Retroalimentación Implícita

IIC 3633 - Sistemas Recomendadores - PUC Chile

Denis Parra
Profesor Asistente, DCC, PUC CHile

Tabla de Contenidos

- Ideas de Proyectos de Investigación
  - Moviecity Dataset
  - Recomendaciones de recetas: Allrecipes.com
  - Modelos gráficos (topic models, content & tags)
  - RecSys challenge 2016: recomendación de ofertas de trabajo
  - Implicit Feedback paper
- Resumen de lecturas
- Retroalimentación implícita

Resumen blogposts de lecturas

- Leer un paper como "crítica". Aún no se entiende la tarea de revisar estos papers en una gran cantidad de casos:
  - Diferenciar entre un "primary study" y un "review". Hacer comentarios de modo diferente.
  - Pocos blogs rescatan los pro y contra de los artículos dados para lectura, se centran exclusivamente en encontrar pequeños errores. Por sí mismo estos no es incorrecto, pero es claramente una revisión incompleta.
  - Si esos ertículos han sido dados a lectura es porque han sido citados una buena cantidad de veces y algo positivo e importante deben tener. La idea es rescatar cosas positivas y negativas en base al contexto en que fueron publicados.
  - Sobre content-based recommender systems, se criticó en algunos casos que los autores "...no hubiesen usado técnicas más avanzadas como word2vec para encontrar items parecidos...". Word2vec es el 2012, el paper "Content-based recommendation systems" es el 2007.
- Dos observaciones interesantes:
  - Una observación importante es que varios papers de los publicados antes del 2010 no presentan un análisis de complejidad computacional cuando se presentan algoritmos.
  - Comentario sobre limitaciones de los sistemas basados en contenido: pueden captar expresión de sentimiento que se aplica en la letra de músicas y poesía?

Retroalimentación Implícita

- Hasta hace pocos años, la gran mayoría de los modelos avanzados de recomendación, basados en factorización matricial, dependían de preferencias explícitas del usuario en forma de ratings.
- Pero los ratings (explicit feedback) son difíciles de obtener.
- Por otro lado, tenemos la opción de usar feedback implícito, pero con los siguientes problemas:
    - No hay feedback negativo.
    - Contiene ruido.
    - Es difícil cuantificar preferencia y confianza en esas preferencias.
    - Hay una carencia de métricas de evaluación (RMSE y MAE no funcionarían bien)

ref: Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In ICDM'08. Eighth IEEE Internatioonal Conference on Data Mining (pp. 263-272).

Paper 1

Hu, Y., Koren, Y., & Volinsky, C. (2008).

Collaborative filtering for implicit feedback datasets.

In ICDM'08. Eighth IEEE Internatioonal Conference on Data Mining (pp. 263-272).

Ratings : recurso escaso

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

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

Modelo Implicit Feedback - Hu et al.

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

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

r_{ui} es, en este caso, el implicit feedback (e.g. plays)

- La función que esperamos minizar es, luego

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

7/39

Modelo Implicit Feedback - Hu et al. II

- Aprendizaje de parámetros (factores latentes): 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} (prefencias) 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 período de 4 semanas, 17.000 programas de TV únicos
- r_{ui} : cuantás 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 $rank_{ui}$ = 0%, el programa i ha sido predicho como el más relevante para el usuario u, y si $rank_{ui}$ = 100%, el programa i es el menos deseado. Expected percentile ranking $\overline{rank}$: the smaller the better

$$\overline{rank} = \frac{\sum_{u,i} r^t_{ui} rank_{ui}}{\sum_{u,i} r^t_{ui}}$$

Resultados I

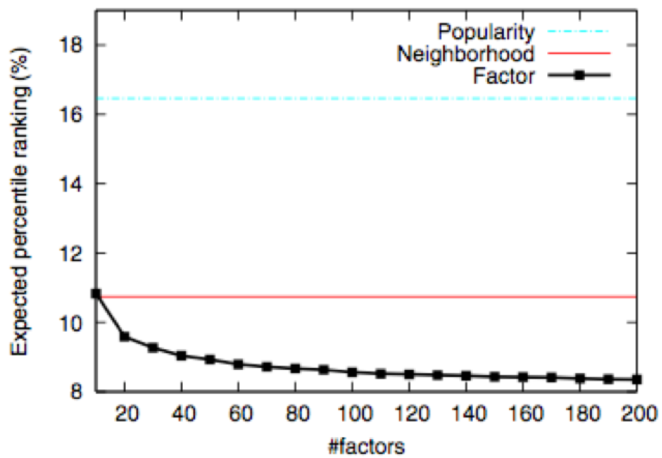


Figure 1. Comparing factor model with popularity ranking and neighborhood model.

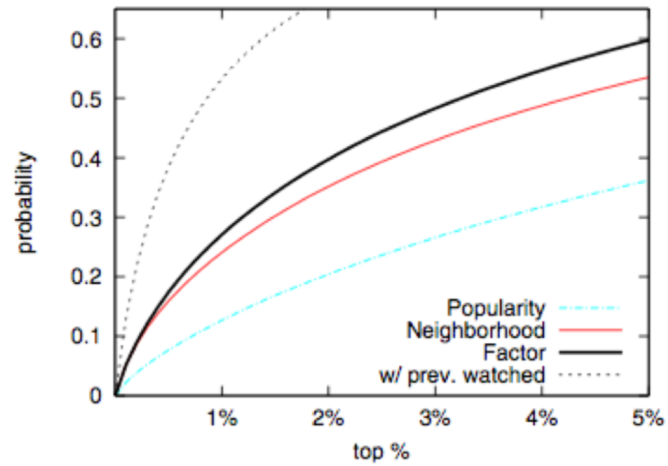


Figure 2. Cumulative distribution function of the probability that a show watched in the test set falls within top x% of recommended shows.

Resultados II

Paper 2

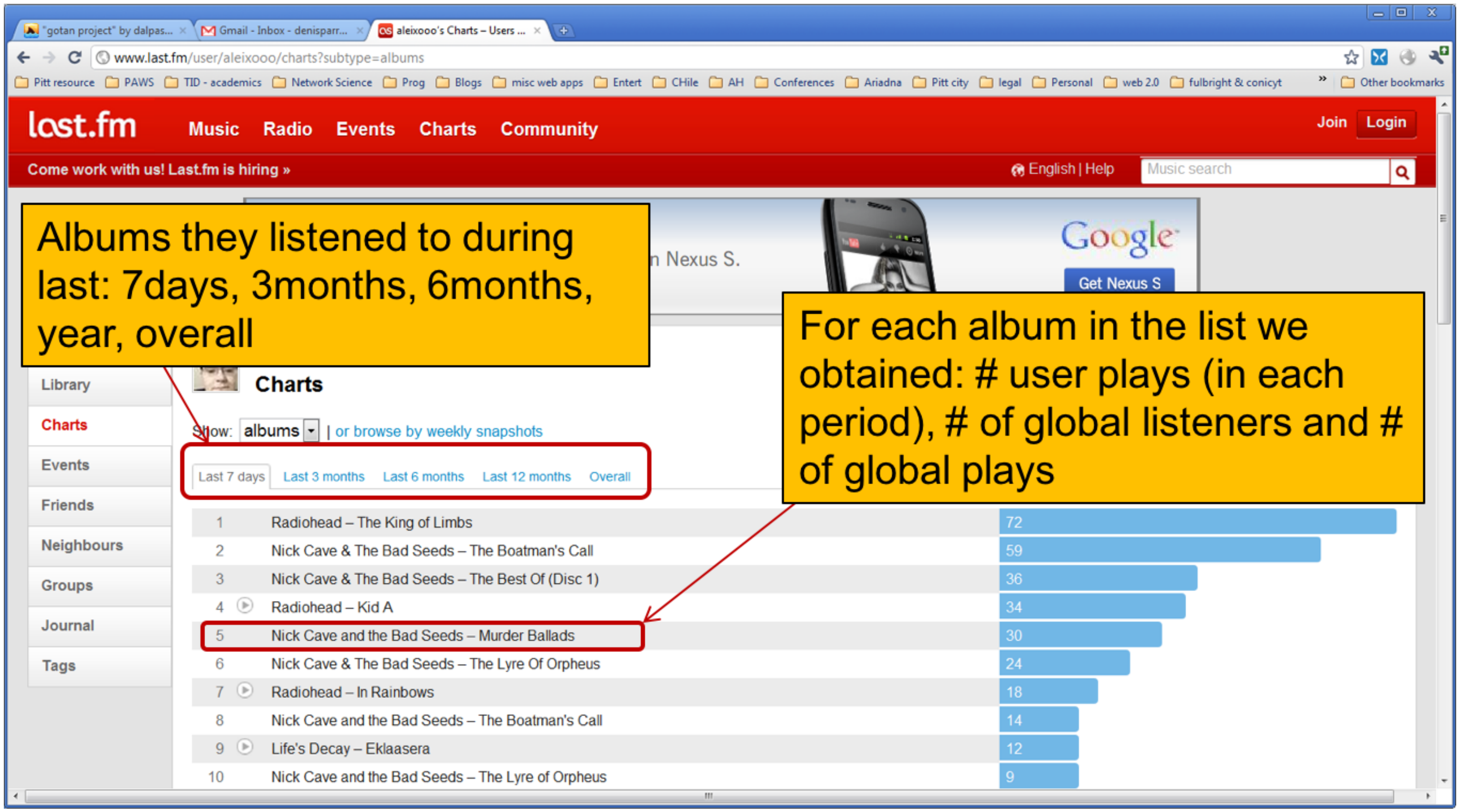
Parra, D., & Amatriain, X. (2011).
Walk the Talk: Analyzing the Relation between Implicit and Explicit Feedback for Preference Elicitation.
In User Modeling, Adaptation and Personalization (pp. 255-268). Springer Berlin Heidelberg.

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

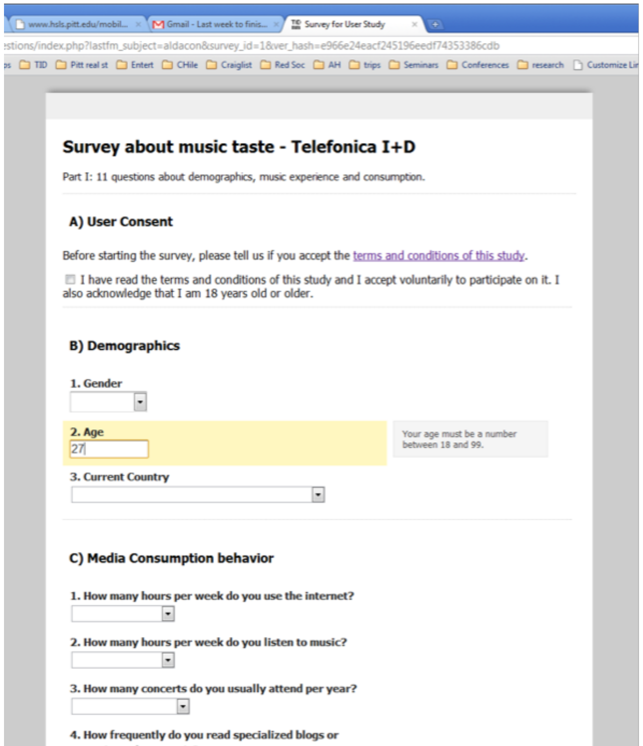
- Is it possible to map implicit behavior to explicit preference (ratings)?
- Which variables better account for the amount of times a user listens to online albums? [Baltrunas & Amatriain CARS '09 workshop – RecSys 2009.]
- OUR APPROACH: Study with Last.fm users
    - Part I: Ask users to rate 100 albums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

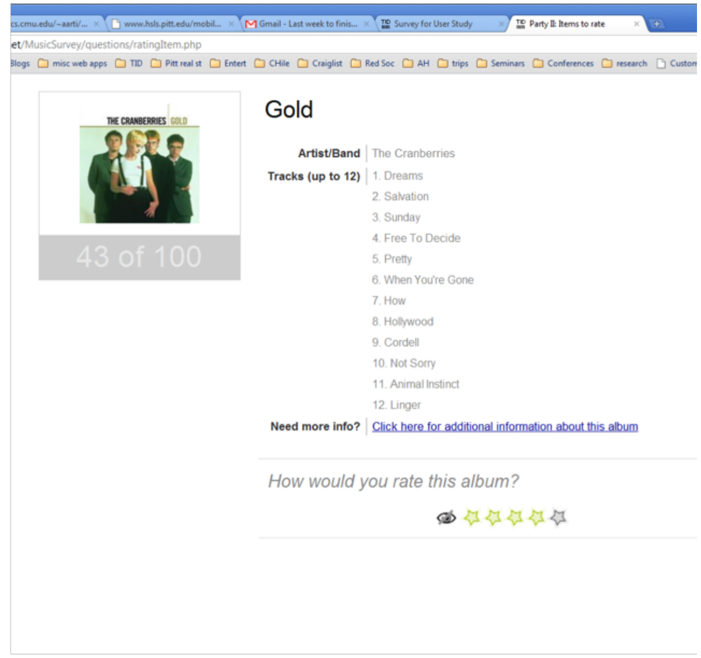
Walk the Talk (2011)



Walk the Talk - II

- Requisitos para participar en estudio: > 18años, scrobblings > 5000





16/39

Muestreo de Datos para estudio de Usuario

- Cuántos y qué items (álbums) deberian ver los usuarios?
  - **Implicit Feedback (IF)**: playcount for a user on a given album. Changed to scale [1-3], 3 means being more listened to.
  - **Global Popularity (GP)**: global playcount for all users on a given album [1-3]. Changed to scale [1-3], 3 means being more listened to.
  - **Recentness (R)** : time elapsed since user played a given album. Changed to scale [1-3], 3 means being listened to more recently.

GP	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3
IF	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3
R	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
%	8,4	6	6,8	5,5	4	5,5	3,7	2,5	4,6	6,4	3,5	4,2	3,8	2	3,3	2	1	2,2	5,8	3	4	3,3	1,6	2,6	1,6	1	2

Table 1. Distribution of items in different bins. GP: global popularity, IF: implicit feedback, R: recentness, #: number of items in the bin, % : percentage of items in the bin.

Análisis de Regresión

– Model 1: $r_{iu} = \beta_0 + \beta_1 \cdot if_{iu}$
– Model 2: $r_{iu} = \beta_0 + \beta_1 \cdot if_{iu} + \beta_2 \cdot re_{iu}$
– Model 3: $r_{iu} = \beta_0 + \beta_1 \cdot if_{iu} + \beta_2 \cdot re_{iu} + \beta_3 \cdot gp_i$
– Model 4: $r_{iu} = \beta_0 + \beta_1 \cdot if_{iu} + \beta_2 \cdot re_{iu} + \beta_3 \cdot if_{iu} \cdot re_{iu}$

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 R2 in more than 10% [1 -> 2]
- Including GP increases R2, 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]

18/39

Análisis de Regresión 2

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

- RMSE1: Considera los ratings = 0.
- 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 recentness 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.

Parte II

- 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

Supuestos en el estudio I

- Linear Regression did not account for the nested nature of ratings

User 1 . . . User n

1 3 5 3 0 4 5 2 2 1 5 4
3 2

3 2 1 0 4 5 2 5 4 3 2 1 3
5

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

Modelo II: 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 Logistic Regression Mapping

- Model

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

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

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

- Predicted values

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

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

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

Paper 3

Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014.

Beyond clicks: dwell time for personalization.

ACM RecSys 2014.

Dwell Time

- Method to consume fine-grained dwell-time at web scale
  - Focus Blur (FB) and Last Event (LE) 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

Dwell Time II

- 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

Eventos: Server y Client-Side

Table 1: Client-side Logging Example

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

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

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 para Distintos Dispositivos

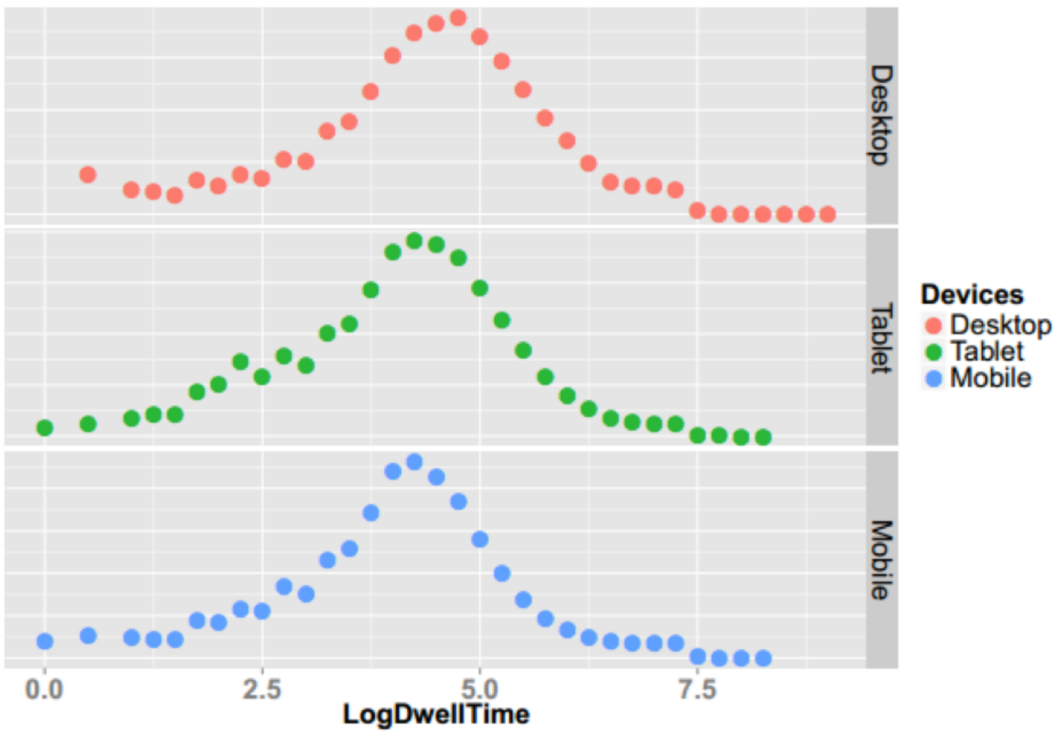
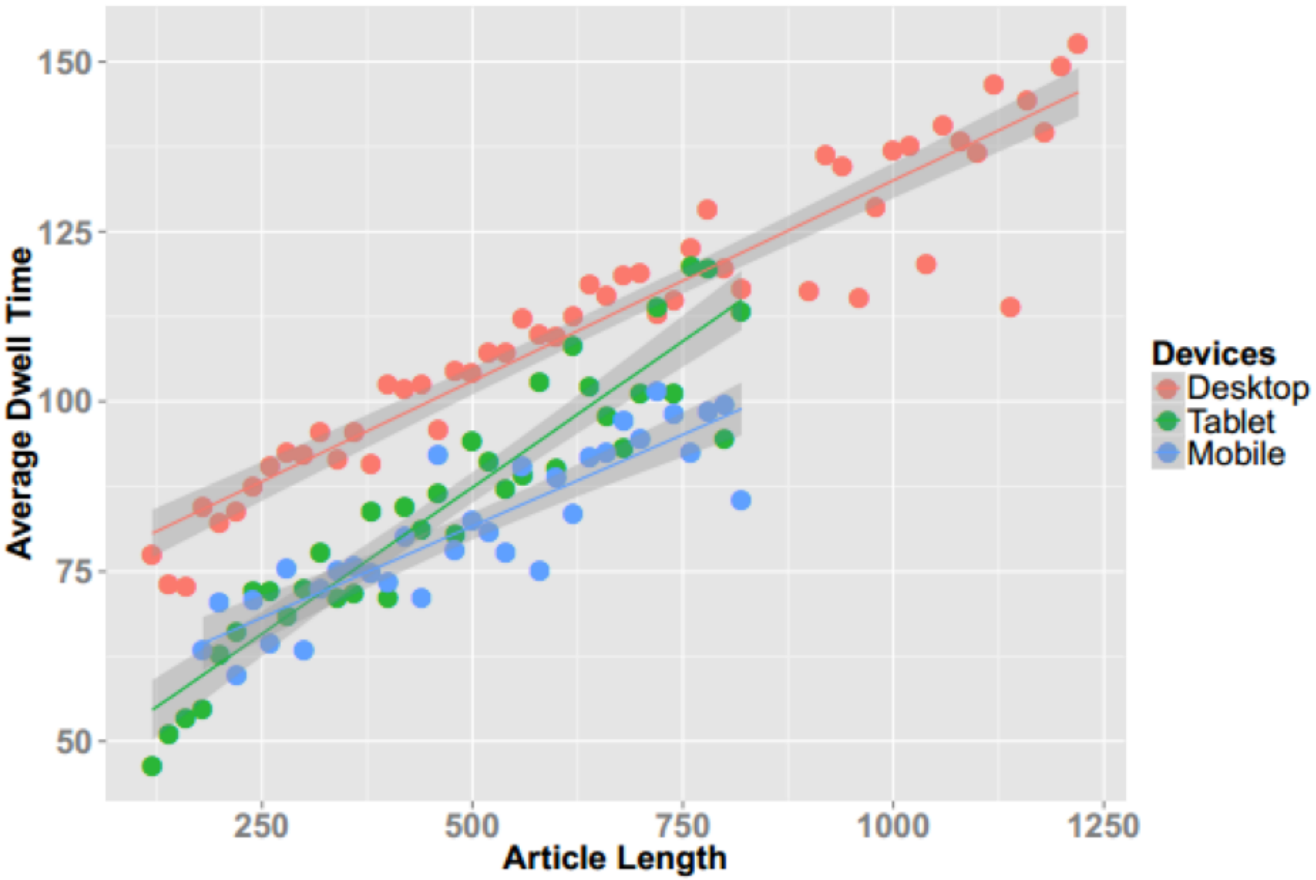


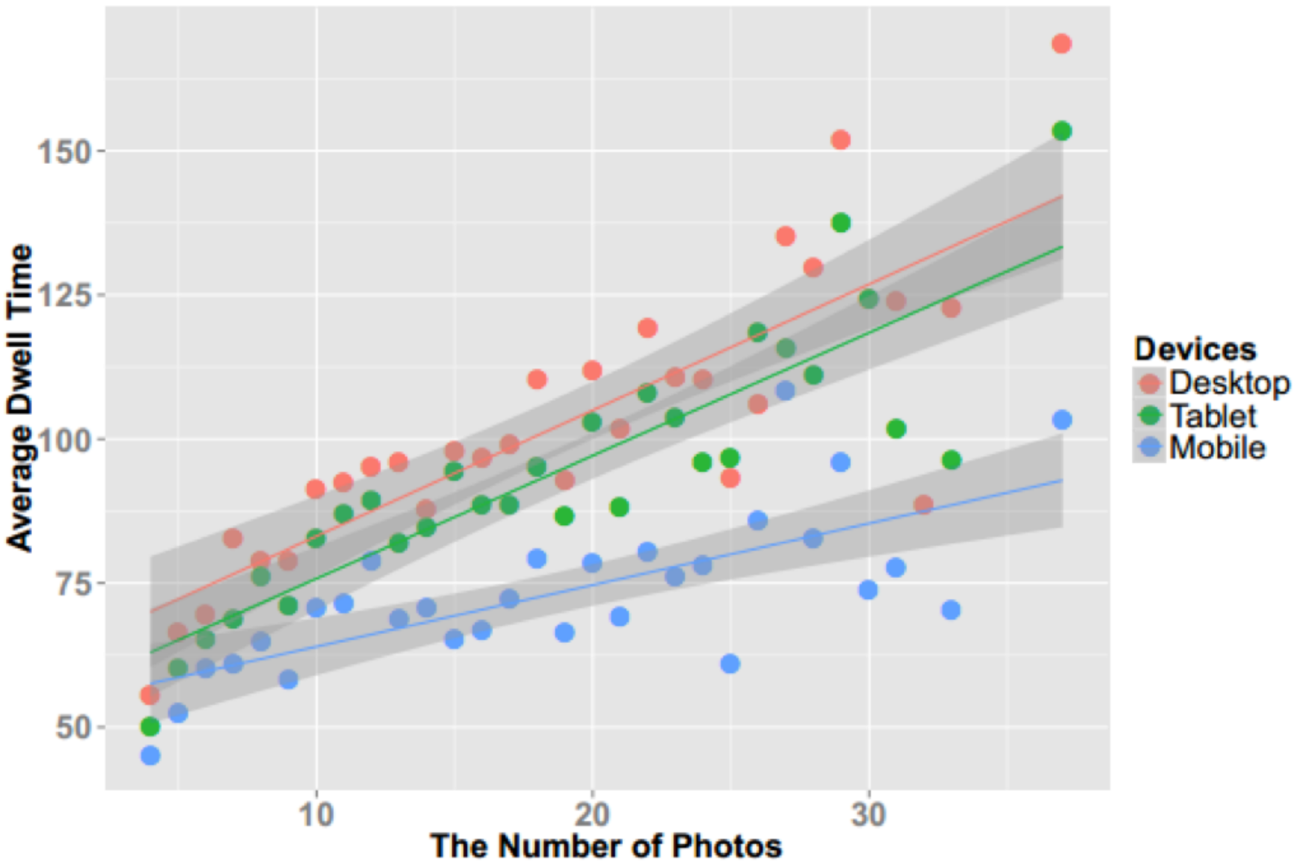
Figure 2: The (un)normalized distribution of log of dwell time for articles across different devices. The X-axis is the log of dwell time and the Y-axis is the counts (removed for proprietary reasons).

32/39

Dwell Time vs. Largo del articulo

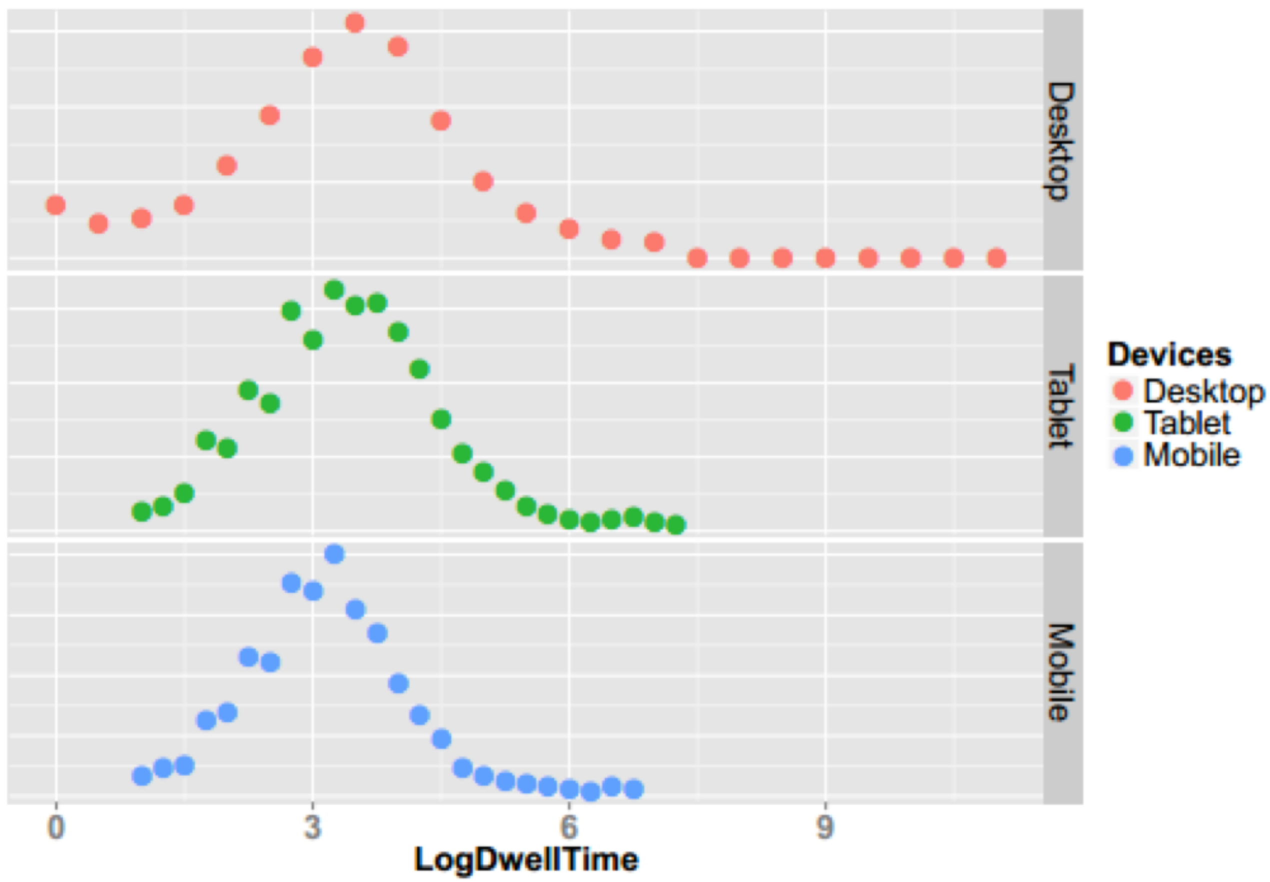


Dwell Time vs. Número de Fotos

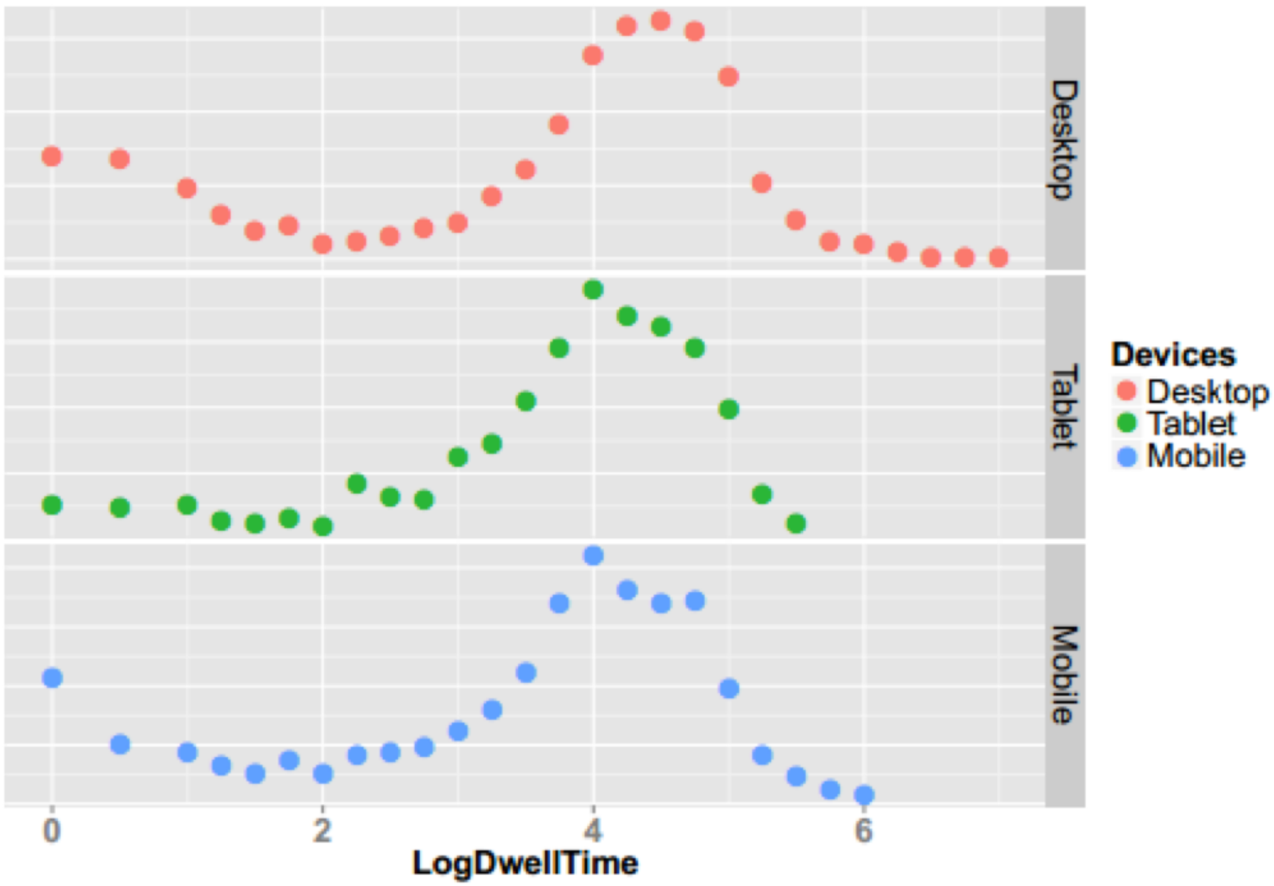


34/39

Slideshows en Distintos Dispositivos



Consumo de Videos en Distintos Dispositivos



36/39

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

Evaluación

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

Referencias

- Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In ICDM'08. Eighth IEEE Internatioonal Conference on Data Mining (pp. 263-272).
- Parra, D., & Amatriain, X. (2011). Walk the Talk: Analyzing the Relation between Implicit and Explicit Feedback for Preference Elicitation. In User Modeling, Adaptation and Personalization (pp. 255-268). Springer Berlin Heidelberg.
- Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. ACM RecSys 2014.