

Implicit Feedback for Recommender Systems

Denis Parra

Sistemas Recomendadores

IIC 3633

Implicit Feedback

- Las siguientes slides son un resumen del Paper Hu, Y., Koren, Y., & Volinsky, C. (2008, December). Collaborative filtering for implicit feedback datasets. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 263-272). IEEE.

Ratings = recurso excaso

- Si bien SVD++ considera implicit feedback, este modelo optimiza específicamente feedback implícito
- Considera, antes que todo, valores binarios de consumo/no consumo del ítem

$$p_{ui} = \begin{cases} 1 & r_{ui} > 0 \\ 0 & r_{ui} = 0 \end{cases}$$

Implicit Feedback – Hu et al.

- Se considera también la confianza de observar p_{ui} con la variable c_{ui} ($\alpha = 40$, uso de CV)

$$c_{ui} = 1 + \alpha r_{ui}$$

r_{ui} es, en este caso, el implicit feedback (eg plays)

- La función que esperamos minimizar es, luego

$$\min_{x_*, y_*} \sum_{u,i} c_{ui} (p_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$

Implicit Feedback – Hu et al.

- Learning: ALS en lugar de SGD
- c_{ui} puede tomar distintas formas. Una alternativa es

$$c_{ui} = 1 + \alpha \log(1 + r_{ui}/\epsilon)$$

- De esta forma, el implicit feedback r_{ui} se descompone en p_{ui} (preferencias) y c_{ui} (nivel de confianza), y
- Maneja todas las combinaciones usuario-item ($n * m$) en tiempo lineal al explotar la estructura algebraica de las variables

Experimento

- Servicio de TV digital, datos recolectados de 300,000 set top boxes.
- En un periodo de 4 semanas, 17.000 programas de TV únicos
- r_{ui} : cuántas veces usuario u vio programa i en un período de 4 semanas
- Luego de una agregación y limpieza de datos, $|r_{ui}|$: 32 millones

Evaluación y Resultados

- Rank_{ui} : percentil-ranking de un programa i en la lista de recomendación de u .
- Si $\text{rank}_{ui} = 0\%$, el programa i ha sido predicho como el más relevante para el usuario u , y si $\text{rank}_{ui} = 100\%$, el programa i es el menos deseado.
- Expected percentile ranking $\overline{\text{rank}}_{ui}$: the smaller the better

$$\overline{\text{rank}} = \frac{\sum_{u,i} r_{ui}^t \text{rank}_{ui}}{\sum_{u,i} r_{ui}^t}$$

Resultados

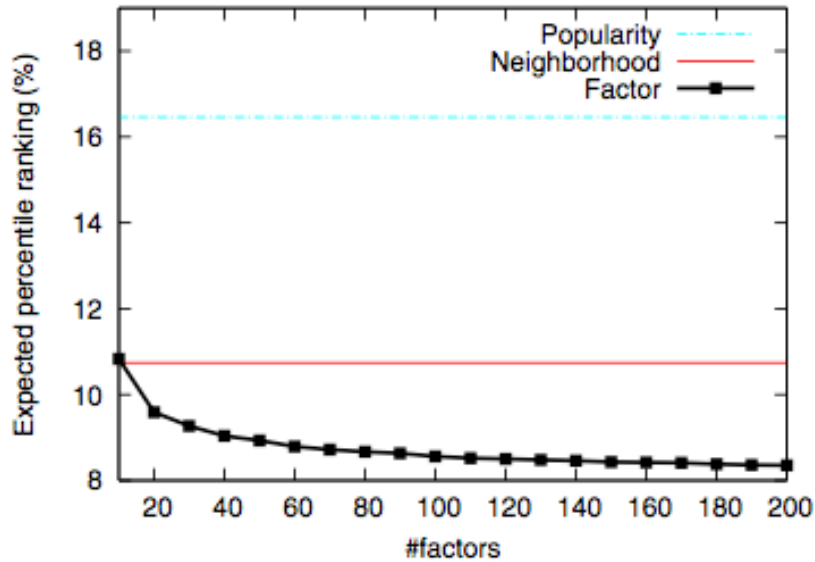


Figure 1. Comparing factor model with popularity ranking and neighborhood model.

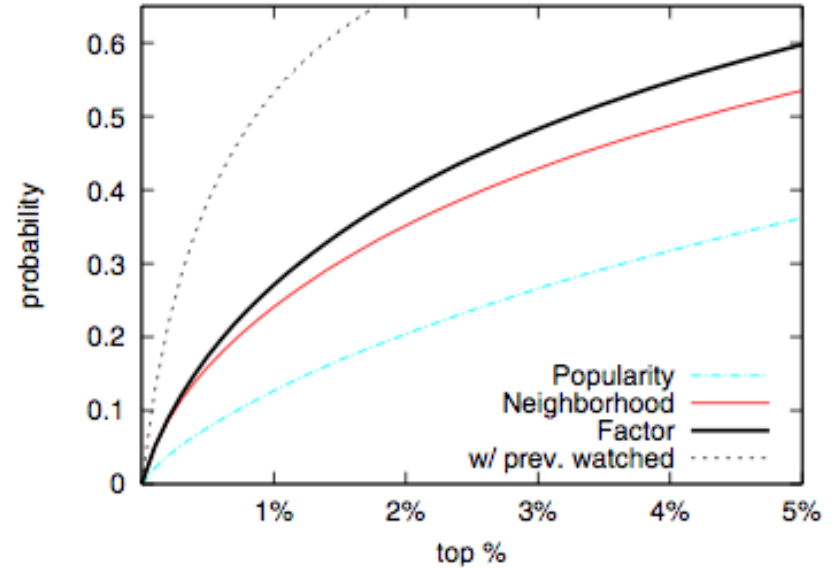


Figure 2. Cumulative distribution function of the probability that a show watched in the test set falls within top x% of recommended shows.

Resultados II

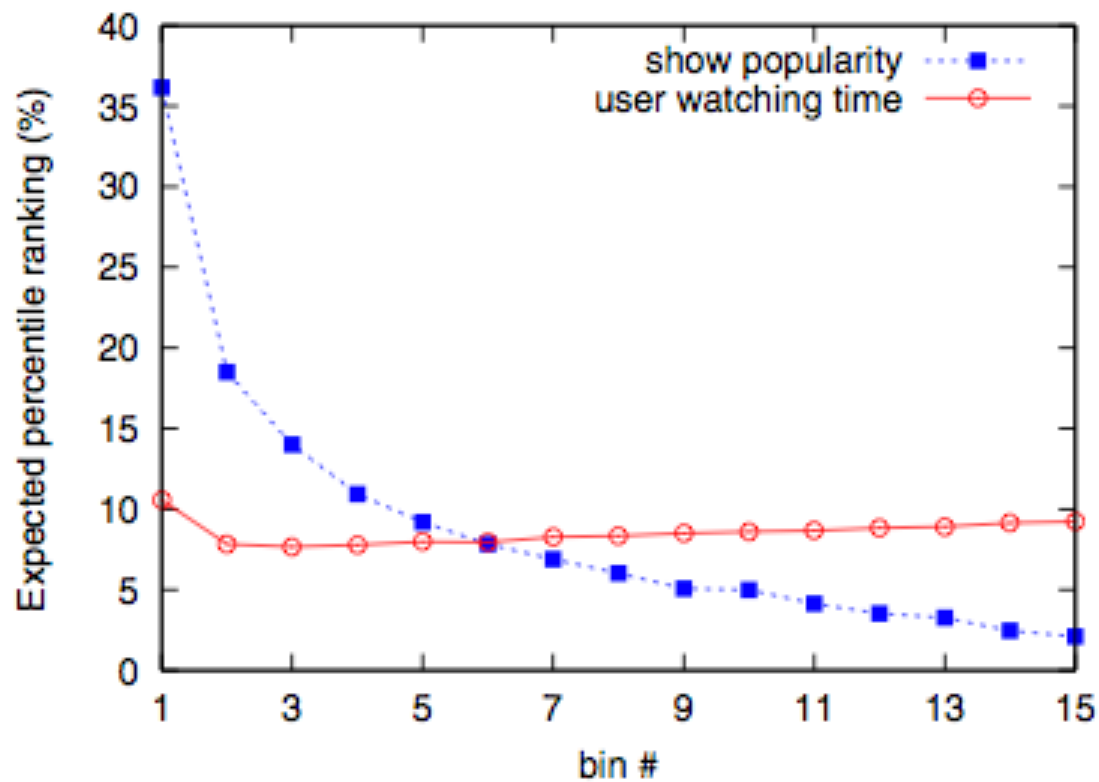


Figure 3. Analyzing the performance of the factor model by segregating users/shows based on different criteria.

Implicit Feedback

- Parra, D., & Amatriain, X. (2011). Walk the Talk. In User Modeling, Adaption and Personalization (pp. 255-268). Springer Berlin Heidelberg.

Implicit-Feedback

- Slides are based on two articles:
 - Parra-Santander, D., & Amatriain, X. (2011). Walk the Talk: Analyzing the relation between implicit and explicit feedback for preference elicitation. Proceedings of UMAP 2011, Girona, Spain
 - Parra, D., Karatzoglou, A., Amatriain, X., & Yavuz, I. (2011). Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping. Proceedings of the CARS Workshop, Chicago, IL, USA, 2011.

Introduction (1/2)

- Most of recommender system approaches rely on explicit information of the users, but...
- Explicit feedback: scarce (people are not especially eager to rate or to provide personal info)
- Implicit feedback: Is less scarce, but ([Hu et al., 2008](#))

There's no negative feedback	... and if you watch a TV program just once or twice?
Noisy	... but explicit feedback is also noisy (Amatriain et al., 2009)
Preference & Confidence	... we aim to map the I.F. to preference (our main goal)
Lack of evaluation metrics	... if we can map I.F. and E.F., we can have a comparable evaluation

Introduction (2/2)

- Is it possible to map implicit behavior to explicit preference (ratings)?
- Which variables better account for the amount of times a user listens to online albums? [Baltrunas & Amatriain CARS '09 workshop – RecSys 2009.]
- **OUR APPROACH: Study with Last.fm users**
 - Part I: Ask users to rate 100 albums (how to sample)
 - Part II: Build a model to map collected implicit feedback and context to explicit feedback

Walk the Talk (2011)

Albums they listened to during last:
7days, 3months, 6months, year,
overall

For each album in the list we
obtained: # user plays (in each
period), # of global listeners and # of
global plays

Rank	Album	Plays
1	Radiohead – The King of Limbs	72
2	Nick Cave & The Bad Seeds – The Boatman's Call	59
3	Nick Cave & The Bad Seeds – The Best Of (Disc 1)	36
4	Radiohead – Kid A	34
5	Nick Cave and the Bad Seeds – Murder Ballads	30
6	Nick Cave & The Bad Seeds – The Lyre Of Orpheus	24
7	Radiohead – In Rainbows	18
8	Nick Cave and the Bad Seeds – The Boatman's Call	14
9	Life's Decay – Eklaasera	12
10	Nick Cave and the Bad Seeds – The Lyre of Orpheus	9

Walk the Talk - 2

- Requirements: 18 y.o., scrobblings > 5000

Survey about music taste - Telefonica I+D

Part I: 11 questions about demographics, music experience and consumption.

A) User Consent

Before starting the survey, please tell us if you accept the [terms and conditions of this study](#).

I have read the terms and conditions of this study and I accept voluntarily to participate on it. I also acknowledge that I am 18 years old or older.

B) Demographics

1. Gender
[Dropdown menu]

2. Age
[Input field: 27] Your age must be a number between 18 and 99.

3. Current Country
[Dropdown menu]

C) Media Consumption behavior

1. How many hours per week do you use the internet?
[Dropdown menu]

2. How many hours per week do you listen to music?
[Dropdown menu]

3. How many concerts do you usually attend per year?
[Dropdown menu]

4. How frequently do you read specialized blogs or

Gold

Artist/Band | The Cranberries

Tracks (up to 12) |

1. Dreams
2. Salvation
3. Sunday
4. Free To Decide
5. Pretty
6. When You're Gone
7. How
8. Hollywood
9. Cordell
10. Not Sorry
11. Animal Instinct
12. Linger

Need more info? [Click here for additional information about this album](#)

How would you rate this album?

👁️ ⭐⭐⭐⭐

M2:
implicit
feedback &
recentness

4 Regression Analysis

M4:
Interaction of
implicit
feedback &
recentness

- Model 1: $r_{iu} = \beta_0 + \beta_1 \cdot ifiu$
- Model 2: $r_{iu} = \beta_0 + \beta_1 \cdot ifiu + \beta_2 \cdot reiu$
- Model 3: $r_{iu} = \beta_0 + \beta_1 \cdot ifiu + \beta_2 \cdot reiu + \beta_3 \cdot gpi$
- Model 4: $r_{iu} = \beta_0 + \beta_1 \cdot ifiu + \beta_2 \cdot reiu + \beta_3 \cdot ifiu \cdot reiu$

M1: implicit feedback

M3: implicit
feedback,
recentness,
global
popularity

Model	R^2	F -value	p -value	β_0	β_1	β_2	β_3
1	0.125	$F(1, 10120) = 1146$	$< 2.2 \cdot 10^{-16}$	2.726	0.499	-	-
2	0.1358	$F(2, 10019) = 794.8$	$< 2.2 \cdot 10^{-16}$	2.491	0.484	0.133	-
3	0.1362	$F(3, 10018) = 531.8$	$< 2.2 \cdot 10^{-16}$	2.435	0.486	0.134	0.0285
4	0.1368	$F(3, 10018) = 534.7$	$< 2.2 \cdot 10^{-16}$	2.677	0.379	0.038	0.053

Table 1. Regression Results. R^2 , F -value, and p -value for the 5 models.

- Including Recentness increases R^2 in more than 10% [1 -> 2]
- Including GP increases R^2 , not much compared to RE + IF [1 -> 3]
- Not Including GP, but including interaction between IF and RE improves the variance of the DV explained by the regression model. [2 -> 4]

4.1 Regression Analysis

Model	RMSE1	RMSE2
User average	1.5308	1.1051
M1: Implicit feedback	1.4206	1.0402
M2: Implicit feedback + recentness	1.4136	1.034
M3: Implicit feedback + recentness + global popularity	1.4130	1.0338
M4: Interaction of Implicit feedback * recentness	1.4127	1.0332

- We tested conclusions of regression analysis by predicting the score, checking RMSE in 10-fold cross validation.
- **Results of regression analysis are supported.**

Conclusions of Part I

- Using a linear model, Implicit feedback and recency can help to predict explicit feedback (in the form of ratings)
- Global popularity doesn't show a significant improvement in the prediction task
- Our model can help to relate implicit and explicit feedback, helping to evaluate and compare explicit and implicit recommender systems.

Part II: Extension of Walk the Talk

- Implicit Feedback Recommendation via Implicit-to-Explicit OLR Mapping (Recsys 2011, CARS Workshop)
 - Consider ratings as ordinal variables
 - Use mixed-models to account for non-independence of observations
 - Compare with state-of-the-art implicit feedback algorithm

Recalling the 1st study (5/5)

- Prediction of rating by **multiple Linear Regression evaluated with RMSE**.
- Results showed that **Implicit feedback** (play count of the album by a specific user) and **recentness** (how recently an album was listened to) were important factors, **global popularity** had a weaker effect.
- Results also showed that **listening style** (if user preferred to listen to single tracks, CDs, or either) was also an important factor, and not the other ones.

... but

- Linear Regression didn't account for the **nested nature of ratings**



- And **ratings** were treated as **continuous**, when they are **actually ordinal**.

So, Ordinal Logistic Regression!

- Actually *Mixed-Effects Ordinal Multinomial Logistic Regression*
- **Mixed-effects**: Nested nature of ratings
- We obtain a distribution over ratings (**ordinal multinomial**) per each pair USER, ITEM -> we **predict** the rating using the **expected value**.
- ... And we can **compare the inferred ratings with a method that directly uses implicit information (playcounts) to recommend** (by Hu, Koren et al. 2007)

Ordinal Regression for Mapping

- Model

$$\text{logit}(P(r_{ui} \leq k)) = \alpha_k + X\beta + g_u$$

where $k = \{1, 2, 3, 4\}$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

- Predicted value

$$E[r_{ui}] = \sum_{k=1}^5 k \cdot P(r_{ui} = k)$$

$$P(r_{ui} = k) = \begin{cases} P(r_{ui} \leq k) & , k = 1 \\ P(r_{ui} \leq k) - P(r_{ui} \leq k - 1) & , 1 < k < 5 \\ 1 - P(r_{ui} \leq k - 1) & , k = 5 \end{cases}$$

Datasets

- D1: users, albums, if, re, gp, ratings, demographics/consumption
- D2: users, albums, if, re, gp, NO RATINGS.

	Dataset1 (Implicit Explicit)	Dataset2 (Implicit)
users	114	2549
albums	6037	6037
entries	10122	111815
density	1.47%	0.73%
avg albums/user	88.79	43.87
avg user/album	1.71	18.52

Table 3: Description of the datasets

Results

	MAP (D1)	nDCG(D1)	MAP(D2)	nDCG(D2)
HK	0.02315	0.14831	0.1014	0.2718
HKlog	0.02742	0.15447	0.1234	0.2954
logit3	0.02636	0.15319	0.1223	0.2944
logit4	0.02601	0.15268	N/A	N/A
popularity	0.48331	0.54378	0.0178	0.1367

Table 4: Results of MAP and nDCG after 5-fold Cross validation on dataset 1 (D1) and dataset 2 (D2)

Conclusions & Current Work

Problem/ Challenge

1. Ground truth: How many Playcounts to relevancy?
> Sensibility Analysis needed

2. Quantization of playcounts (implicit feedback), recentness, and overall number of listeners of an album (global popularity) **[1-3] scale v/s raw playcounts > modify and compare**

3. Additional/Alternative metrics for evaluation [MAP and nDCG used in the paper]

Dwell Time

Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. **Beyond clicks: dwell time for personalization**

-- method to consume fine-grained dwell-time at web scale

* Focus Blur and Last Event methods: server side methods

* Focus blur closer to client side, so is the one used

-- dwell times varies by device (correlation between)

-- Raw dwell time distributions change considerably on content type, but at least log-raw distributions are bell shaped

* challenge: dwell time normalization, to extract an engagement signal which is comparable across devices -> they normalize

-- Dwell time is used in a learning to rank approach (using dwell time as target) to rank items

Evaluation on Yahoo! logs

-- Option 2 is using directly dwell time in a CF-based recommendation

Dwell Time for personalization

- La idea principal es usar dwell-time en lugar de explicit feedback (ratings) o implicit feedback (clicks) para hacer recomendaciones
- Temas:
 - Calculo del Dwell-time desde el server-side
 - Normalización entre distintos dispositivos
 - Recomendación basada en Learning to Rank
 - Recomendación basada en MF/CF

Dwell Time for personalization

Table 1: Client-side Logging Example

User Behaviors	Client-side Events
A user opens a news article page.	{DOM-ready, t_1 }
He reads the article for several seconds.	{Focus, t_2 }
He switches to another browser tab or a window to read other articles.	{Blur, t_3 }
He goes back to the article page and comments on it.	{Focus, t_4 }
He closes the article page, or clicks the back button to go to another page.	{BeforeUnload, t_5 }

$\{i, \text{Click}, t_1\} \rightarrow \{j, \text{Click}, t_2\} \rightarrow \{k, \text{Click}, t_3\} \rightarrow \{i, \text{Comment}, t_4\} \rightarrow \{n, \text{Click}, t_5\}$

Table 2: Comparison of dwell time measurement. The first two columns are for LE, the middle two columns are for FB and the last two columns are for client-side logs. Each row contains data from a day.

#	DT. (LE)	#	DT. (FB)	#	DT. (C)
3,322	86.5	3,197	134.4	3,410	130.3
5,711	85.4	5,392	132.6	5,829	124.0

Dwell Time for personalization

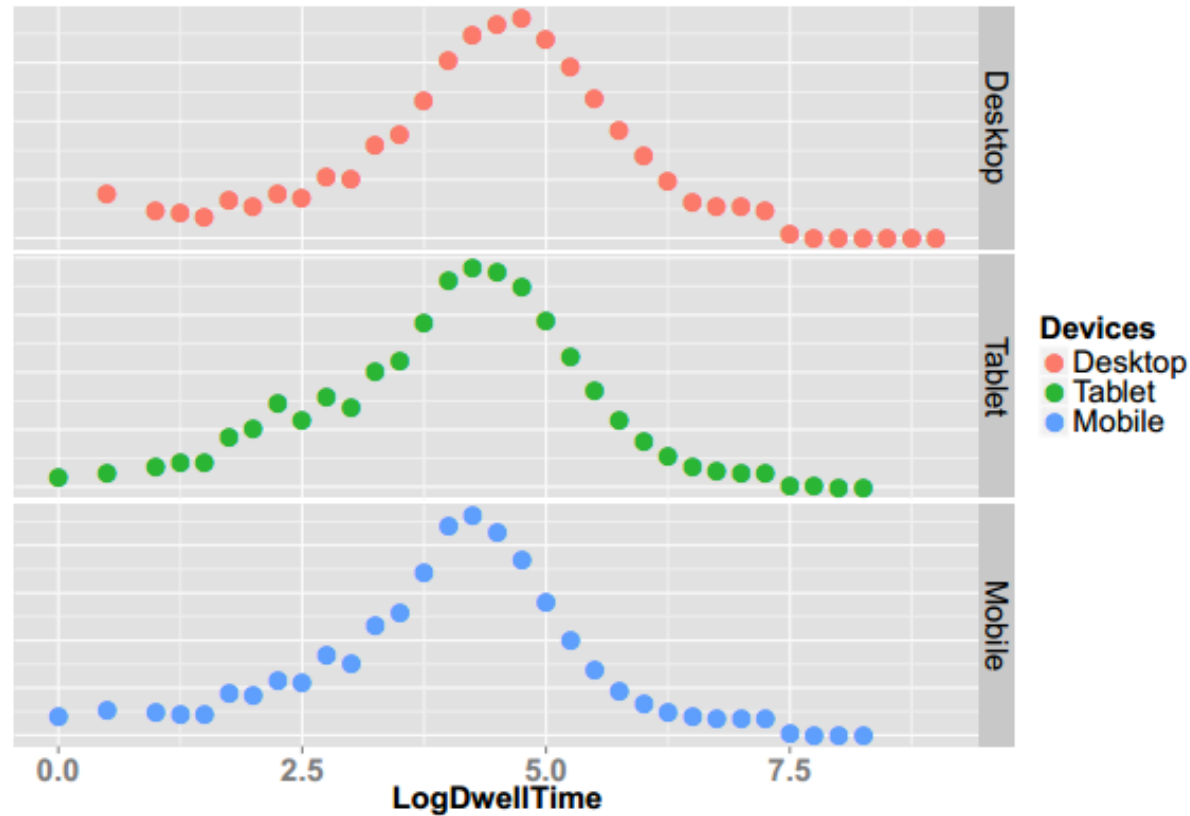
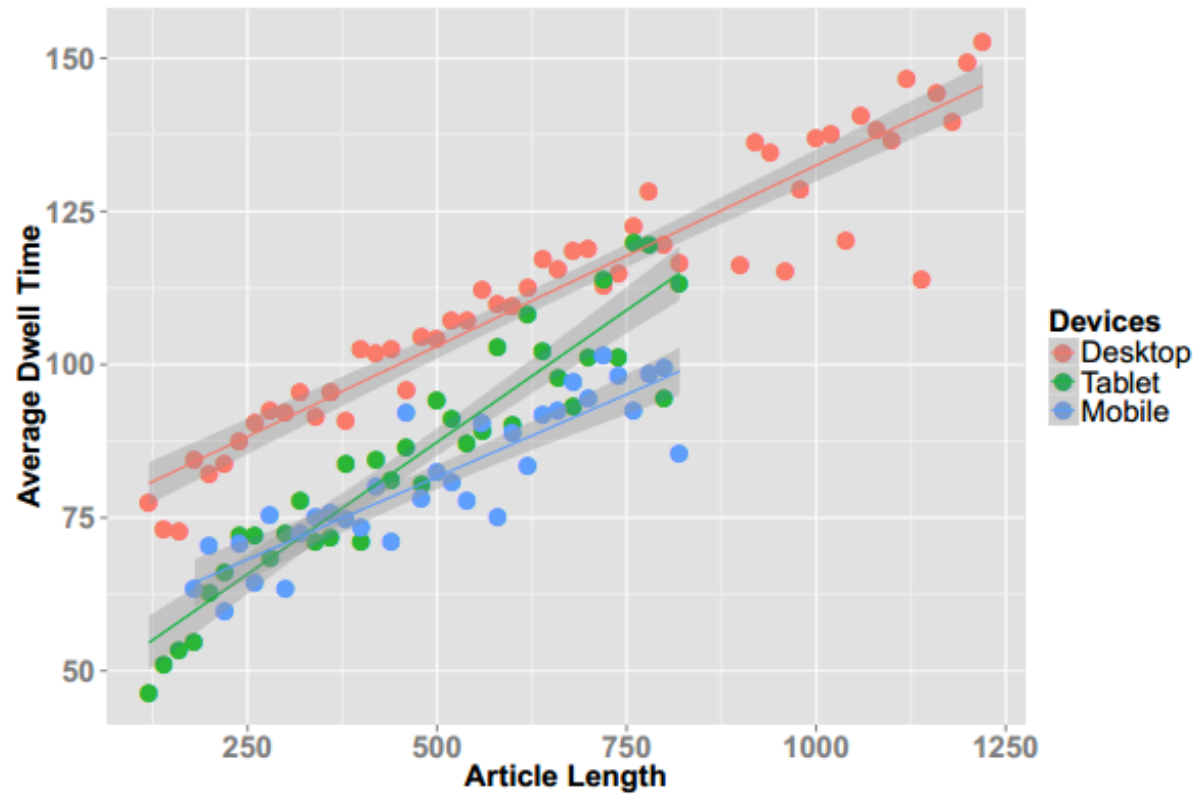


Figure 2: The (un)normalized distribution of log of dwell time for articles across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

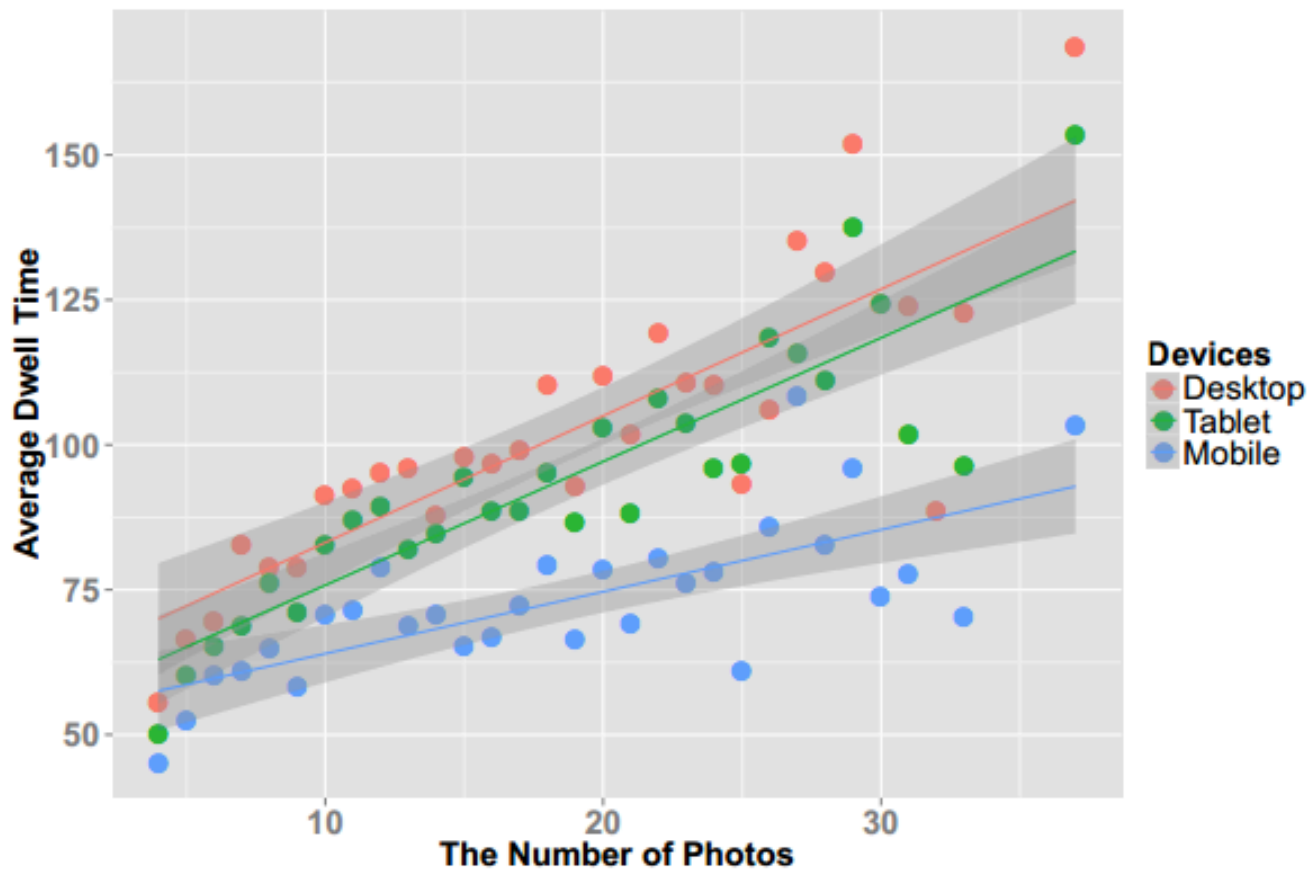
Dwell Time for personalization

- Relation with article length



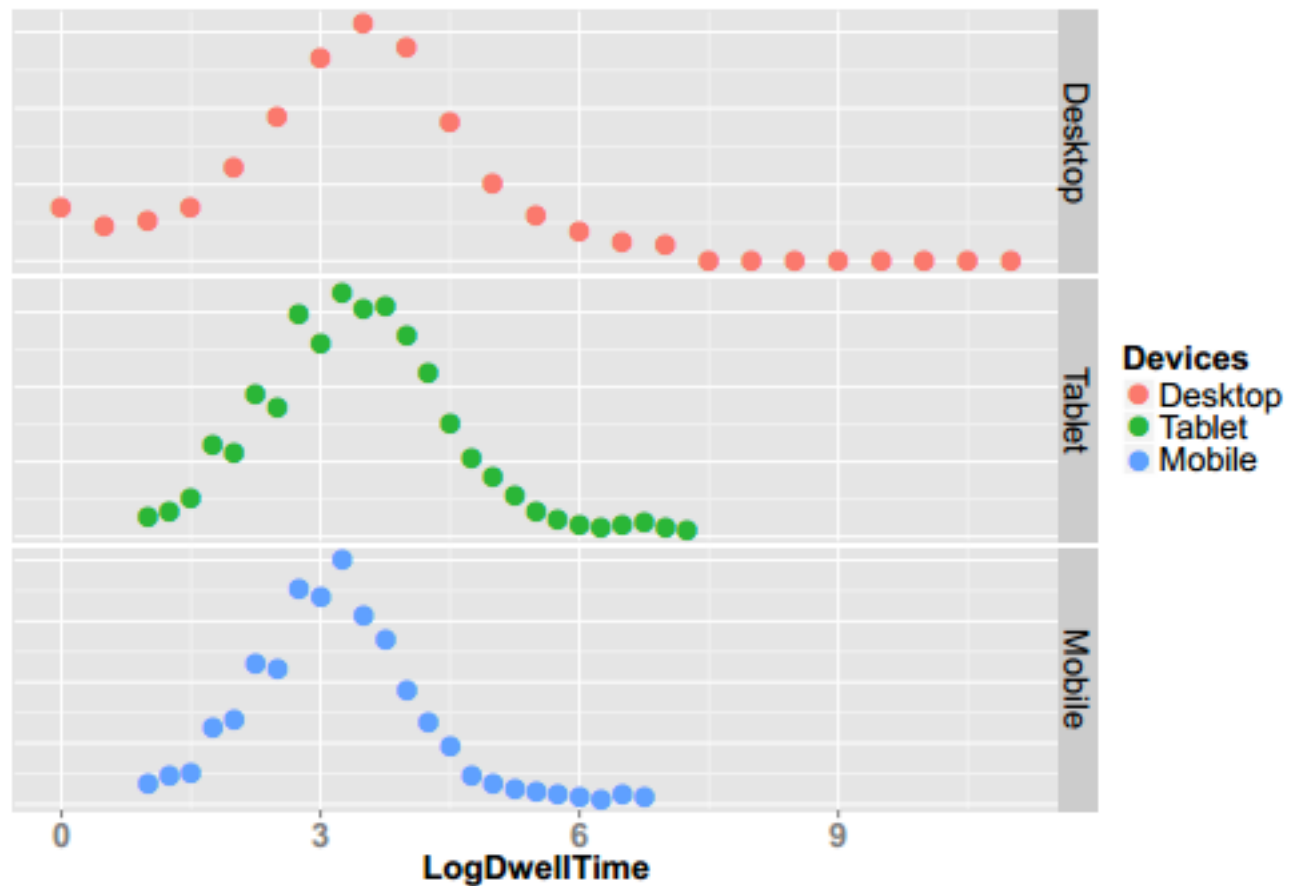
Dwell Time for personalization

- Relation with number of photos



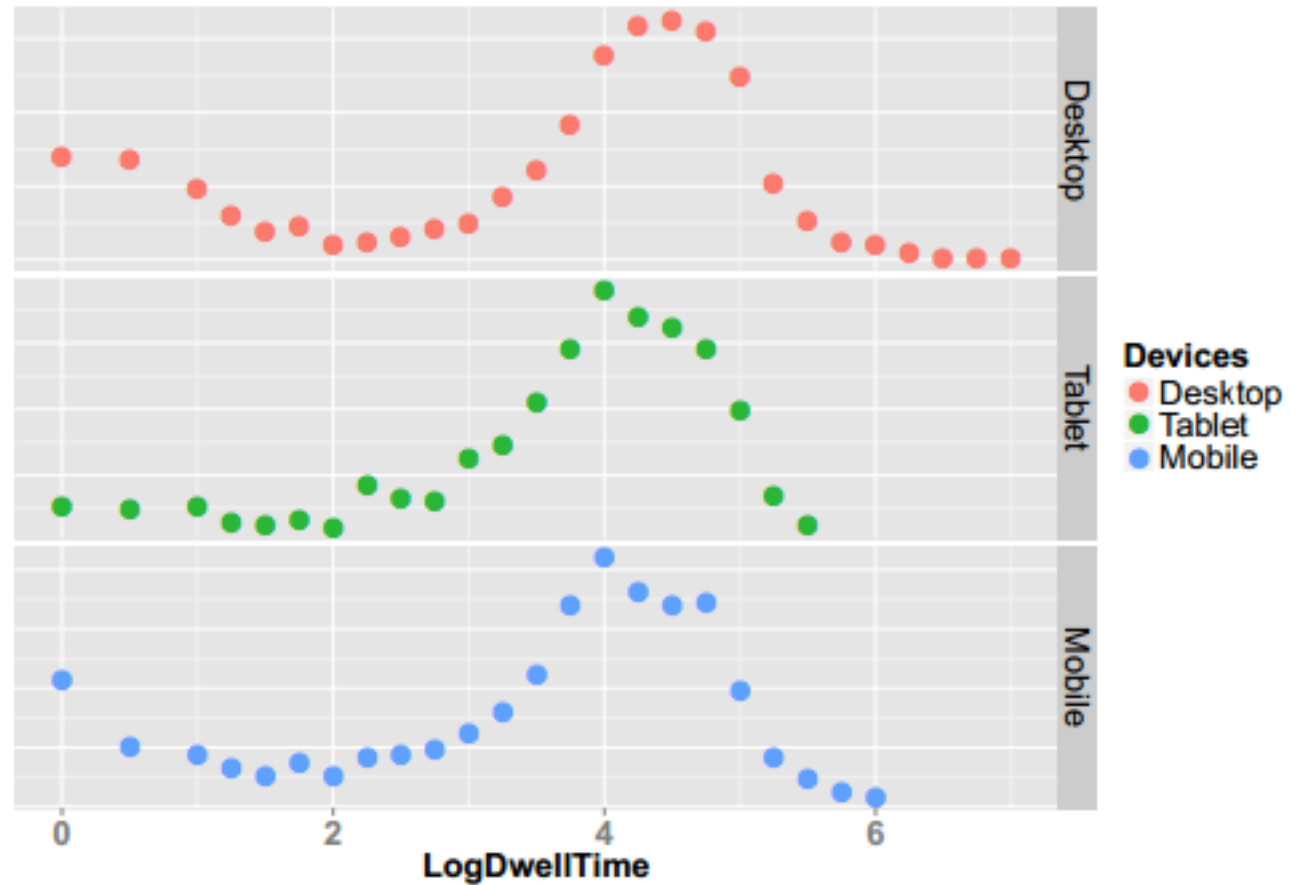
Dwell Time for personalization

- Slideshows



Dwell Time for personalization

- Videos



Evaluacion: Features

Table 3: Features and corresponding weights for predicted dwell time. The features are shown in the order of magnitude of weights. The left column shows positive weights and the right negative weights.

Name	Weight	Name	Weight
Desktop	1.280	Apparel	-0.001
Mobile	1.033	Hobbies	-0.010
Tablet	0.946	Travel & Tourism	-0.039
Content Length	0.218	Technology	-0.040
Transportation	0.136	Environment	-0.065
Politics	0.130	Beauty	-0.094
Science	0.111	Finance	-0.151
Culture	0.100	Food	-0.173
Real Estate	0.088	Entertainment	-0.191

Evaluacion

Table 4: Offline Performance for Learning to Rank

Signal	MAP	NDCG	NDCG@10
Click as Target	0.4111	0.6125	0.5680
Dwell Time as Target	0.4210	0.6201	0.5793
Dwell Time as Weight	0.4232	0.6226	0.5820

Table 5: Performance for Collaborative Filtering

Performance for Monthly Prediction

Signal	MAP	NDCG	NDCG@10
Click as Target	0.3773	0.7439	0.7434
Dwell Time as Target	0.3779	0.7457	0.7451

Performance for Weekly Prediction

Signal	MAP	NDCG	NDCG@10
Click as Target	0.6275	0.5820	0.5813
Dwell Time as Target	0.6287	0.5832	0.5826

Performance for Daily Prediction

Signal	MAP	NDCG	NDCG@10
Click as Target	0.6275	0.5578	0.5570
Dwell Time as Target	0.6648	0.5596	0.5589

Johnson's NIPS 2014

$$c = 1 + \alpha \log(1 + r_{ui}/\epsilon)$$

By making the assumption that all entries of R are independent we derive the likelihood of our observations R given the parameters X , Y , and β as:

$$\mathcal{L}(R | X, Y, \beta) = \prod_{u,i} p(l_{ui} | x_u, y_i, \beta_u, \beta_i)^{\alpha r_{ui}} (1 - p(l_{ui} | x_u, y_i, \beta_u, \beta_i)) \quad (2)$$

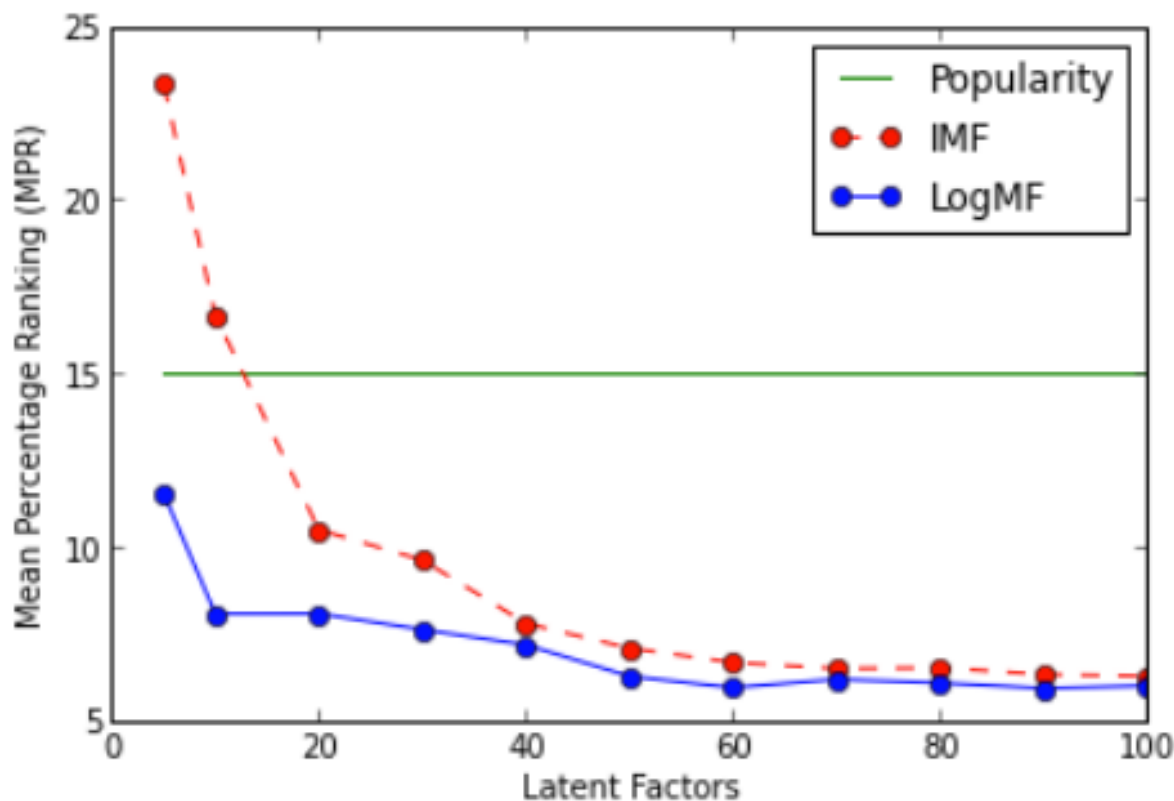
Additionally, we place zero-mean spherical Gaussian priors on user and item latent factor vectors to help regularize the model and avoid over fitting to the training data.

$$p(X | \sigma^2) = \prod_u \mathcal{N}(x_u | 0, \sigma_u^2 I), \quad p(Y | \sigma^2) = \prod_i \mathcal{N}(y_i | 0, \sigma_i^2 I)$$

Taking the log of the posterior and replacing constant terms with a scaling parameter λ we arrive at the following.

$$\log p(X, Y, \beta | R) = \sum_{u,i} \alpha r_{ui} (x_u y_i^T + \beta_u + \beta_i) - (1 + \alpha r_{ui}) \log(1 + \exp(x_u y_i^T + \beta_u + \beta_i)) - \frac{\lambda}{2} \|x_u\|^2 - \frac{\lambda}{2} \|y_i\|^2 \quad (3)$$

Evaluación Logistic Matrix Factorization



$$MPR = \frac{\sum_{ui} r_{ui}^t \text{rank}_{ui}}{\sum_{ui} r_{ui}^t}$$

Figure 2: MPR for popularity baseline, IMF, and Logistic MF using streaming count data for 50k users and the top 10k artists on Spotify

Resumen

- Feedback Implícito es una importante variable a considerar para hacer recomendaciones, pero también incorpora ruido y la evaluación se puede hacer menos clara de interpretar.
- Podemos considerar muchos tipos de implicit feedback: clicks, tiempo en la página, y otras acciones.
- Distintos dispositivos permiten contextualizar formas distintas de personalización.