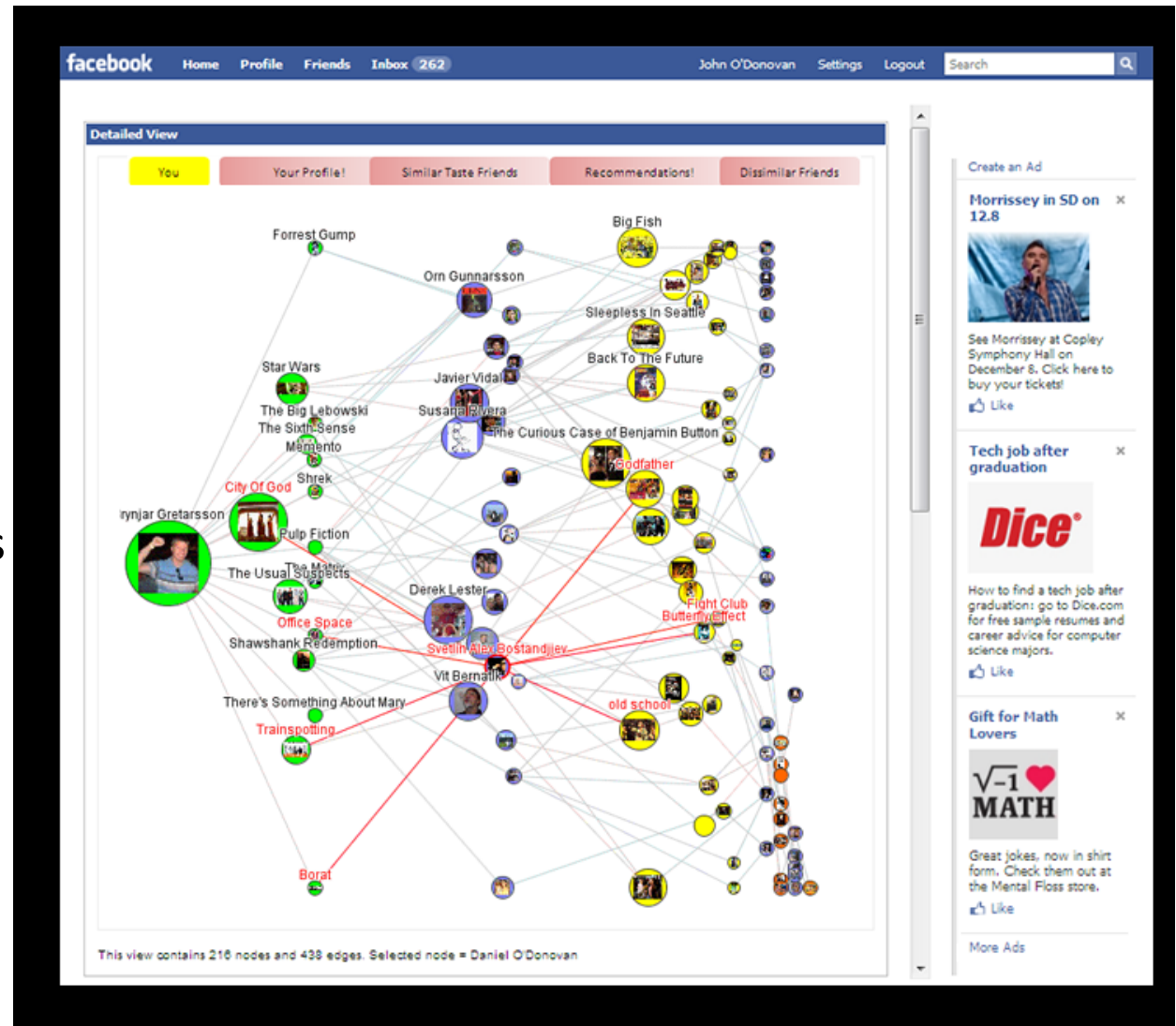
- 2009: Behavior-driven Visualization Recommendations (IUI 2009)
- David Gotz, Zhen Wen (IBM Research)

Given certain tasks inferred from user's behavior, recommend visualizations to accomplish those tasks more efficiently



Related work on Visual RS – 2

- 2010: “SmallWorlds: Visualizing Social Recommendations” IEEE-VGTC 2010
- Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, Tobias Höllerer(UCSB)
- User study with 17 users



Related work on Visual RS - 3

- 2010: Pharos “Who is Talking about What: Social Map-based Recommendation for Content-Centric Social Websites” (RecSys 2010)
- Zhao et al.(IBM China)

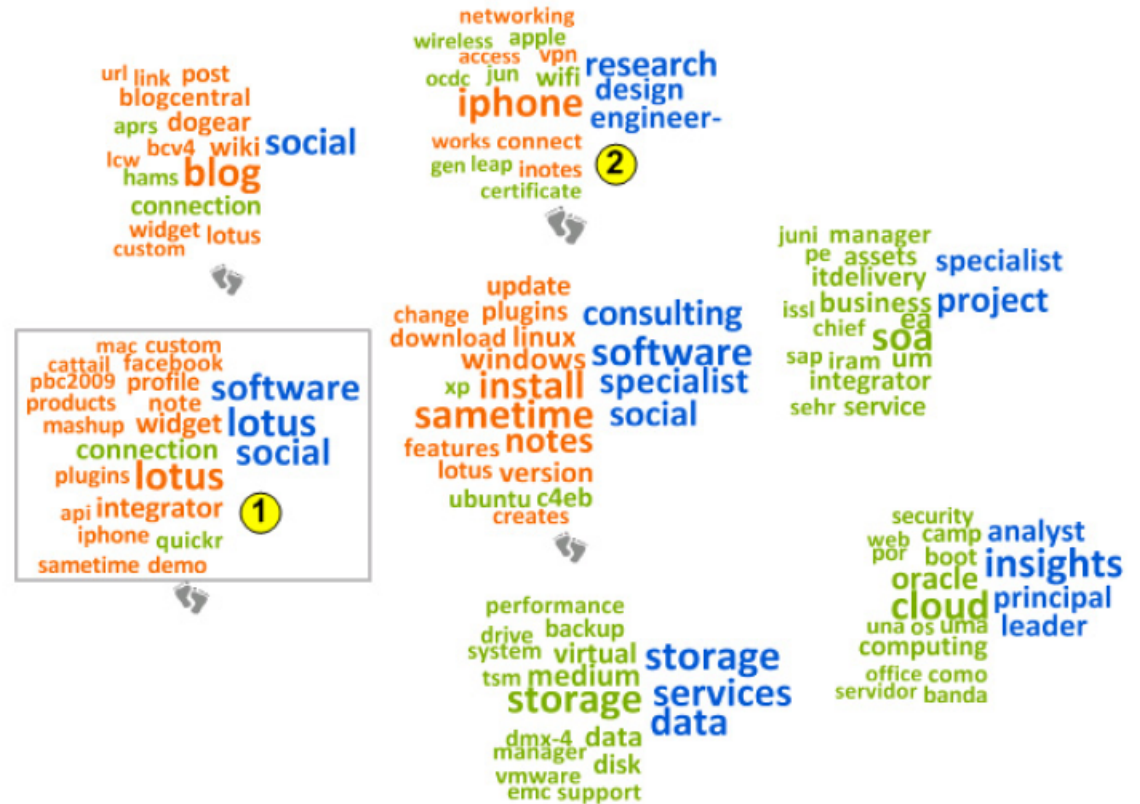


Figure 2: Highlight a user’s activities (keywords in orange) in multiple communities. The size of the footprint indicates how active the user is in the attached community.

Related Work – 4 😊

- **2010: Opinion Space: A Scalable Tool for Browsing Online Comments**
- Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, Ken Goldberg
- Software sponsored by US Government to diversify political opinions

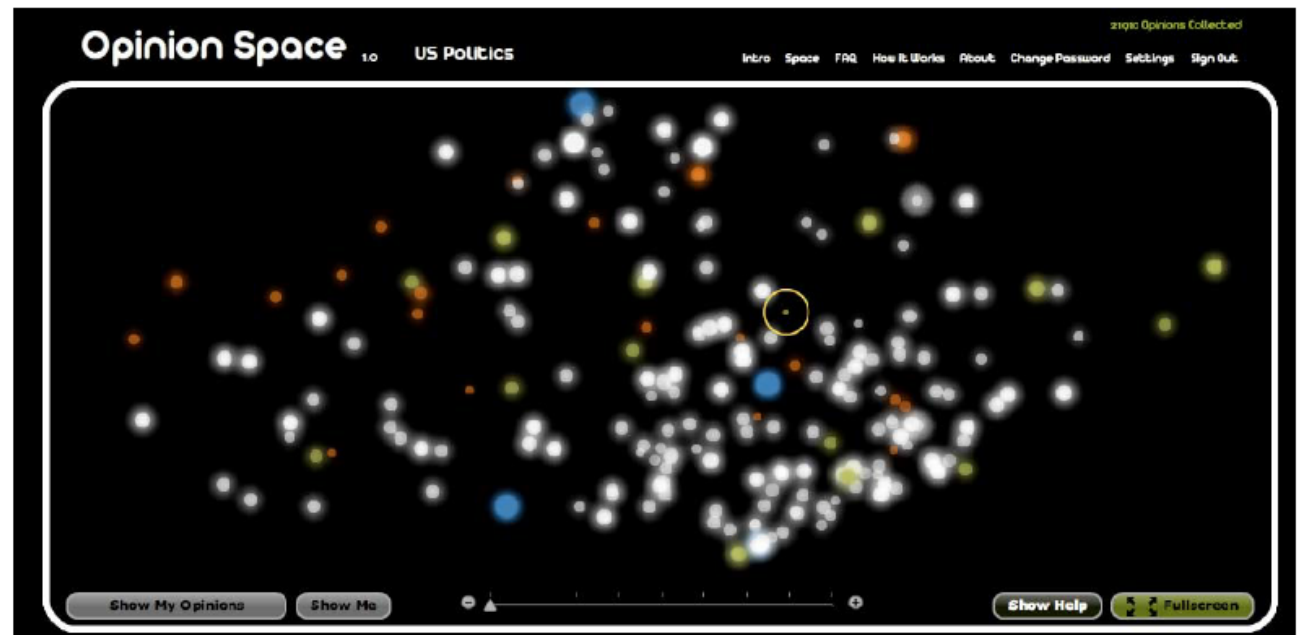
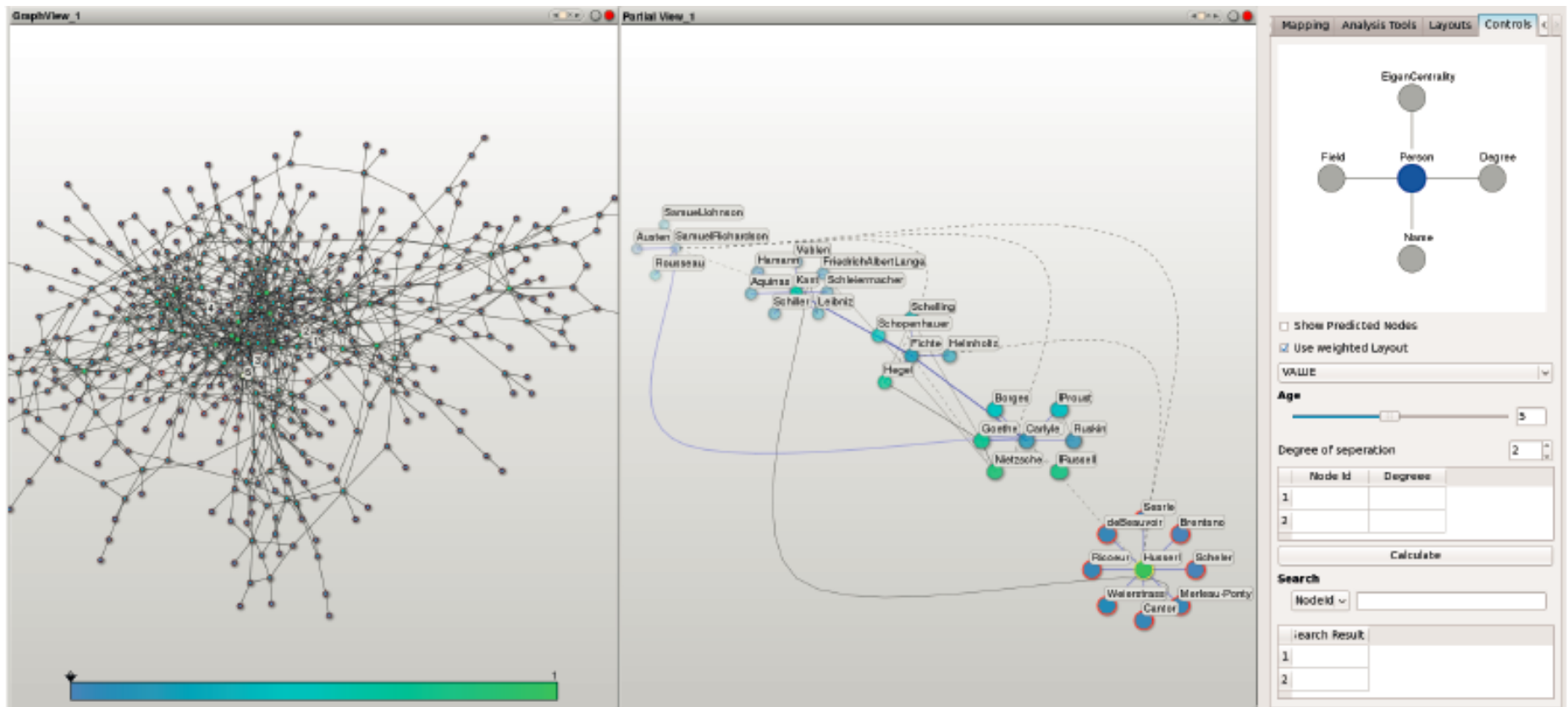


Figure 1. A screenshot of the Opinion Space 1.0 interactive map. Each point corresponds to a user and comment. The point with the halo indicates the position of the active user; green points correspond to comments rated positively by the active user, and red points correspond to comments rated negatively. Larger and brighter points are associated with the comments that are rated more positively by the user community.

Related work on Visual RS - 5

- 2011: *Visual Recommendations for Network Navigation*. IEEE Symposium on Visualization . Tarik Crnovrsanin, Isaac Liao, Yingcai Wu, Kwan-Liu Ma
- Build on top of netzen: <http://vis.cs.ucdavis.edu/~correac/netzen/index.html>



Related work on Visual RS - 6

- 2011: SFViz: interest-based friends exploration and recommendation in social networks SFVIZ (VINCI 2011)
- Gou, You (?) et al.



Figure 14. Friendship patterns at the top level in the tag tree.

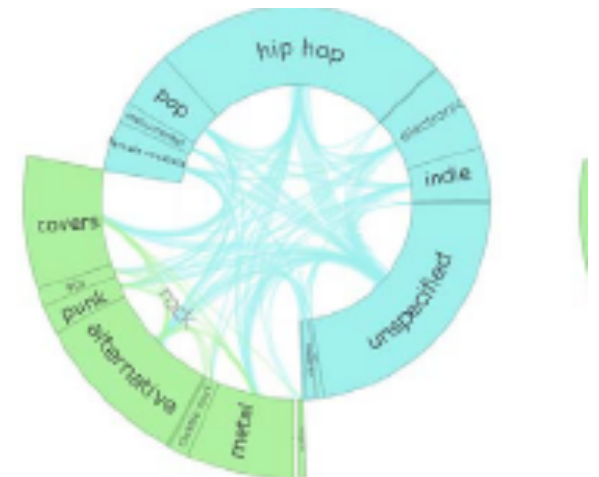


Figure 15. A cross-scale view of category under "rock" with other category from the first level.



Figure 17. A social network of a center user all levels with DOI = 1.

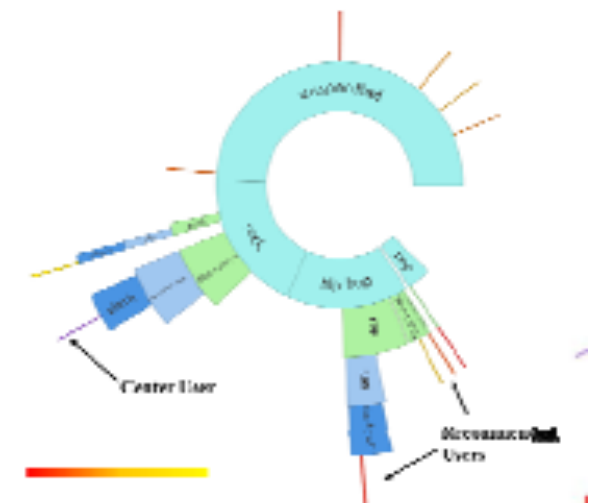
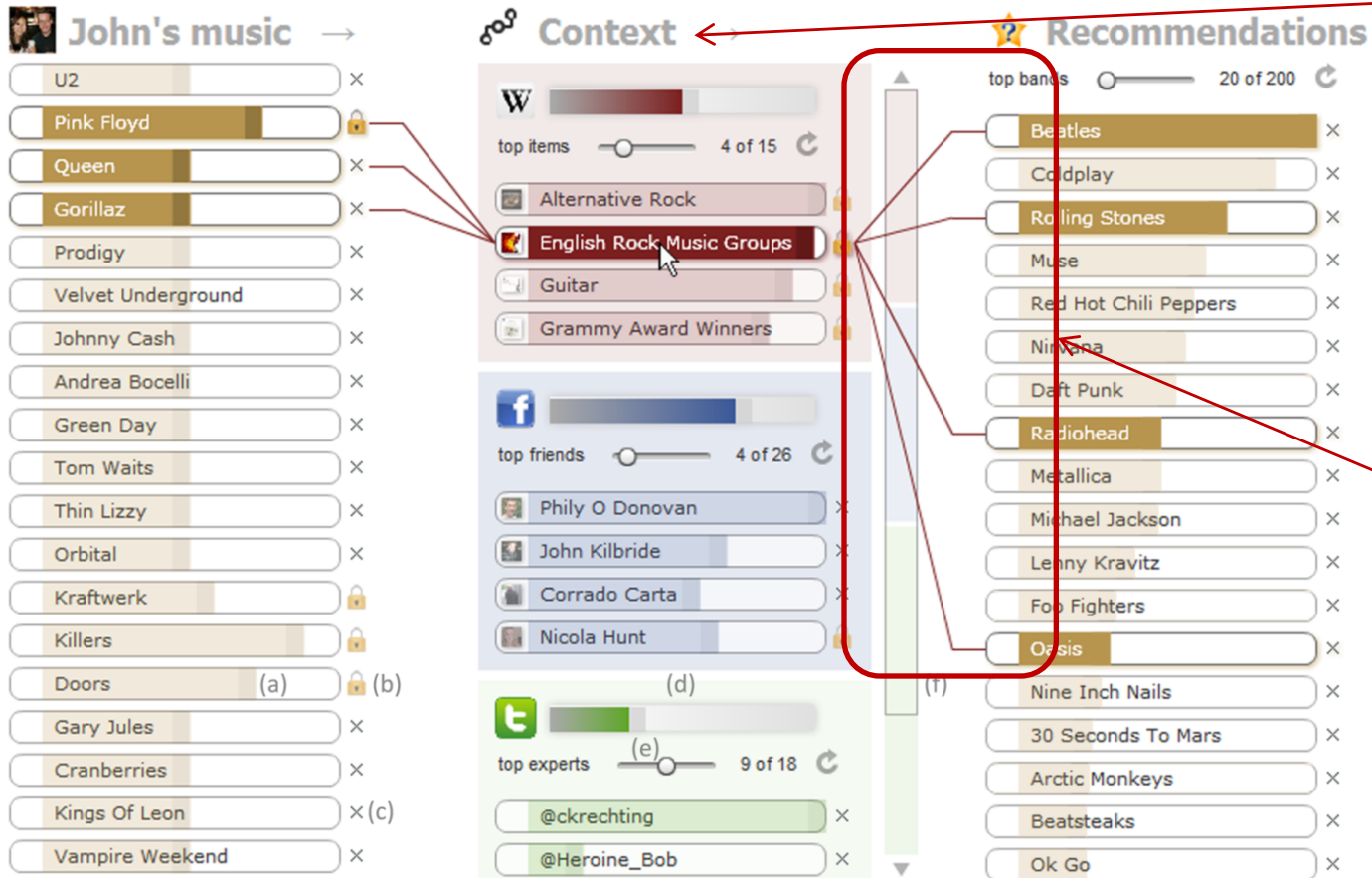


Figure 18. Top 10 recommended friends without a category of interest.

Related work on Visual RS - 8



Controllability:
Sliders that let users control the importance of preferences and contexts

Inspectability: lines that connect recommended items with contexts and user preferences

Related work on Visual RS - 8

- SetFusion
- Denis Parra, Peter Brusilovsky, and Christoph Trattner. 2014. See what you want to see: visual user-driven approach for hybrid recommendation (IUI 2014)

The screenshot displays the SetFusion interface, which is divided into three main sections: (a), (b), and (c).

(b) Tune weights of the recommender methods: This section allows users to adjust the weights of three recommendation methods using sliders. The current settings are: Most bookmarked papers (0.4), Similar to your favorite articles (0.8), and Frequently cited authors in ACM DL (0.4). An "Update Recommendation List" button is located below the sliders.

(c) Similar to your favorite articles: This section features a Venn diagram with three overlapping circles representing the recommendation methods. The top circle is yellow and labeled "Articles in top30", the bottom-left circle is blue and labeled "Articles not in top30", and the bottom-right circle is red. A tooltip is visible over the intersection of the yellow and red circles, displaying the article title "2. Can't see the forest for the trees? A citation recommendation system".

(a) List of recommended articles: This section shows a list of 16 articles with their titles, authors, and a "see abstract" link. The top article is "2. Can't see the forest for the trees? A citation recommendation system" by C. Lee Giles, Cornelia Caragea, Adrian Silvescu, and Prasenjit Mitra.

SetFusion: A Controllable Hybrid Recommender

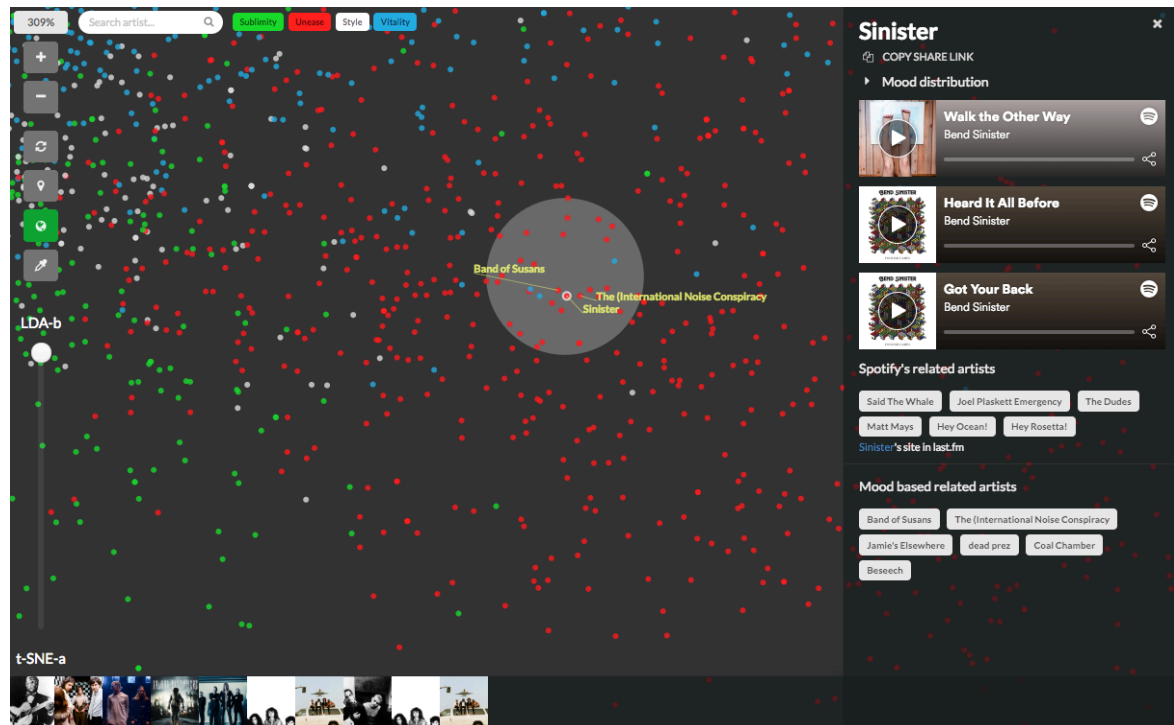
Parra, D., Brusilovsky, P., Trattner, C.

IUI 2014, Haifa, Israel

<https://www.youtube.com/watch?v=9LwSx1V6Yxk>

Related work on Visual RS - 9

- Moodplay
- Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2016. Moodplay: Interactive Mood-based Music Discovery and Recommendation. (UMAP 2016)



<http://moodplay.pythonanywhere.com/>

<https://www.youtube.com/watch?v=eEdo32oOmcE>

Controlabilidad

TasteWeights

¿Por qué controlabilidad?

- Beyond prediction accuracy, transparency and explainability in **#recsys** have proved to be related to user satisfaction.
- Studies show an effect of controllability on user satisfaction (papers I, II, III) ~ now the details are still not completely clear
- What has not been studied?
 - Insights from our TalkExplorer studies (submitted to IUI)

Paper I

Bart P. Knijnenburg, Niels J.M. Reijmer, and Martijn C. Willemsen. 2011. **Each to his own: how different users call for different interaction methods in recommender systems.** In *Proceedings of the fifth ACM conference on Recommender systems (RecSys '11)*.

Paper I

- Recommender for Energy-saving measures
- **Main message:** Controllability matters, but mainly for experts. For novices, a TopN recommendation without too much control led to better user satisfaction

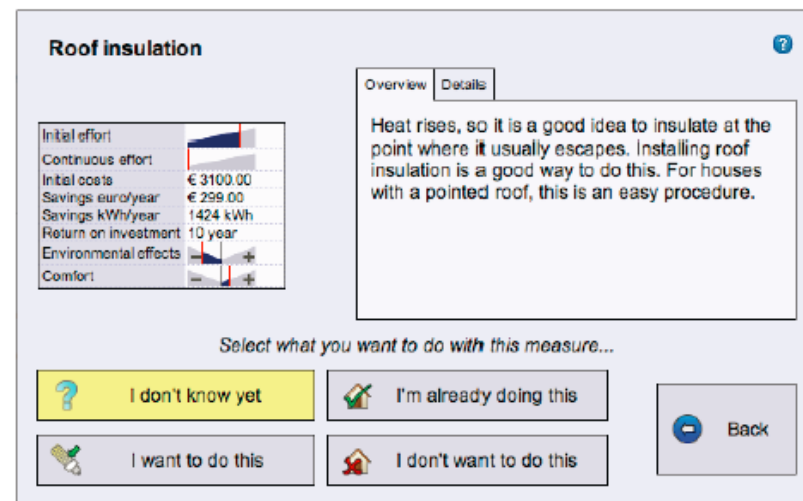


Figure 2. Screen shown to users when they click on an item

Paper II

- Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012.
Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems (RecSys '12)*.

Paper II

- Study on **TasteWeights**: New System introduced at RecSys 2012
- Facebook music recommender
- Gives user controls and explains how they came about
- Study with 267 (recruited in craigslist and mechanical turk)

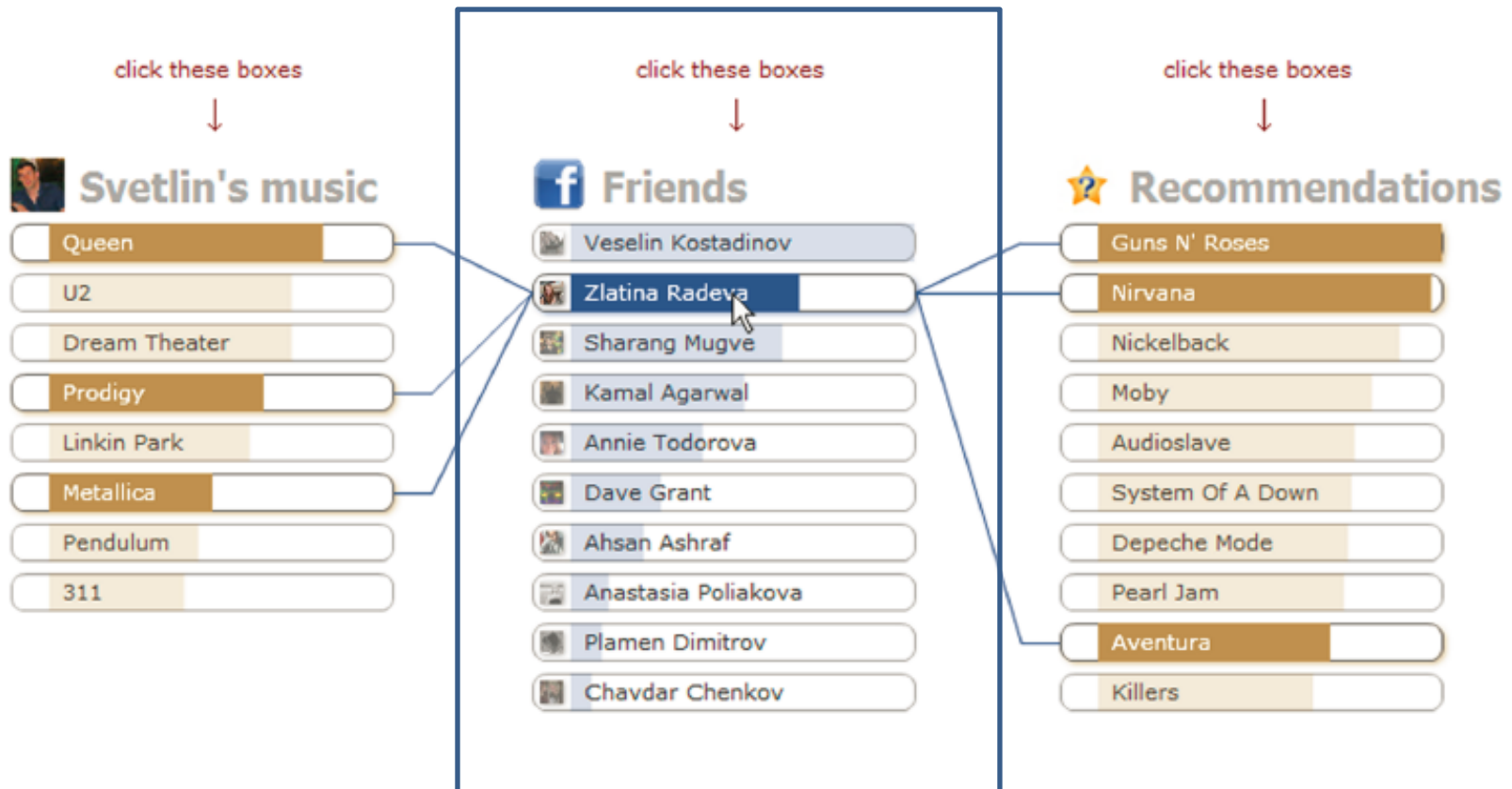
Paper II



Instructions

Inspectability

- By clicking on the boxes below, you can see how your likes are linked to your friends, and how your friends are linked to the recommendations.
- Please carefully inspect the visualization and the recommendations by clicking on the boxes below.
- When you are done, click "Next".



Paper II

- Summary of Results
 - Positive effects of inspectability and control, but several nuances
 - In the full graph condition, people “recognize” more recommendation, leading to better trust but lower system satisfaction (diff than recomm. Quality)
- Personal Characteristics:
 - Trusting propensity positively correlated with user satisfaction
 - Music experts feel less in control (bands to filter might be too rough) but have an overall positive perception of the system

Paper III

- Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. 2012. **The relation between user intervention and user satisfaction for information recommendation.** In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12)*

Paper III

- Terms: User Intervention instead of Control
- Study on Music Recommendation, 84 users
- Methods of user intervention
 - Rating: usual explicit feedback
 - (CI) Context Input: When / Where / With Whom
 - (CAS) Context attribute selection: country, gender, sex, unit, year
 - (PE) Profile Editing: not clear, but the highest level of intervention

Paper III

- Condition: 100 songs used for learning, 1000 for testing (experiment itself)
- 1st step: gather data from user to build recommendations
- 2nd step: randomly assign to each user 2 of the conditions: ratings, CI, CAS, PE

Paper III - results

- “... Therefore, results show that the changes of recommendation results by user interventions improve the precision...”

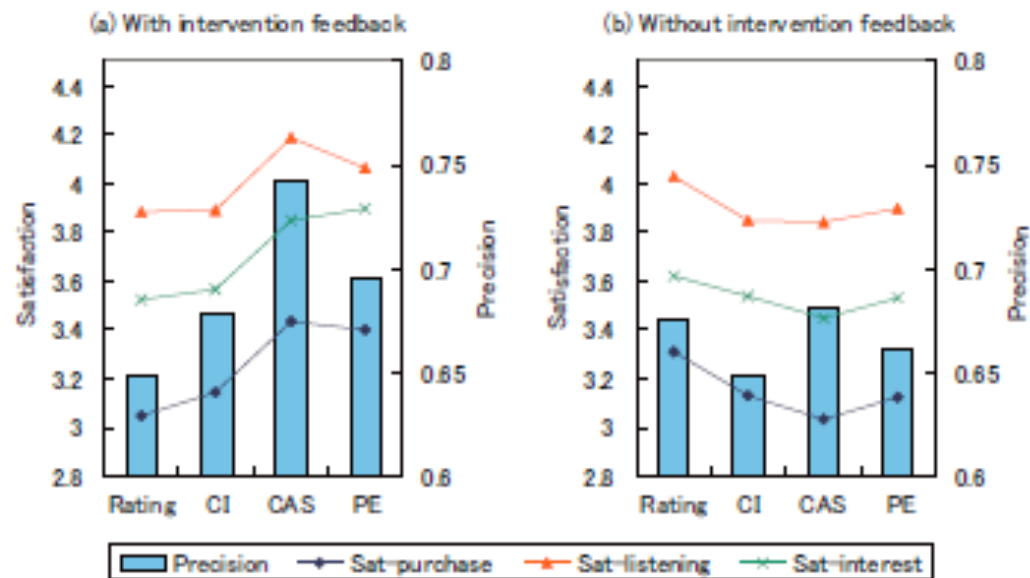


Figure 3: Relation between user intervention, precision and user satisfaction

Paper III - results

- Considering group of people with feedback effect of interest degree

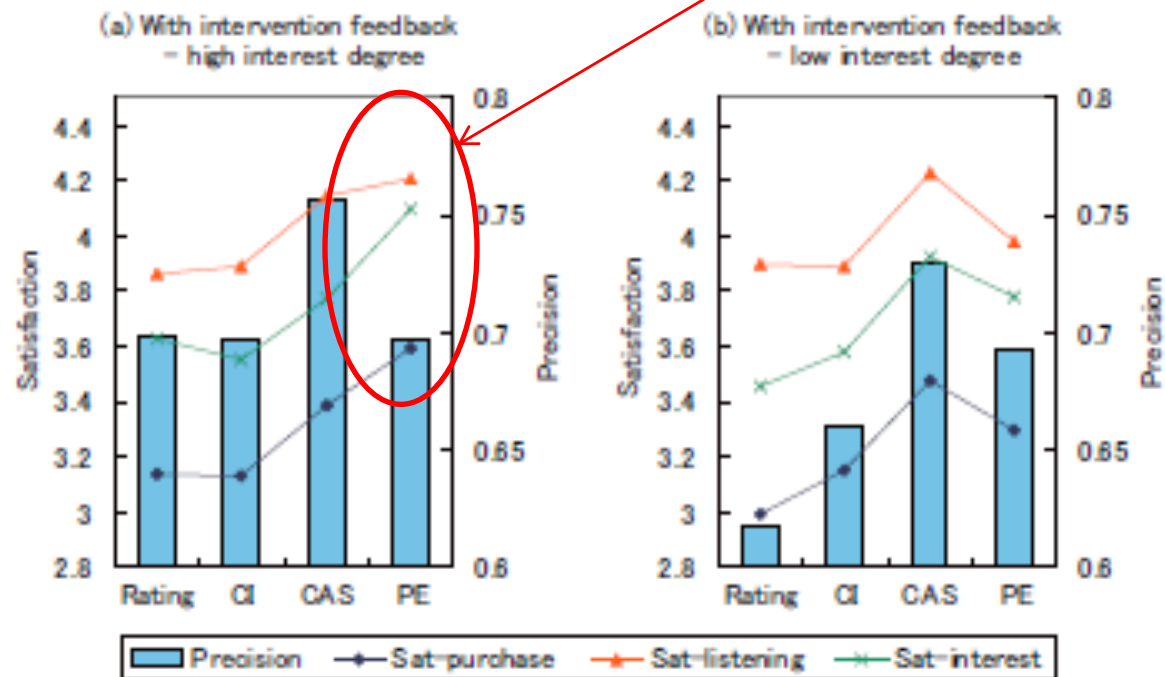


Figure 4: Relation between user intervention, precision and user satisfaction in the group with intervention feedback

Summary paper III

- When system recommends items with high precision to users with high interest in music, the more the user intervenes -> the better the user satisfaction
- **NEVERTHELESS**, It is still unclear whether user intervention itself influences user satisfaction

PAWS insights

- Ahn, Jae-wook and Brusilovsky, Peter and Grady, Jonathan and He, Daqing and Syn, Sue Yeon. 2007. **Open user profiles for adaptive news systems: help or harm?** WWW 2007
- Verbert, Parra, Brusilovsky. 2013. **Visualizing Recommendations to Support Exploration, Transparency and Controllability**

Talk Explorer

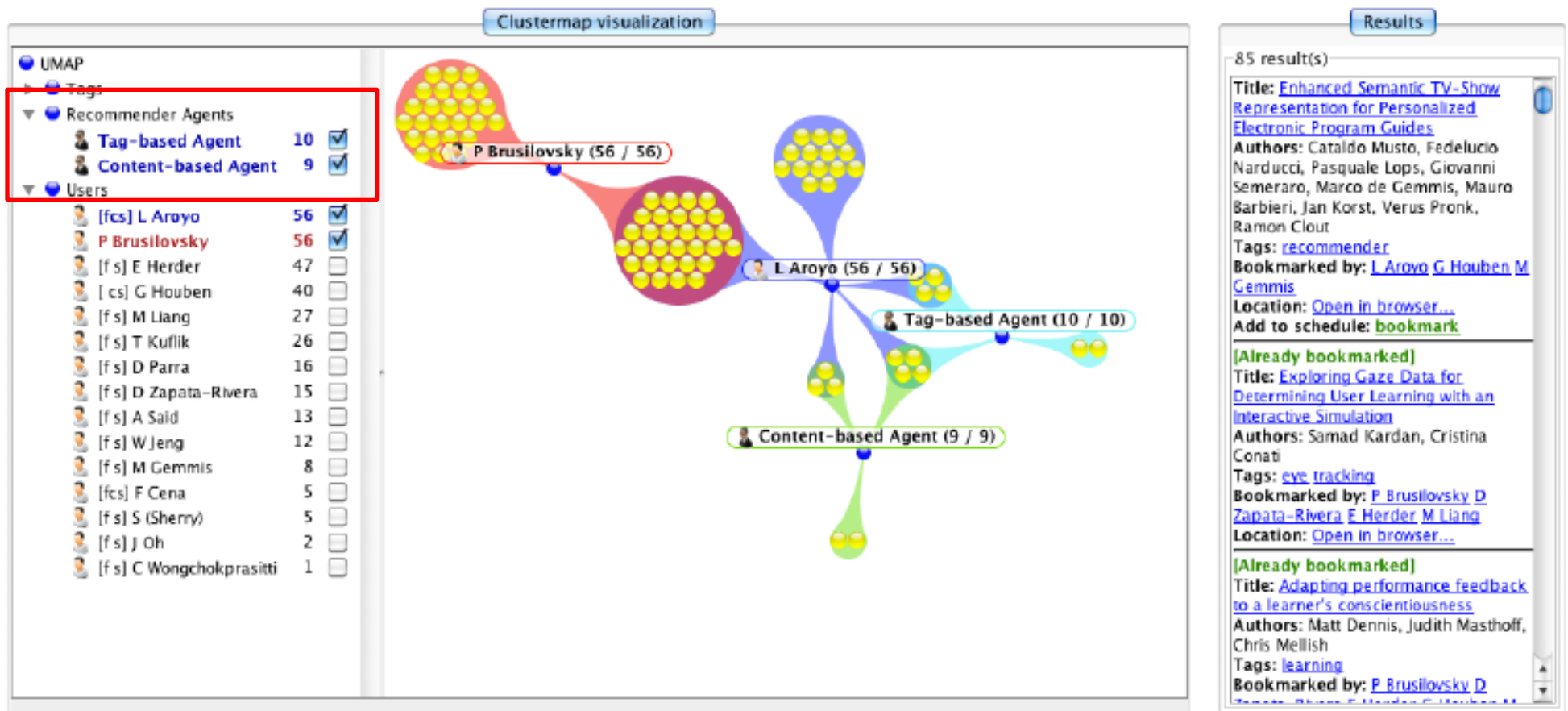
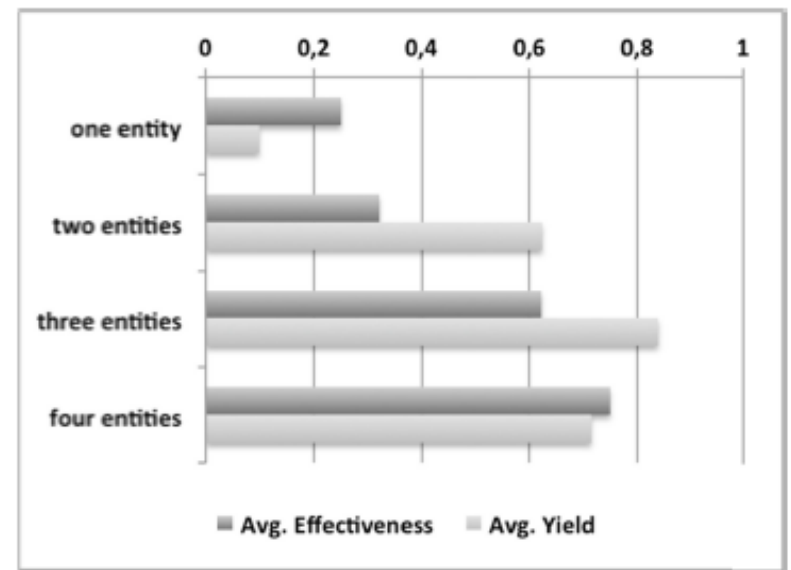
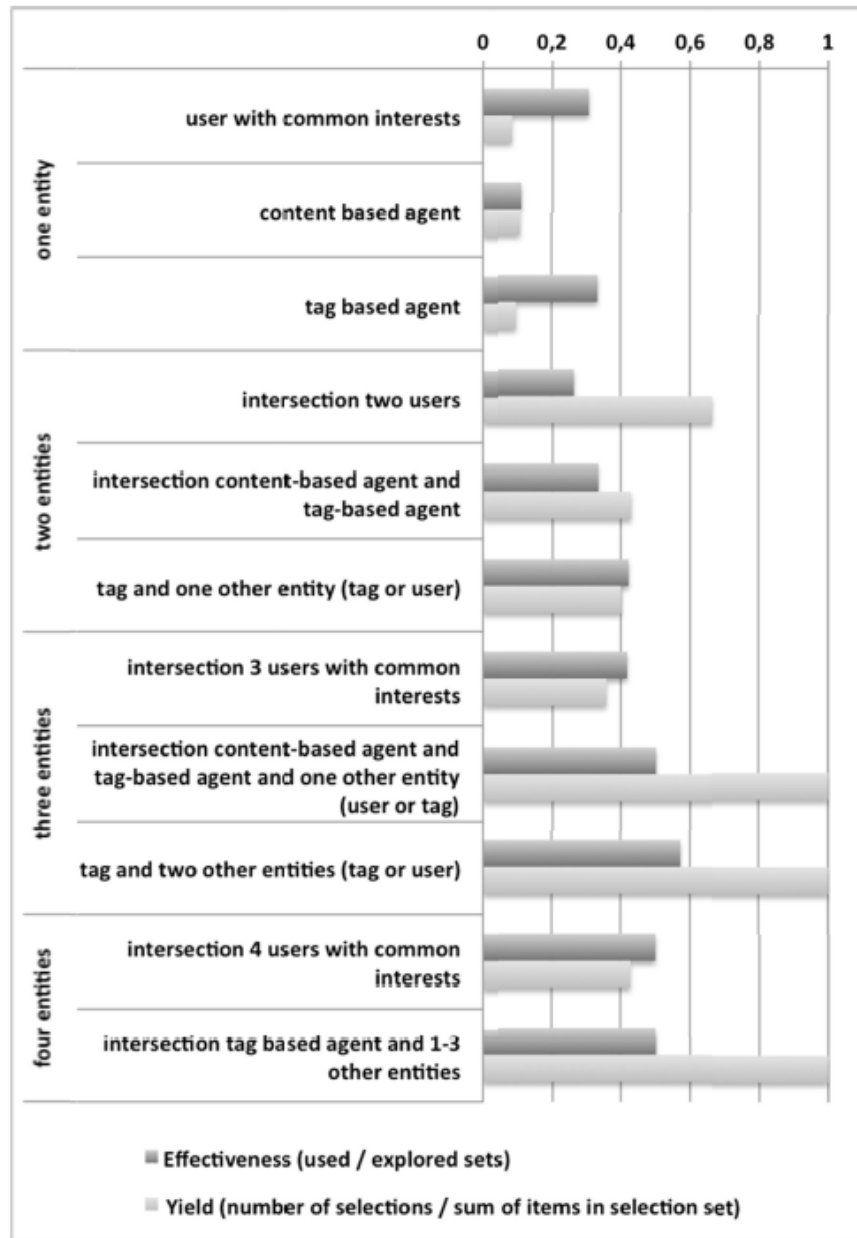


Figure 2: TalkExplorer

Talk Explorer

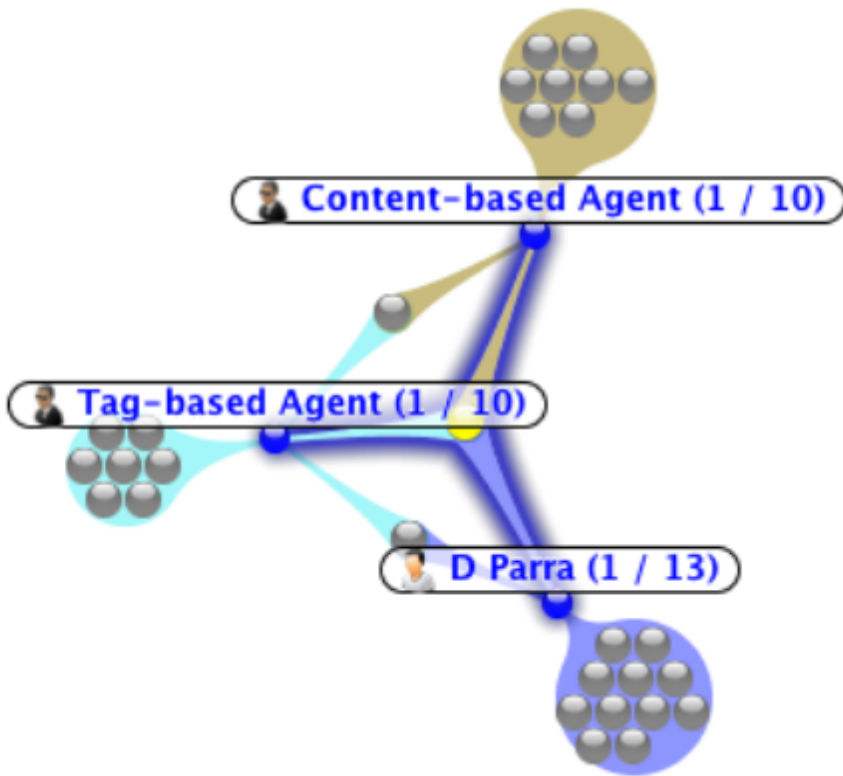


(b)

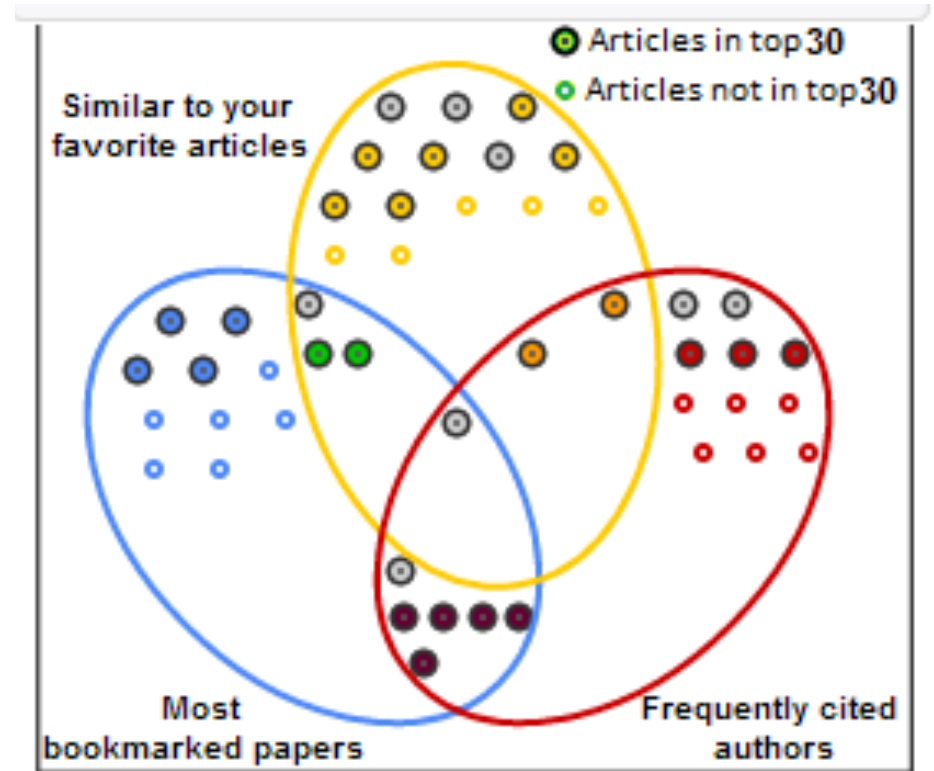
SetFusion vs. TalkExplorer

Drawback: Visualizing Intersections

- **Venn diagram:** more natural way to visualize intersections



Clustermap



Venn diagram

Evaluation: Intersections & Effectiveness

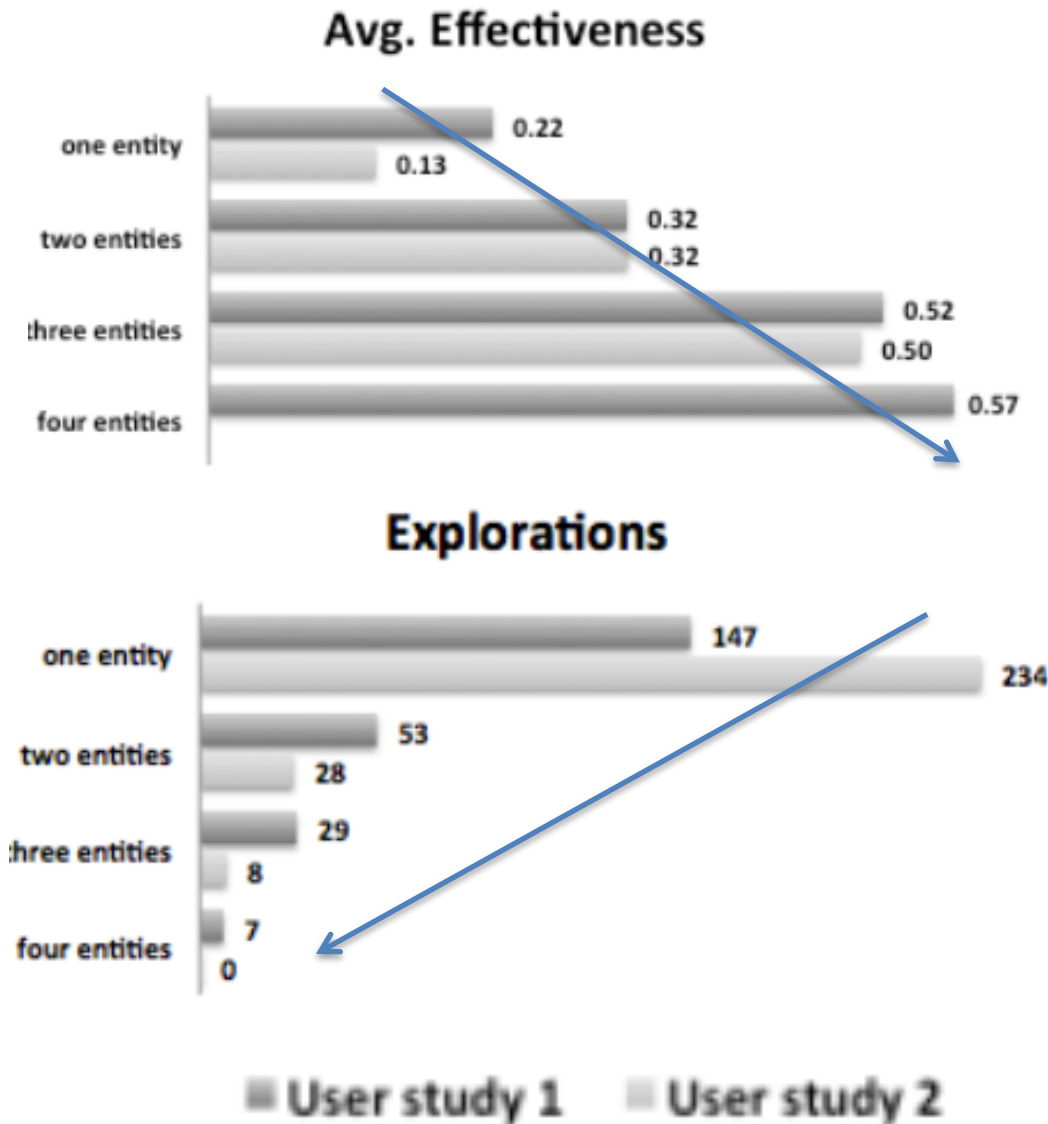
- What do we call an “Intersection”?



- We used #explorations on intersections and their effectiveness, defined as:

$$effectiveness = \frac{|bookmarked\ items|}{|intersections\ explored|}$$

Results of Studies I & II



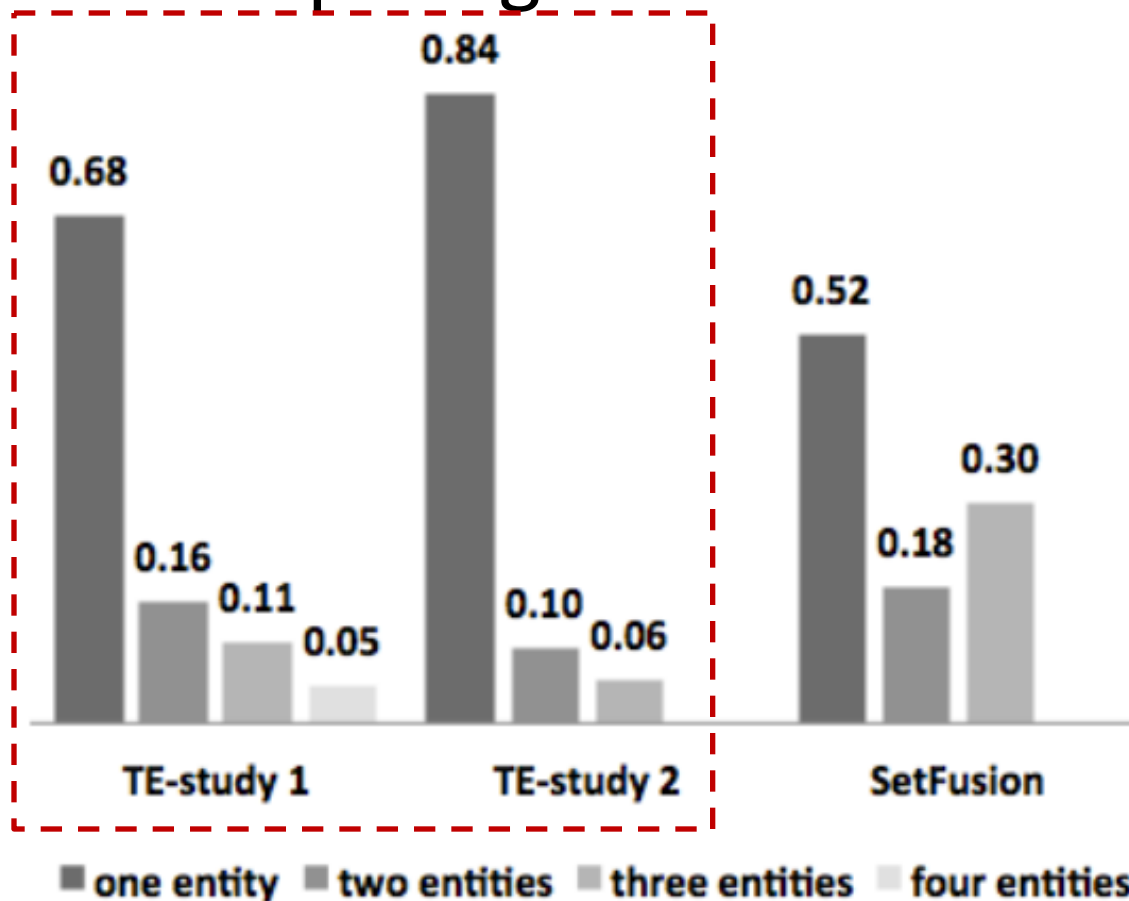
- Effectiveness increases with intersections of more entities
- Effectiveness wasn't affected in the field study (study 2)
- ... but exploration distribution was affected

More Details About TalkExplorer

- Verbert, K., Parra, D., Brusilovsky, P., & Duval, E. (2013). **Visualizing recommendations to support exploration, transparency and controllability.** In Proceedings of the 2013 international conference on Intelligent user interfaces (pp. 351-362). ACM.
- Verbert, K., Parra, D., & Brusilovsky, P. (2016). **Agents Vs. Users: Visual Recommendation of Research Talks with Multiple Dimension of Relevance.** ACM Transactions on Interactive Intelligent Systems (TiiS), 6(2), 11.

TalkExplorer vs. SetFusion

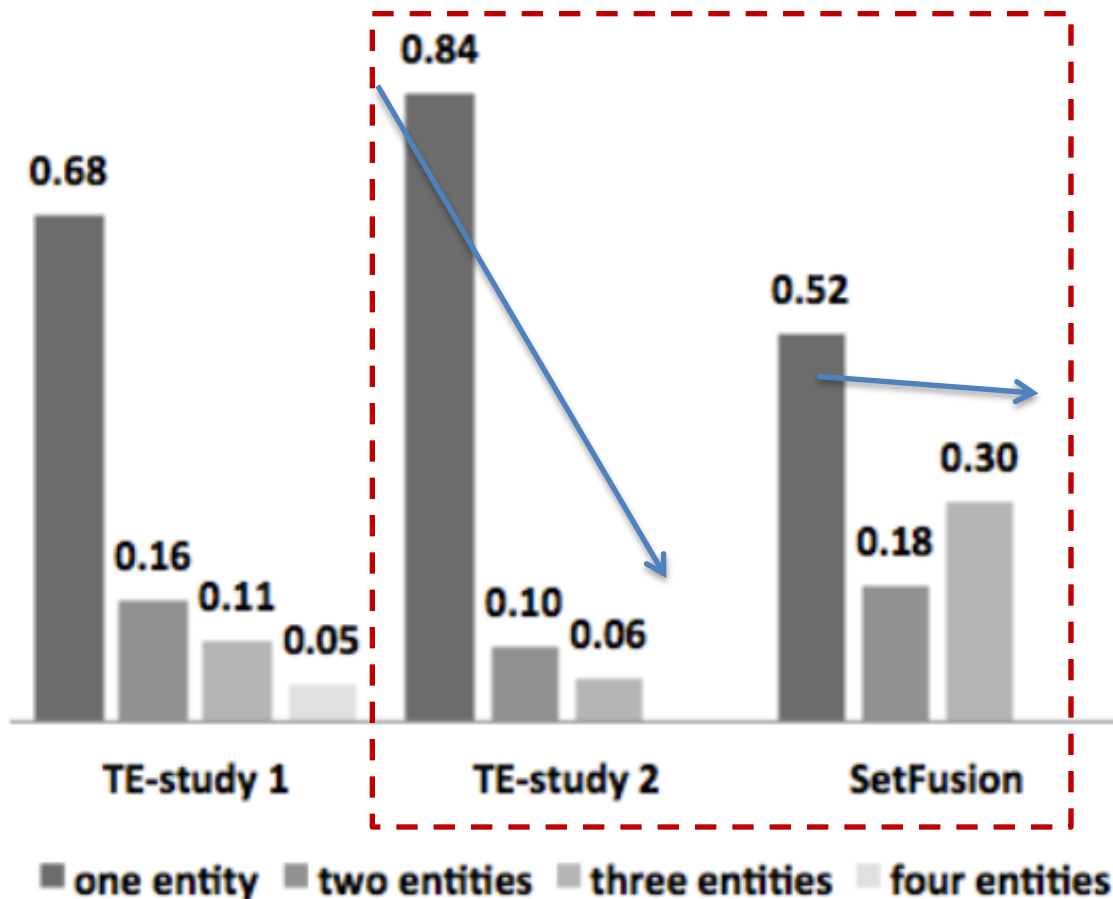
- Comparing distributions of explorations



In studies 1 and 2 over talkEplorer we observed an important change in the distribution of explorations.

TalkExplorer vs. SetFusion

- Comparing distributions of explorations



- Comparing the field studies:
- In TalkExplorer, 84% of the explorations over intersections were performed over clusters of 1 item
 - In SetFusion, was only 52%, compared to 48% (18% + 30%) of multiple intersections, diff. not statistically significant



PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Sistemas Recomendadores

Evaluación centrada en el usuario II: Frameworks de Recomendación

Denis Parra

IIC3633 – Sistemas Recomendadores

OutLine

- Evaluación centrada en el Usuario:
 - Xiao y Benbasat: Resumen de estudios empíricos sobre “Agentes de Recomendación”
 - Framework I: Resque (Pearl Pu)
 - Framework II: Knijnenburg et al.

Frameworks de Evaluación Centrada en el Usuario

- Xiao y Benbasat (MIS Quartely paper) 2007 (act. 2012)
 - Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: use, characteristics, and impact. *Mis Quarterly*, 31(1), 137-209.
- Pearl Pu (ResQue) – 2011
 - Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 157-164). ACM.
- Bart Knijnenburg – 2012
 - Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5), 441-504.

Xiao y Benbasat

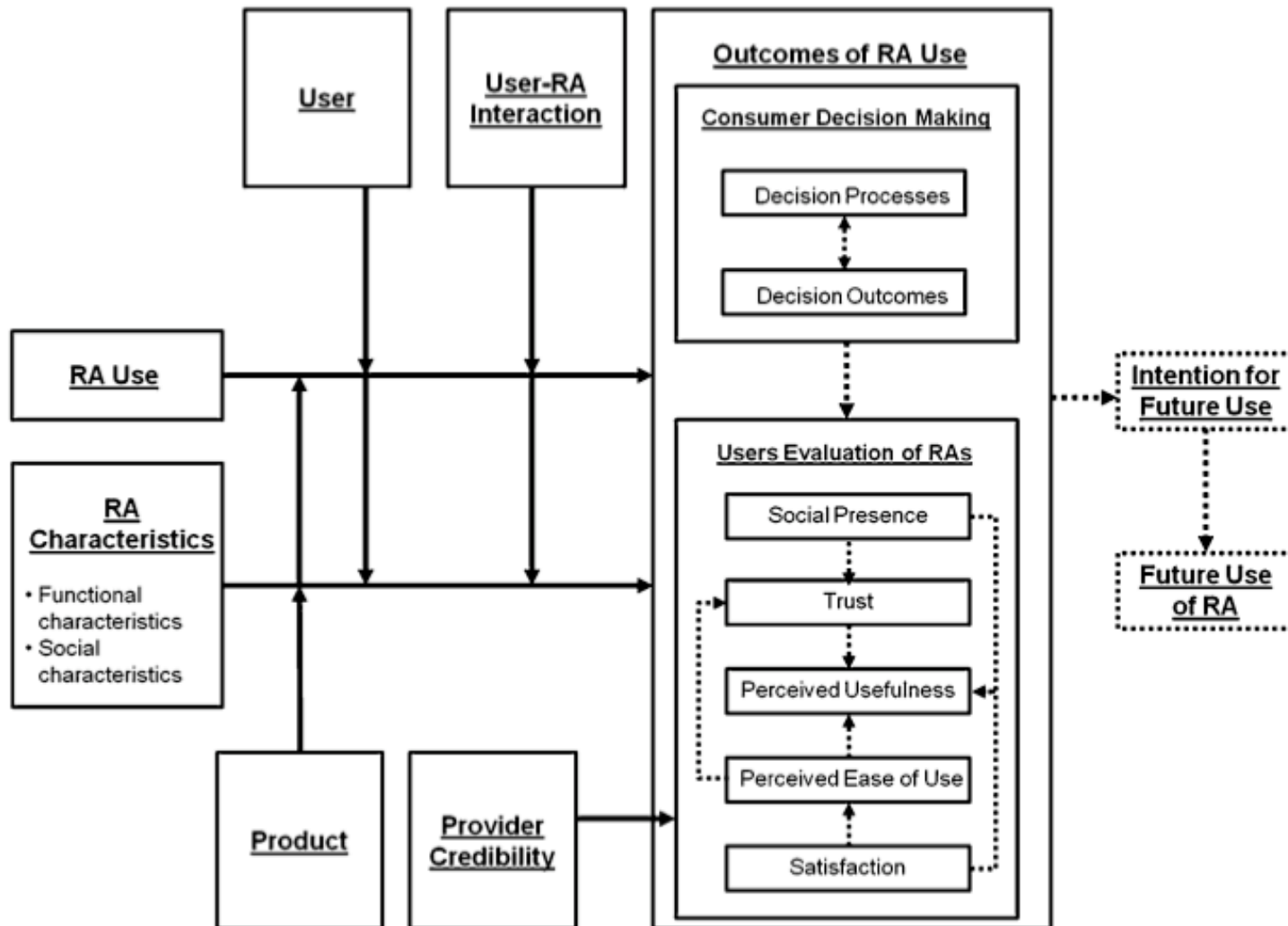


Fig. 2 Updated conceptual model

Resumen de más de 20 estudios

Paper	Type of study Type of RA	Independent variables	Dependent variables
Chang and Chin (2010)	Experiment (lab) RA for mini-notebooks	Recommendation sources: word of mouth (WOM), advertising, or recommendation systems Gender (moderator) Perceived risk (moderator)	Intention to purchase online

Major areas addressed

Major findings

RA use compared to the use of advertising or WOM

A positive recommendation by WOM led to a stronger increase in willingness to purchase online than did advertising and recommendation systems
The effect of WOM, advertising, and recommendation systems on online purchase intentions was greater for female consumers, who perceived higher risks in purchasing.

Resumen de más de 20 estudios

Paper	Type of study Type of RA	Independent variables	Dependent variables
Wang and Doong (2010aa)	Experiment (lab) RA for eBooks	Argument form (claim only, claim plus data and warrant, and claim plus data and backing) Spokesperson type (Web itself, expert, customer)	Argument quality Source credibility Purchase intention
Major areas addressed		Major findings	
RA output characteristics → explanation		<p>Customers' perceptions of the argument quality and source credibility of the RA's recommendations were found to effectively influence their purchase intentions at the Webstore</p> <p>Customers' perceptions of argument quality and source credibility differed significantly as a result of the varied argument forms</p> <p>Although the various spokesperson types generated significantly different levels of source credibility, argument quality remained unchanged</p>	

Framework I - ResQue

- Identifica qué variables (constructos) definen la experiencia de un usuario con un sistema recomendador
- Desarrollado en base a modelos existentes para evaluar (TAM y SUMI) y a resultados de estudios relacionados
 - TAM: perceived ease of use of a system, its perceived usefulness and users' intention to use the system
 - TAM v2 (UTAUT): performance expectancy, effort expectancy, social influence, and facilitating conditions
 - SUMI (Software Usability Measurement Inventory) : efficiency, affect, helpfulness, control, learnability

Framework I - ResQue

- Identifica qué variables definen la experiencia de un usuario con un sistema recomendador

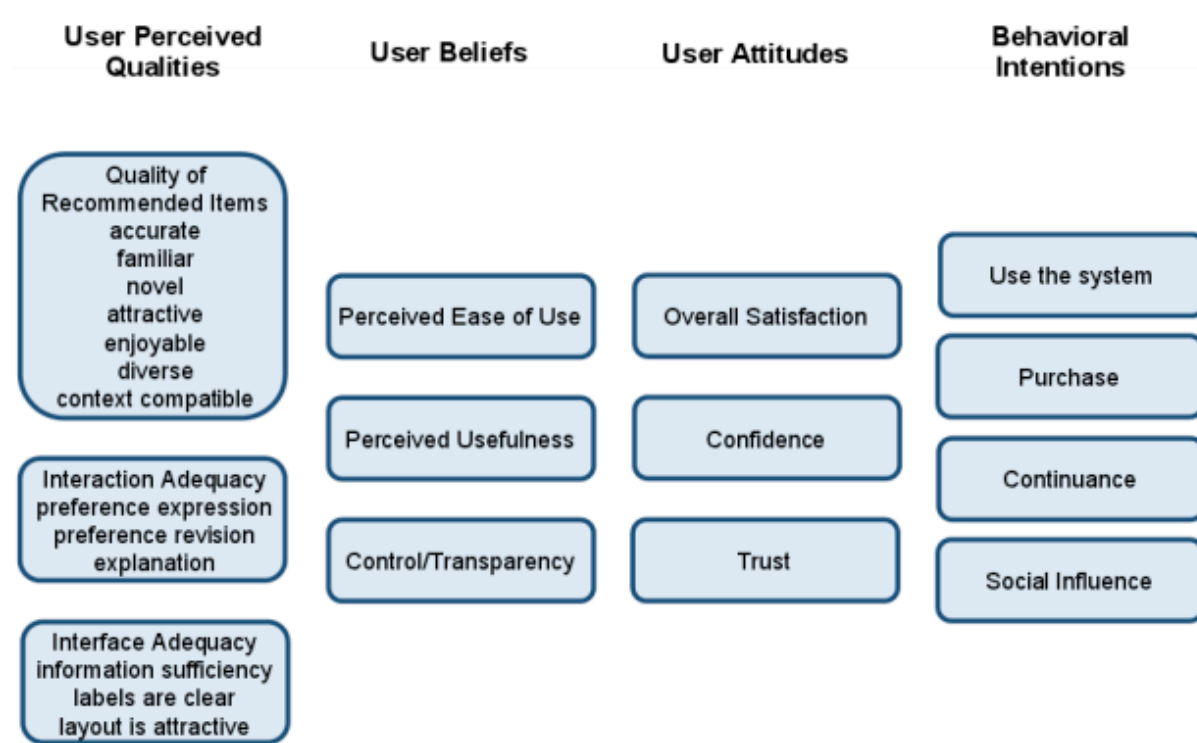





Figure 1: Constructs of an Evaluation Framework on the Perceived Qualities of Recommenders (ResQue).



Encuesta

A1. Quality of Recommended Items



A.1.1 Accuracy

-  The items recommended to me matched my interests.*
-  The recommender gave me good suggestions.
-  I am not interested in the items recommended to me (reverse scale).

A.1.2 Relative Accuracy

-  The recommendation I received better fits my interests than what I may receive from a friend.
-  A recommendation from my friends better suits my interests than the recommendation from this system (reverse scale).

A.1.3 Familiarity

-  Some of the recommended items are familiar to me.
-  I am not familiar with the items that were recommended to me (reverse scale).

Encuesta II


A.1.4 Attractiveness

 The items recommended to me are attractive.

A.1.5 Enjoyability

 I enjoyed the items recommended to me.

A.1.6 Novelty

 The items recommended to me are novel and interesting.*

 The recommender system is educational.

 The recommender system helps me discover new products.

 I could not find new items through the recommender (reverse scale).

A.1.6 Diversity

 The items recommended to me are diverse.*

 The items recommended to me are similar to each other (reverse scale).*

Encuesta III

A.1.7 Context Compatibility

- 🎬 I was only provided with general recommendations.
- 🎬 The items recommended to me took my personal context requirements into consideration.
- 🎬 The recommendations are timely.

A2. Interaction Adequacy

- 🎬 The recommender provides an adequate way for me to express my preferences.
- 🎬 The recommender provides an adequate way for me to revise my preferences.
- 🎬 The recommender explains why the products are recommended to me.*

A3. Interface Adequacy



- 🎬 The recommender's interface provides sufficient information.
- 🎬 The information provided for the recommended items is sufficient for me.
- 🎬 The labels of the recommender interface are clear and adequate.
- 🎬 The layout of the recommender interface is attractive and adequate.*

Encuesta IV




A4. Perceived Ease of Use

A.4.1 Ease of Initial Learning






I became familiar with the recommender system very quickly.

-  I easily found the recommended items.
-  Looking for a recommended item required too much effort (reverse scale).

A.4.2 Ease of Preference Elicitation

-  I found it easy to tell the system about my preferences.
-  It is easy to learn to tell the system what I like.
-  It required too much effort to tell the system what I like (reversed scale).

A.4.3 Ease of Preference Revision

-  I found it easy to make the system recommend different things to me.
-  It is easy to train the system to update my preferences.
-  I found it easy to alter the outcome of the recommended items due to my preference changes.
-  It is easy for me to inform the system if I dislike/like the recommended item.
-  It is easy for me to get a new set of recommendations.

Encuesta V

A.4.4 Ease of Decision Making

- 🎬 Using the recommender to find what I like is easy.
- 🎬 I was able to take advantage of the recommender very quickly.
- 🎬 I quickly became productive with the recommender.
- 🎬 Finding an item to buy with the help of the recommender is easy.*
- 🎬 Finding an item to buy, even with the help of the recommender, consumes too much time.

A5. Perceived Usefulness

- 🎬 The recommended items effectively helped me find the ideal product.*
- 🎬 The recommended items influence my selection of products.
- 🎬 I feel supported to find what I like with the help of the recommender.*
- 🎬 I feel supported in selecting the items to buy with the help of the recommender.

Encuesta VI

A6. Control/Transparency

- 🎬 I feel in control of telling the recommender what I want.
- 🎬 I don't feel in control of telling the system what I want.
- 🎬 I don't feel in control of specifying and changing my preferences (reverse scale).
- 🎬 I understood why the items were recommended to me.
- 🎬 The system helps me understand why the items were recommended to me.
- 🎬 The system seems to control my decision process rather than me (reverse scale).

A7. Attitudes

- 🎬 Overall, I am satisfied with the recommender.*
- 🎬 I am convinced of the products recommended to me.*
- 🎬 I am confident I will like the items recommended to me. *
- 🎬 The recommender made me more confident about my selection/decision.
- 🎬 The recommended items made me confused about my choice (reverse scale).
- 🎬 The recommender can be trusted.

Encuesta VII

A8. Behavioral Intentions

A.8.1 Intention to Use the System

 If a recommender such as this exists, I will use it to find products to buy.

A.8.2 Continuance and Frequency

 I will use this recommender again.*

 I will use this type of recommender frequently.

 I prefer to use this type of recommender in the future.

A.8.3 Recommendation to Friends

 I will tell my friends about this recommender.*

A.8.4 Purchase Intention

 I would buy the items recommended, given the opportunity.*

Framework II

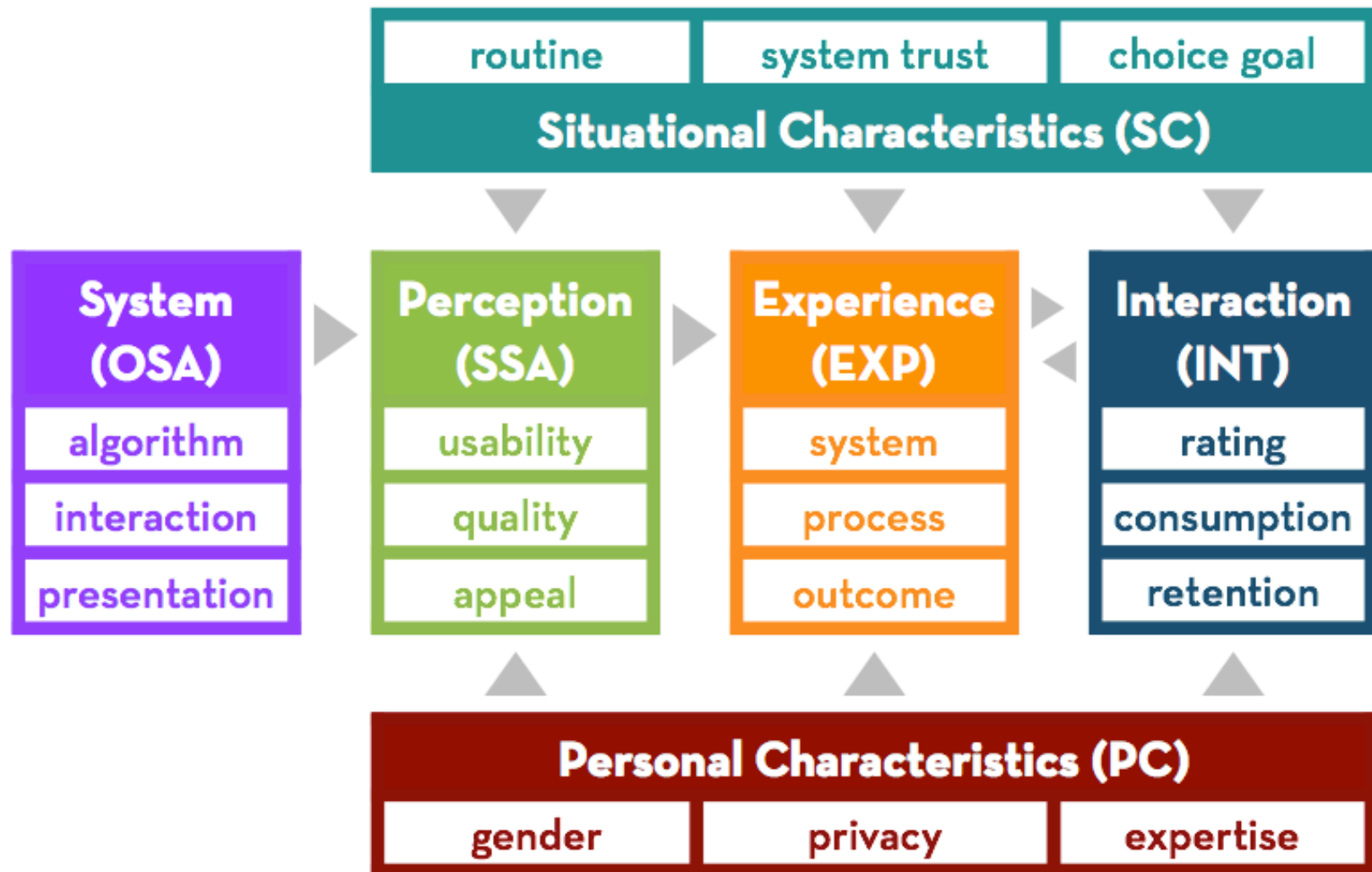


Fig. 1 An updated version of the User-Centric Evaluation Framework [61].

Knijnenburg et al.

- En este modelo, el evaluador debe identificar las variables específicas y a qué dimensiones y/o categorías de aspectos correspondan.
- Una vez identificadas y medidas, se cotejan con el modelo estructural para ver si corresponden.

Ejemplo de Aplicación

- Estudio de TasteWeights: Inspectability & Controlability



Figure 1. The TasteWeights system as used in the online user experiment. This is the inspection phase of the “full graph” condition. Users can click on items, friends and recommendations to see the links between them. The inspection phase of the “list only” condition shows the rightmost list (recommendations) only.

Inspectability & Controlability

- Condiciones de Control

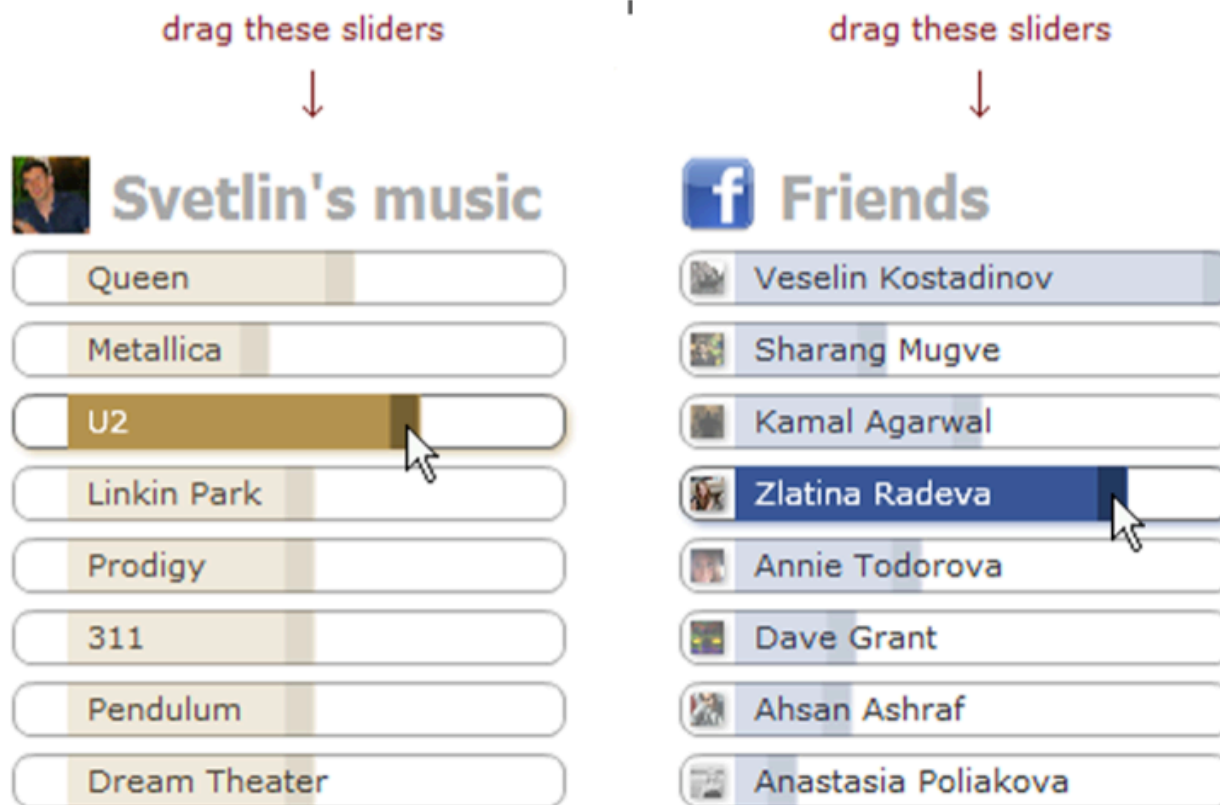


Figure 2. The control phase of item control (left) and friend control (right) conditions.

Inspectability & Controlability

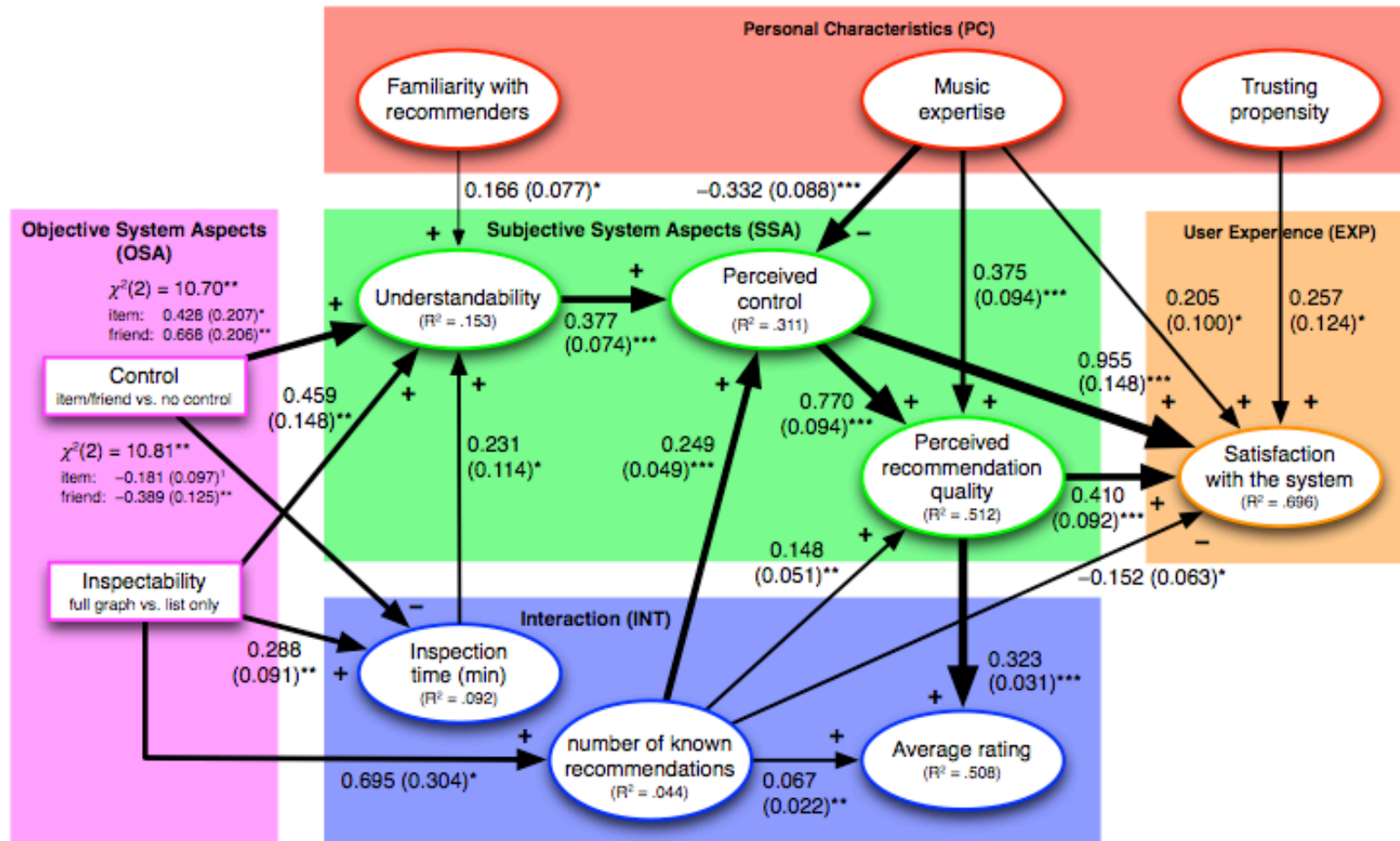


Figure 3. The structural equation model for the data of the experiment. Significance levels: $*** p < .001$, $** p < .01$, 'ns' $p > .05$. R^2 is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent the β coefficients (and standard error) of the effect. Factors are scaled to have an SD of 1.

Resultados

- Control e Inspectability tienen un efecto positivo sobre “Comprensión del Sistema” (understandability)
- “Comprensión del Sistema” influye a la vez sobre la “Percepción de Control” (PC) y la “Percepción de Calidad de las Recomendaciones” (PQR)
- PC y PQR influyen sobre la satisfacción final con el sistema

Efectos Marginales

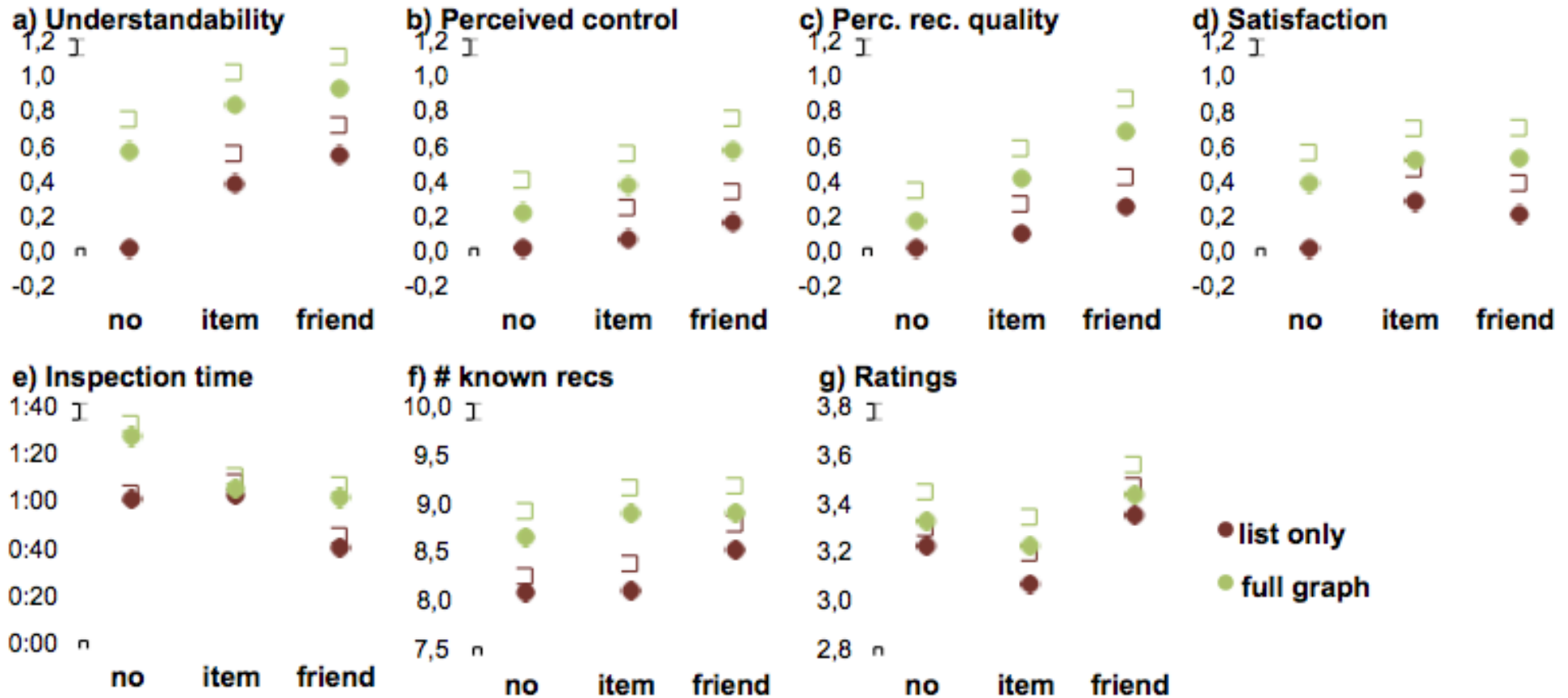


Figure 4. Marginal effects of inspectability and control on the subjective factors (top) and on behaviors (bottom). For the subjective factors, the effects of the “no control, list only” condition is set to zero, and the y-axis is scaled by the sample standard deviation.

Resumen de Resultados

- Visualización tipo “grafo de recomendación” mejora la experiencia del usuario al dejarlo inspeccionar las recomendaciones:
 - Comprensión, percepción de control, percepción de calidad de recomendación, satisfacción con el sistema
- Control sobre los pesos de “amigos” produce mayor efecto que control sobre los “items”
- Inspección y control son sumativos: puede incrementar escrutabilidad.

Proyecto Reciente

- Efecto de Explicabilidad y de Algoritmo sobre distintos aspectos de percepción de los usuarios respecto de recomendaciones de arte

Cheers!

@denisparra



Towards Explanations for Visual Recommender Systems of Artistic Images

Vicente Dominguez, Pablo Messina, Denis Parra, Christoph Trattner
CS Department
School of Engineering
Pontificia Universidad Católica de Chile

IntRS Workshop, October 7th 2018

Artwork Recommendation

- Online artwork market: Growing since 2008, despite global crises!
 - In 2011, art received \$11.57 billion in total global annual revenue, over \$2 billion versus 2010 (*forbes)
- Previous recommendation projects date for as long as 2007, such as the CHIP project to recommend paintings from Rijksmuseum.
- Little use of recent advances in Deep Neural Networks for Computer Vision.

[forbes] The World's Strongest Economy? the Global Art Market. <https://www.forbes.com/sites/abigailesman/2012/02/29/the-worlds-strongest-economy-the-global-art-market/> (2012)

Image Recommendation

- Since 2017 we have been working on recommending art images, using data from the online store UGallery.
- Two papers published:
 - DLRS 2017: Dominguez, V., Messina, P., Parra, D., Mery, D., Trattner, C., & Soto, A. (2017, August). Comparing Neural and Attractivenessbased Visual Features for Artwork Recommendation. In Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems(pp. 55-59). ACM.
 - UMUAI 2018: Messina, P., Dominguez, V., Parra, D., Trattner, C., & Soto, A. (2018). Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. User Modeling and User-Adapted Interaction, 1-40.

Data: UGallery

- Online Artwork Store, based on CA, USA.
- Mostly sales one-of-a-kind physical artwork.

Orientation —

Horizontal (496)

Vertical (162)

Square (145)

Size —

Height: 0" - 18"

0" 60"+

Width: 0" - 45"

0" 60"+

Medium —

Oil Painting (537)

Acrylic Painting (125)

Watercolor Painting (116)

Drawing Artwork (10)

Mixed Media Artwork (8)

Other Media (6)


Photography (1)

Style +

Color +


Price +

Sort By ▼




NEW

Oksana Johnson
14" x 11", oil painting
Evening Stroll: \$600




NEW

Suren Nersisyan
12" x 16", oil painting
Lake in the Mountains (Sunny Day):
\$400




NEW

Catherine McCargar
15" x 21", watercolor painting
Mt. Diablo, Port Costa View: \$825




NEW

Valerie Berkely
11" x 14", oil painting
Across Yellow Fields: \$300



NEW

Tami Cardnella
12" x 18", oil painting
Emerald Marsh: \$600

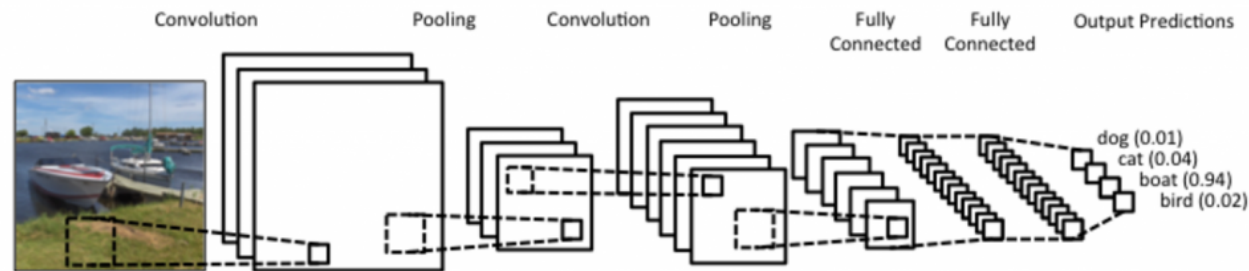


NEW

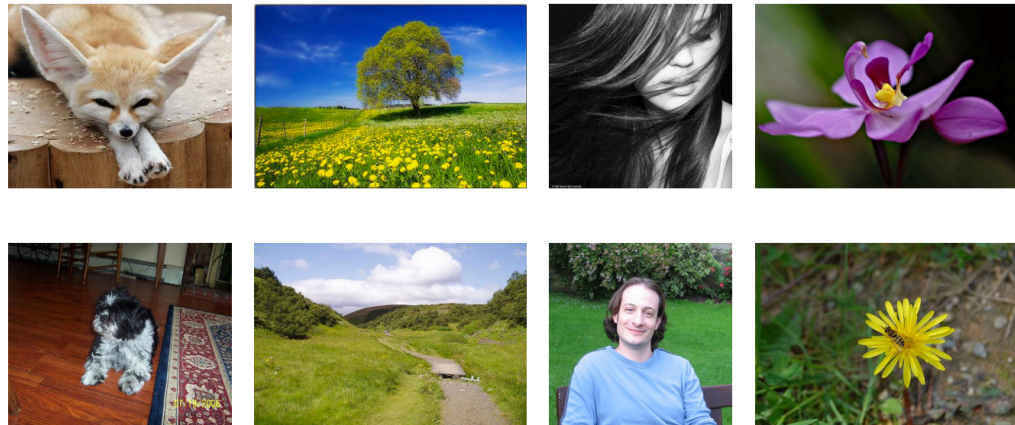
Tami Cardnella
18" x 24", oil painting
Sky Series #15: \$1725

Visual Features

- (DNN) Deep Neural Networks



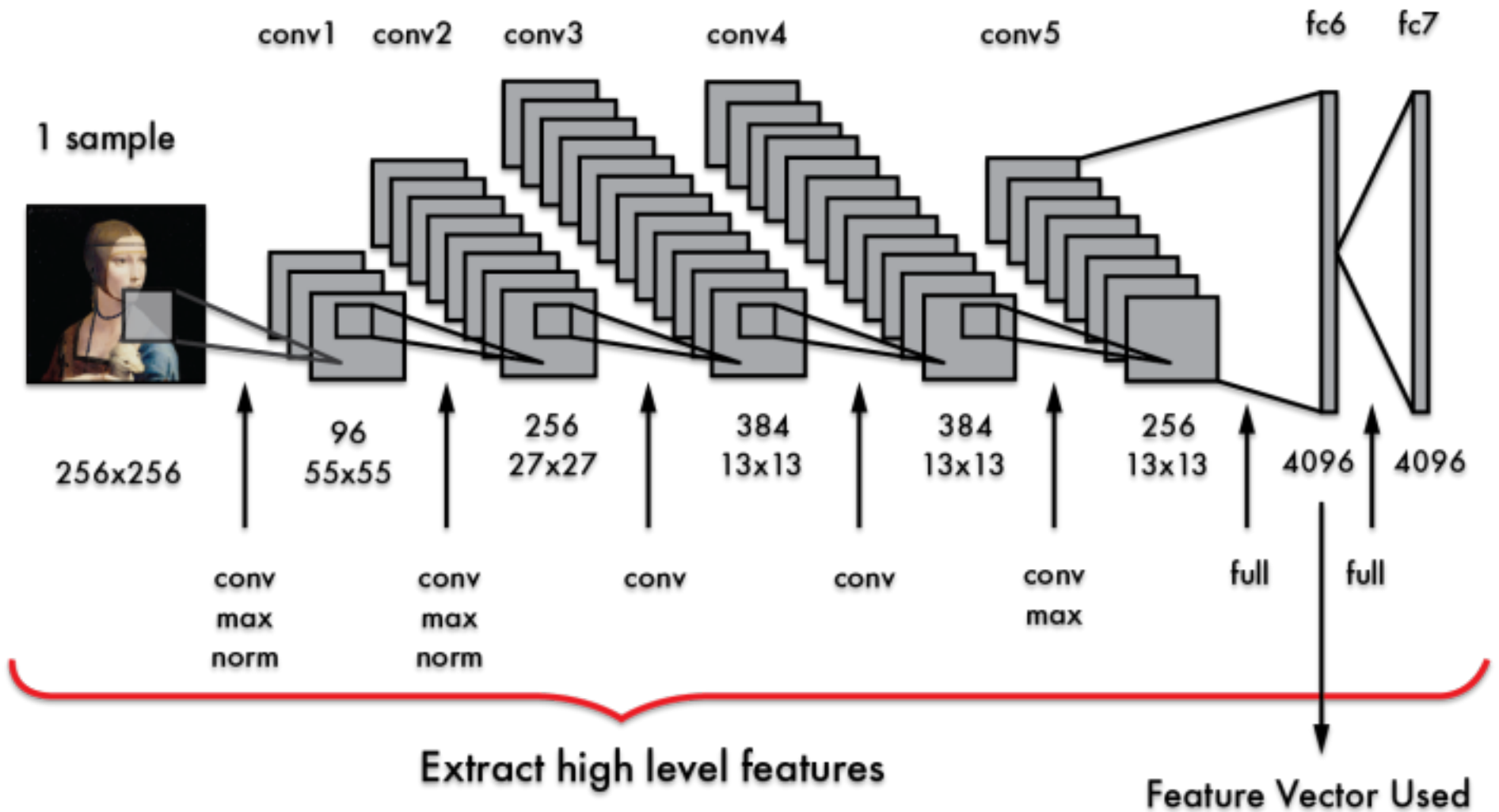
- (AVF) Attractiveness-based



Visual Features: DNN vs. AVF

- Deep Neural Networks (DNN): we used an AlexNet DNN (pre-trained with ImageNet ILSVRC 2012 dataset) (Krizhevsky et al, 2012) to map each artwork image to its corresponding latent vector of 4,096 dimensions obtained at the fc6 layer of the AlexNet network.

Visual Features: DNN vs. AVF



UGallery Data: Visual Features

- Average brightness,
- Saturation,
- Sharpness,
- Entropy,
- RGB-contrast,
- Colorfulness,
- Naturalness

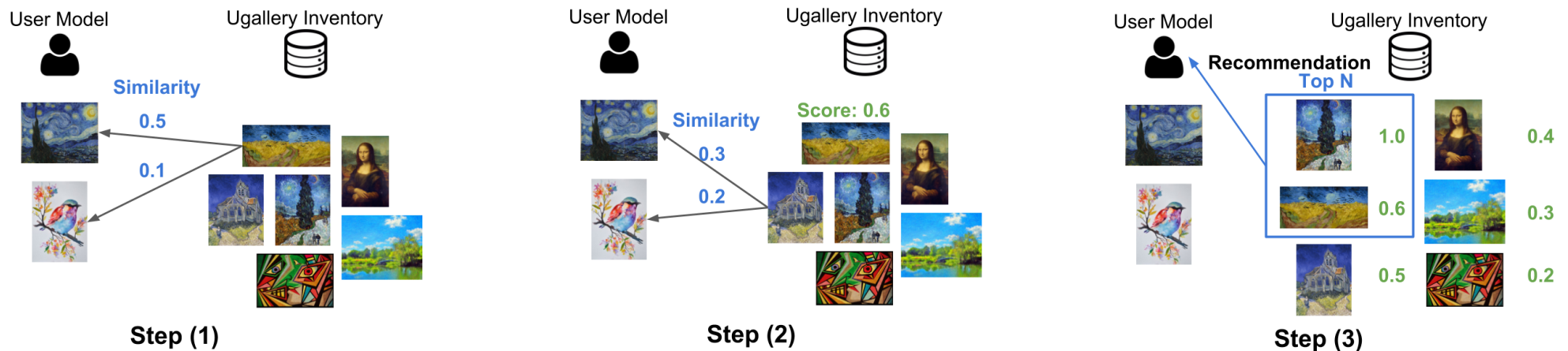
*Jose San Pedro and Stefan Siersdorfer.
2009.*

*Ranking and Classifying Attractiveness of
Photos in Folksonomies.*

*In Proceedings of the 18th International
Conference on World Wide Web (WWW
'09).*

Ugallery: Making Recommendations

- Scoring items based on cosine similarity between user model and item model:



$$\text{sim}(V_i, V_j) = \cos(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\| \|V_j\|}$$

Motivation & Research Questions

- Explaining recommendations is an active area of research, but there is little research on explaining image recommendation.

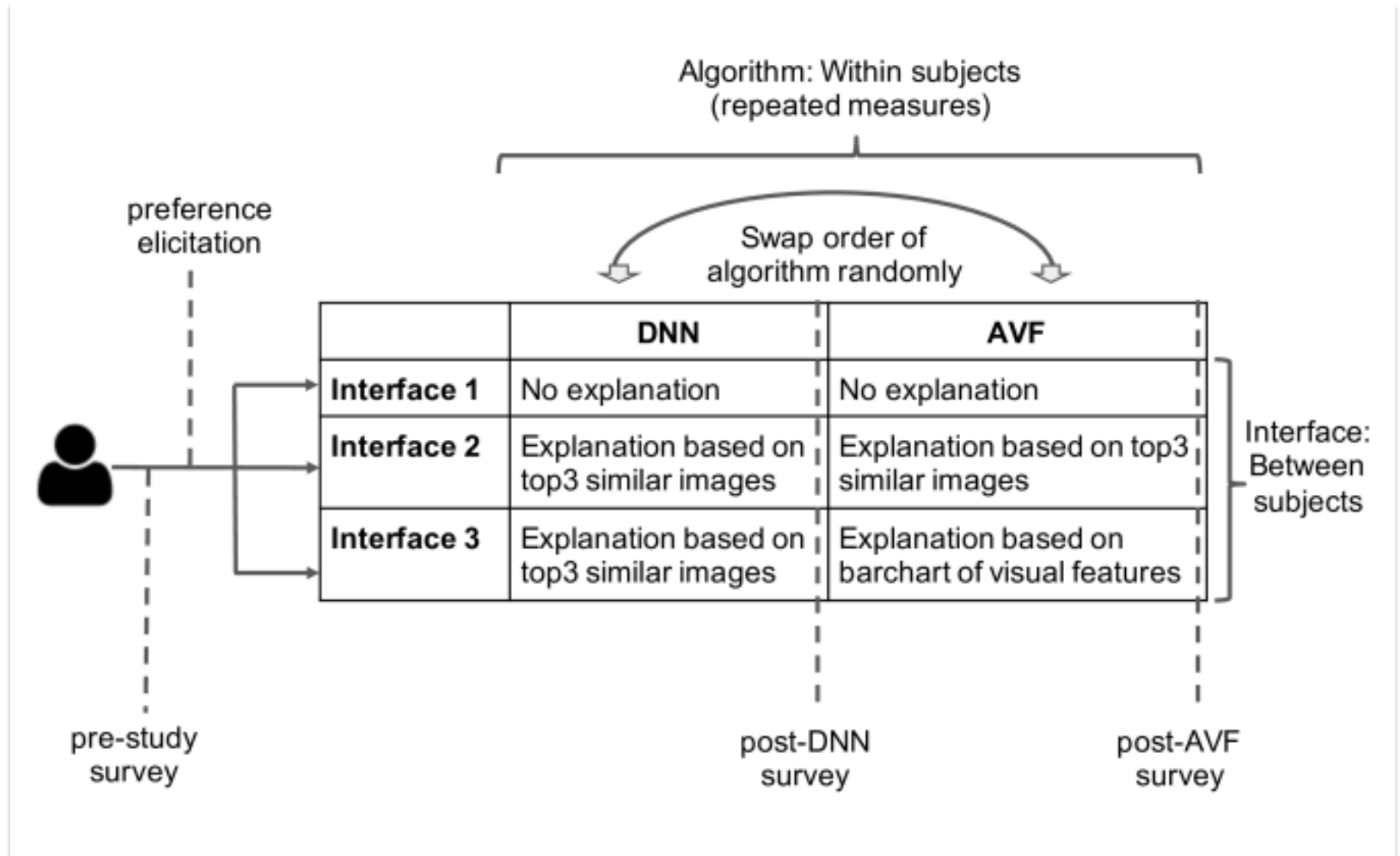
Motivation & Research Questions

- Research questions:
 - Which is the impact of an image recommendation explanation interface? Which explanation has more impact ?
 - Is there an interaction between the image recommendation algorithm and the explanation interface?

Methodology

- User study conducted in Amazon Mechanical Turk.
- Image data from UGallery, a web e-commerce for one-of-a-kind physical artworks.

Study Procedure

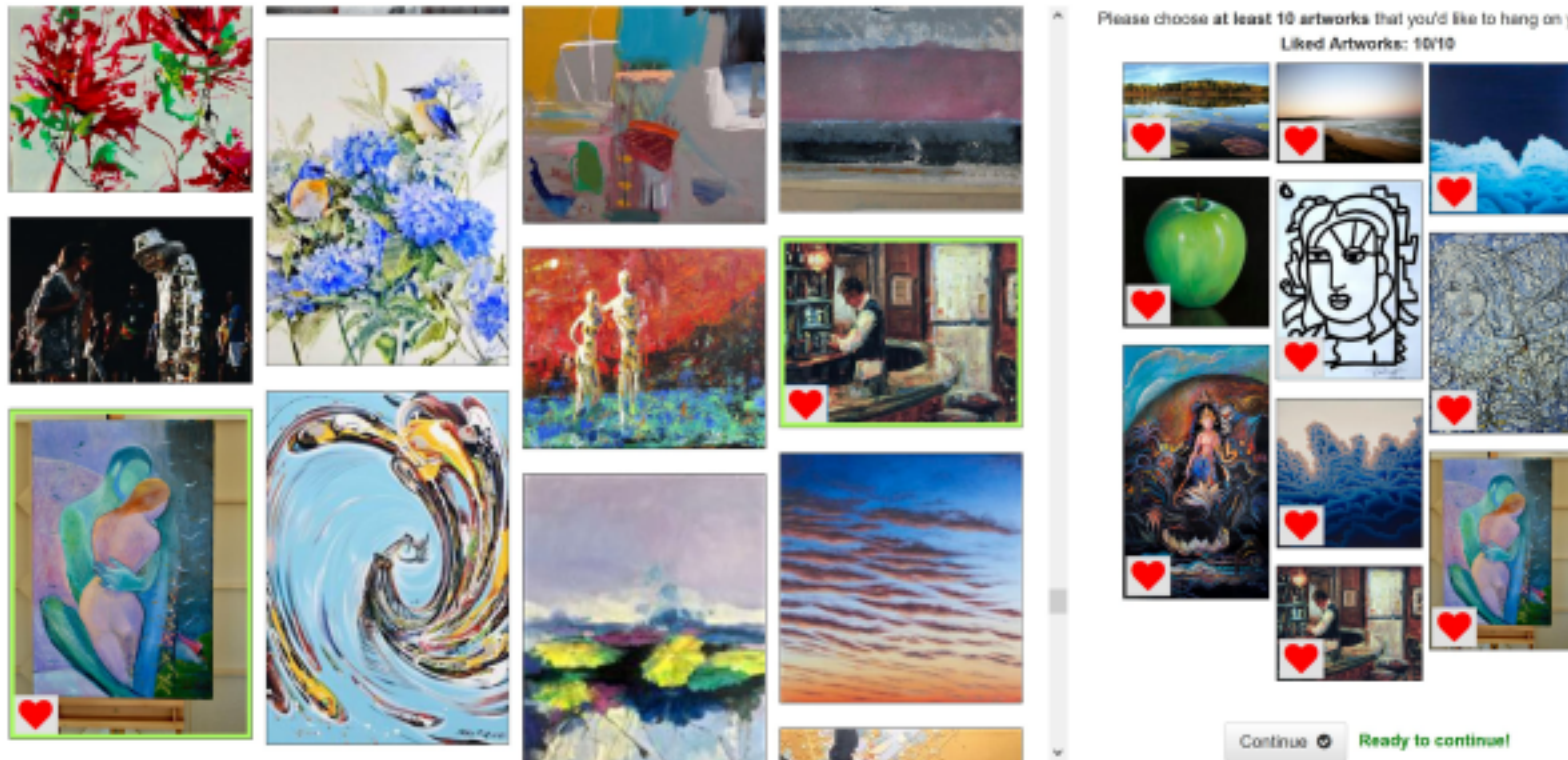


Preference Elicitation

- We collect user preferences from a Pinterest-like interface

User Study: (step 3 of 5) Exploration

user Logout

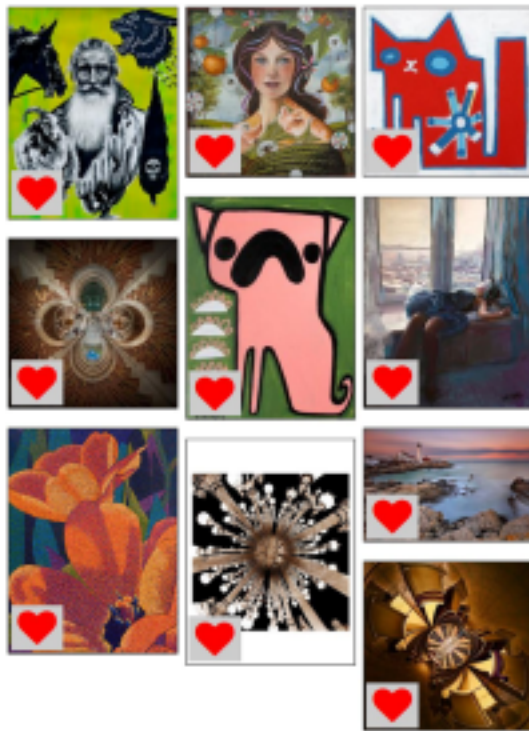


Interface 1 : no explanation, no transparency

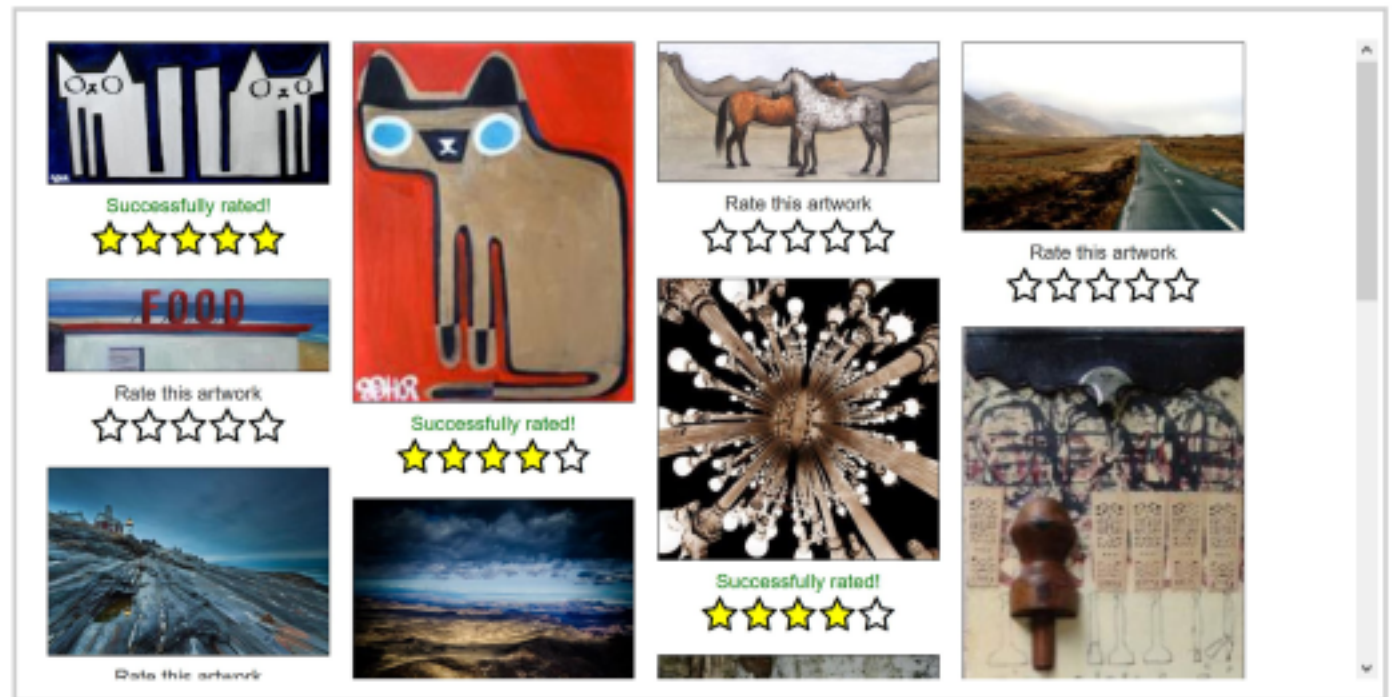
User Study: (step 4 of 5) Recommendation

_user Logout

Recommender 2 of 2



Artworks rated: 3/10



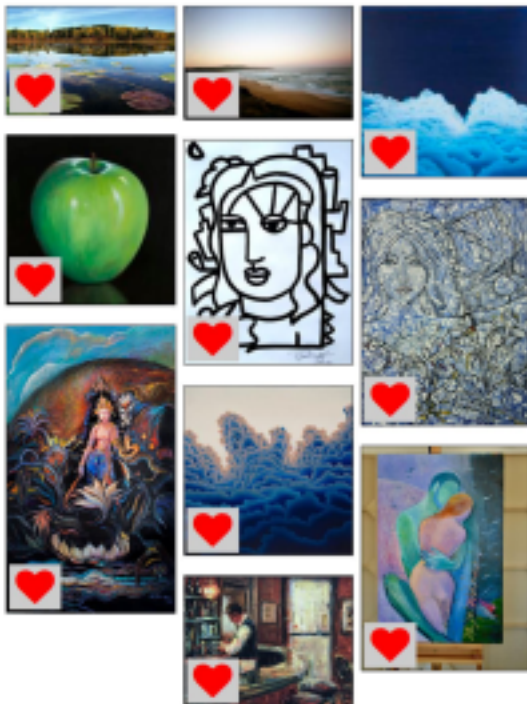
Continue to survey You still have to rate 7 artworks before continuing

Interface 2: explainable, no transparency

User Study: (step 4 of 5) Recommendation

user Logout

Recommender 2 of 2



Artworks rated: 2/10

Recommended Artwork	Explanation
<p>Successfully rated!</p> <p>★★★★★</p>	<p>Recommended because:</p> <p>it's 81.98% similar to this artwork that you like it's 70.10% similar to this artwork that you like it's 88.52% similar to this artwork that you like</p> <p>With an average of 73.53%</p>
<p>Rate this artwork</p> <p>☆☆☆☆☆</p>	<p>Recommended because:</p> <p>it's 75.99% similar to this artwork that you like it's 74.11% similar to this artwork that you like it's 70.11% similar to this artwork that you like</p>

Continue to survey

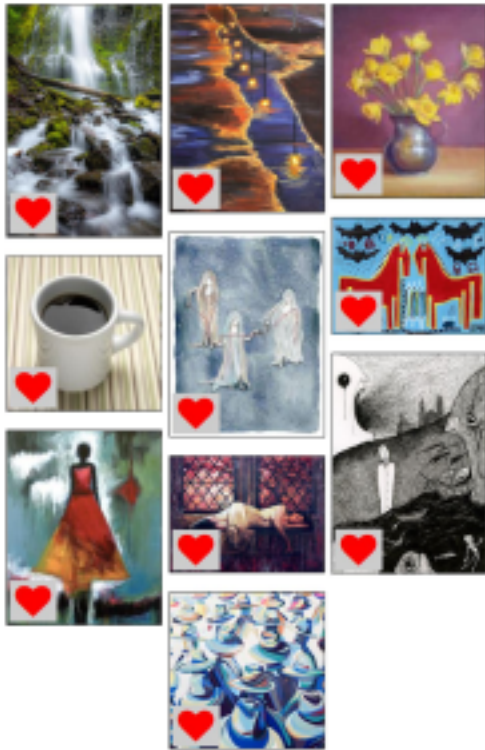
You still have to rate 8 artworks before continuing

Interface 3: explainable & transparent







User Study: (step 4 of 5) Recommendation


user Logout

Recommender 1 of 2



Artworks rated: 0/10

Recommended Artwork	Explanation
 <p>Rate this artwork ☆☆☆☆☆</p>	<p>Recommended because:</p>  <p>it's 96.32% similar to this artwork that you like</p> 
 <p>Rate this artwork ☆☆☆☆☆</p>	<p>Recommended because:</p>  <p>it's 96.27% similar to this artwork that you like</p> 

Continue to survey  You still have to rate 10 artworks before continuing

Evaluation & Results

Study on Amazon Mechanical Turk:

- 121 valid users completed correctly the study.
- Task took them around 10 minutes to complete.
- ~56% female, 44% male.
- 80% attended to 1 or more art classes at high school level or above.
- 80% visited museums or art galleries at least once a year.

Evaluation & Results

Post-study survey aspects (agreement 1-100)

- **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.

Evaluation & Results

- Result 1: DNN better than AVF except on diversity

Condition	Evaluation Dimensions													
	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
Interface 2 (DNN & AVF: Top-3 similar images)	83.5* ^{†1}	74.0 ^{†1}	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
Interface 3 (DNN: Top-3 similar, AVF: feature bar chart)	84.2* ^{†1}	70.4 ^{†1}	82.3* ^{†1}	56.2	65.3 ^{†1}	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

Evaluation & Results

- Result 1: DNN better than AVF except on diversity

Condition	Evaluation Dimensions													
	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
Interface 2 (DNN & AVF: Top-3 similar images)	83.5* ^{†1}	74.0 ^{†1}	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
Interface 3 (DNN: Top-3 similar, AVF: feature bar chart)	84.2* ^{†1}	70.4 ^{†1}	82.3* ^{†1}	56.2	65.3 ^{†1}	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

Evaluation & Results

Result 2: Boost of several dimensions by using explanations, but..

Condition	Evaluation Dimensions													
	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
Interface 2 (DNN & AVF: Top-3 similar images)	83.5* ^{†1}	74.0* ^{†1}	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
Interface 3 (DNN: Top-3 similar, AVF: feature bar chart)	84.2* ^{†1}	70.4* ^{†1}	82.3* ^{†1}	56.2	65.3* ^{†1}	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

- Result expected: people perceive it as more explainable using the explainable interfaces

Evaluation & Results

- Result 1: DNN better than AVF except on diversity

Condition	Evaluation Dimensions													
	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
Interface 2 (DNN & AVF: Top-3 similar images)	83.5* ^{†1}	74.0 ^{†1}	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
Interface 3 (DNN: Top-3 similar, AVF: feature bar chart)	84.2* ^{†1}	70.4 ^{†1}	82.3* ^{†1}	56.2	65.3 ^{†1}	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

- Algorithm is the same (DNN), but by adding explanations people perceive it as more relevant

Evaluation & Results

Result 2: Perception of Diversity did not change for AVF, but it changed for DNN

Condition	Evaluation Dimensions													
	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
Interface 2 (DNN & AVF: Top-3 similar images)	83.5* ^{†1}	74.0 ^{†1}	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
Interface 3 (DNN: Top-3 similar, AVF: feature bar chart)	84.2* ^{†1}	70.4 ^{†1}	82.3* ^{†1}	56.2	65.3 ^{†1}	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

Evaluation & Results

Result 4: “Trust difference” becomes significant only after explanations

Condition	Evaluation Dimensions													
	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
Interface 2 (DNN & AVF: Top-3 similar images)	83.5* ^{†1}	74.0 ^{†1}	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
Interface 3 (DNN: Top-3 similar, AVF: feature bar chart)	84.2* ^{†1}	70.4 ^{†1}	82.3* ^{†1}	56.2	65.3 ^{†1}	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

Conclusion and Future Work

Conclusions:

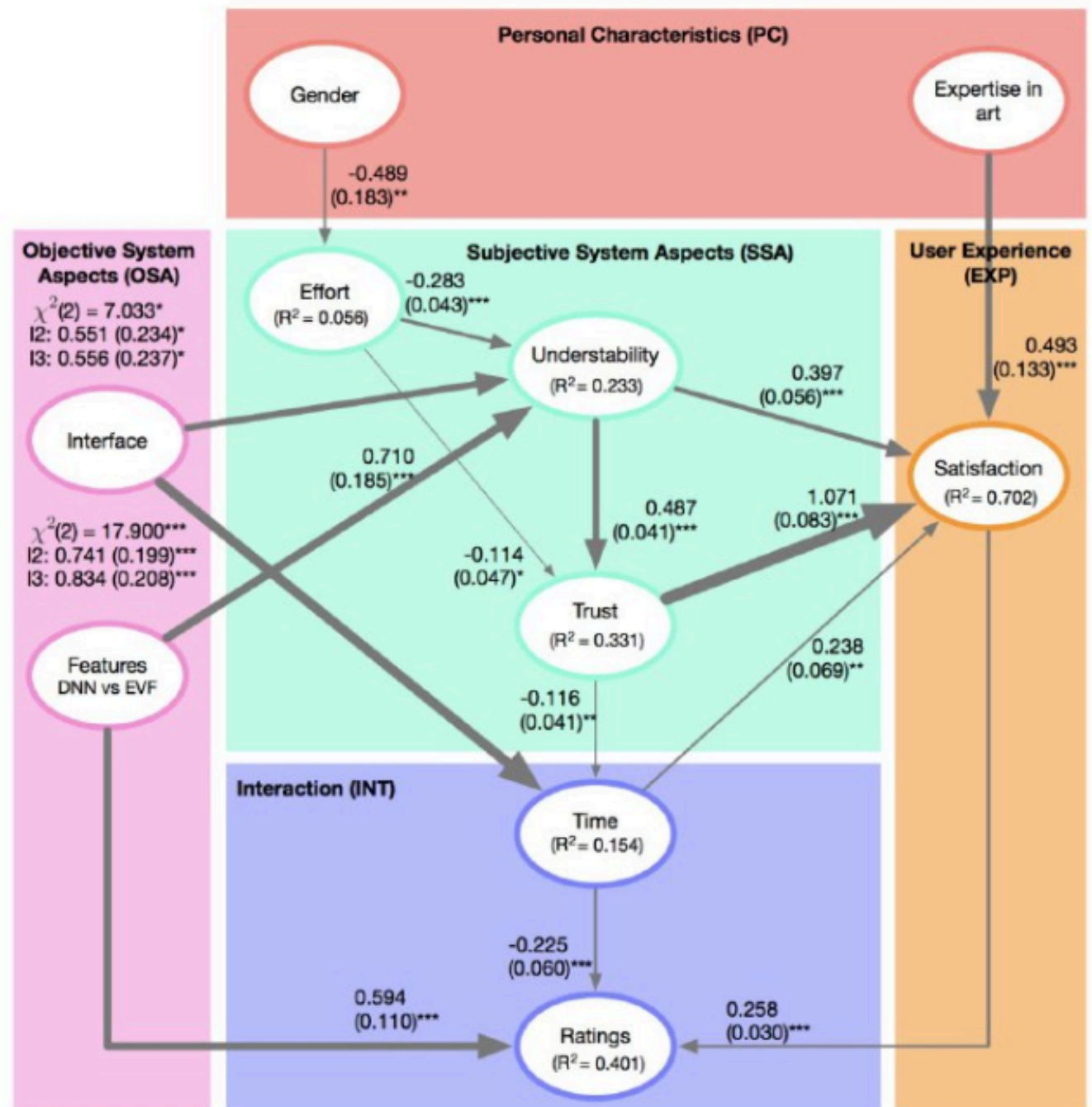
- DNN features performed better than AVF features, probably because AlexNet is able to capture more complex patterns. Results confirm previous work by Messina et al. (2018)
- Providing explanations has a big effect on several dimensions evaluated.
- Results indicate that the users' perception of trust and explainability is not only about the interface's explainability and transparency, is also affected by other factors such as the perception of relevance. A SEM analysis is necessary to understand these relations.

Conclusion and Future Work

Future Work:

- Build a model-based recommender rather than the current one, based on K-NN and heuristics.
- Incorporate other types of information (metadata) in the explanation interface.

Full Model: SEM using Knijnenburg framework



Acknowledgment

- The authors from PUC Chile were funded by PUC Chile, Conicyt, Fondecyt grant 11150783, as well as by the Millennium Institute for Foundational Research on Data (IMFD).

References

- Vicente Dominguez, Pablo Messina, Denis Parra, Domingo Mery, Christoph Trattner, and Alvaro Soto. 2017. Comparing Neural and Attractiveness-based Visual Features for Artwork Recommendation. In Proceedings of the Workshop on Deep Learning for Recommender Systems, co-located at RecSys 2017.
- 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).
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems.
- Pablo Messina, Vicente Dominguez, , Denis Parra, Christoph Trattner, and Alvaro Soto. 2018. Content-Based Artwork Recommendation: Integrating Painting Metadata with Neural and Manually-Engineered Visual Features. User Modeling and User-Adapted Interaction (2018).
- Jose San Pedro and Stefan Siersdorfer. 2009. Ranking and Classifying Attractiveness of Photos in Folksonomies. In Proceedings of the 18th International Conference on World Wide Web (WWW '09).

vidominguez@uc.cl

THANKS!

