Métrica *Beyond-Accuracy Personal*: Evaluando Gustos en la Recomendación

Benjamín Fuentes
Pontificia Universidad Católica de Chile
Santiago, Chile
bgfuentes@uc.cl

Álvaro Labarca
Pontificia Universidad Católica de Chile
Santiago, Chile
aalabarca@uc.cl

Resumen

Las métricas *beyond-accuracy* están cobrando una relevancia cada vez más importante en el campo de los sistemas recomendadores. En la presente investigación se propone una nueva métrica *beyond-accuracy* llamada *personal*, la cual intenta medir qué tan personal se siente una recomendación dada a un usuario. Esta métrica es comparada con otras métricas de la literatura, encontrándose una baja correlación con ellas, lo cual valida su existencia junto a las métricas actuales. Finalmente se propone un *framework* que busca identificar para cada usuario la importancia de las métricas estudiadas en su satisfacción, el cual es testeado con ciertos voluntarios para testear el funcionamiento de este.

1. INTRODUCCIÓN

En un inicio, la mayoría de los estudios en el campo de sistemas recomendadores, estaban enfocados únicamente a mejorar más y más la precisión de los algoritmos, buscando que las predicciones realizadas se acerquen lo más posible a los valores reales.[1] Sin embargo, estudios más recientes han indicado que este enfoque no solo puede no entregar la mejor recomendación, sino que incluso puede ser perjudicial para la calidad del algoritmo y sus recomendaciones.[2] Para mejorar la satisfacción del usuario con el sistema recomendador y de esta manera generar recomendaciones más adecuadas al gusto de los usuarios, muchos investigadores recientemente se han enfocado en desarrollar y estudiar métricas *beyond-accuracy*, que buscan medir otras cualidades de las listas de recomendación más allá de su precisión.[1]
Las métricas *beyond-accuracy* más estudiadas en la literatura, se encuentran *diversity*, la cual se enfoca en determinar qué tan diversos son los elementos contenidos dentro de la lista,[3] *novelty*, que estudia si los elementos dentro de la lista de recomendación son "novedosos" o no tan populares[1], *serendipity* la cual estudia si las recomendaciones pueden ser consideradas como una "sorpresa placentera"[4] y *coverage* que observa la proporción del set total de ítems que termina siendo recomendada[5].

En esta investigación se identificó una cualidad en las listas de recomendación que podría también tener un impacto sobre la satisfacción de los usuarios respecto a las listas de recomendación entregadas. Si un algoritmo predice que un usuario calificará una película como *El Padrino* con un 4,75 y predice para una película como *Annihilation* un 4,70. Supongamos ahora que en realidad el usuario conoce estas dos películas y a ambas las calificó de igual manera con un 5,0. Entonces, si un algoritmo de recomendación le entrega a este usuario dos listas, en las cuales una contiene a *El Padrino* y la otra a *Annihilation*, se podría esperar que el usuario quede más satisfecho con la segunda lista, no debido a la popularidad de la película ni a su factor sorpresa, sino que, debido a que el usuario sabe que *El Padrino* es una película altamente aclamada por la crítica, con calificaciones elevadas en todos los sitios, mientras que *Annihilation* es una película que recibió una recepción más variada, de esta manera, la lista que contiene a *Annihilation* se sentirá más *personal*, es decir, se sentirá que fue recomendada debido a sus gustos personales y características como individuo, en lugar de la calidad general y recepción de la película.

Por último, se propondrá un *framework* que permitirá estudiar las diferencias entre cada usuario respecto a qué cualidad de las listas de recomendación son más influyentes en la satisfacción percibida por los usuarios, de esta manera entregando listas de recomendación que sean personalizadas no solo en los elementos que contienen, sino que también en los objetivos que buscan cumplir.

2. REVISIÓN DEL ESTADO DEL ARTE

Se han realizado diversos estudios dentro del campo de las métricas *beyond-accuracy* y de su relación con la satisfacción percibida por los usuarios.

Kaminskas et al. (2016)[1] proporcionan un *survey* con los principales resultados y conclusiones en el área de las métricas *beyond-accuracy* hasta ese momento. También entrega fórmulas para calcular la novelty para un ítem $i(N_i)$ y para el valor de *novelty* (N), para una lista de recomendación R definidos de la siguiente forma:

$$N_i = -\log_2 p(i)$$

$$N(R) = \frac{\sum_{i \in R} N_i}{|R|}$$

Donde $p(i)$ es la fracción de usuarios que han calificado i

Chen et al. (2019)[6] nos entrega una manera de calcular la *serendipity* mediante la siguiente fórmula:

$$SR_i = Unexpectedness(i) \times relevance(i)$$

Donde

$$Unexpectedness(i, H) = \frac{1}{I} \sum_{h \in H} CosineSimilarity(i, h)$$

Donde i es el elemento (película) para la que se está midiendo y H es el set de ítems con los que el usuario ha interactuado. Y $relevance(i)$ indica si el usuario ha interactuado con el ítem en cuestión. Luego, para una lista de ítems recomendados, I, se define

$$SR = \frac{\sum_{i \in I} SR_i}{|I|}$$

Es decir, el valor promedio de la *serendipity* de cada ítem. Estas dos métricas serán usadas en esta investigación para calcular la *serendipity*.

Finalmente, para *diversity*, Smith & McClave (2001), proponen las métricas D_i para un ítem i y $D(R)$ para una lista de recomendación R de la forma:

$$D_i = \frac{\sum_{j \in R \setminus \{i\}} dist(i, j)}{|R| - 1}$$

$$D = \frac{\sum_{i \in R} D_i}{|R|}$$

3. DATASET

El dataset utilizado en esta investigación fue el de *MovieLens 25M*. Este consta con 25 millones de ratings para 62,000 película distintas y 162,000 usuarios. La densidad de los datos es bastante baja y es equivalente a 0,003. Se realizo un *split* de los datos en donde un 25 % de estos fueron considerados para un set de *test*. Para lograr hacer una separación correcta de los datos se consideró la distribución de las interacciones hecha por cada usuario, en donde una interacción equivale al rating hecho de una película. Esto se muestra en la figura 10.

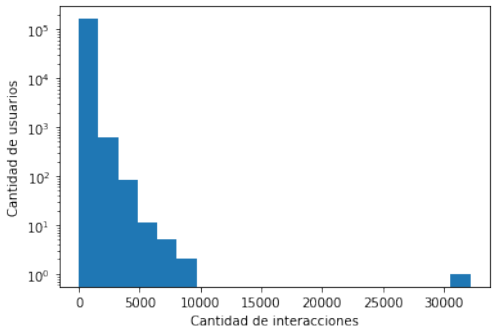

Figura 1: Distribución de interacciones por usuario

El usuario con menos interacciones tiene 20, y de la figura se logra ver que la mayor parte de los usuarios tiene entre 20 y 100. De esta forma, para la separación de los datos se exigió que en el set de entrenamiento se encuentren todos los usuarios con al menos 10 interacciones y todas los *items* al menos una vez. También es importante que existan películas contemporáneas para el entrenamiento, dado que para ciertos experimentos se utilizan los datos de usuarios externos a la base de datos que contienen ratings de películas de los últimos años. La distribución por año de estas se aprecia en la figura 5 en el apéndice, la cual cumple con la condición de tener películas actuales.

4. MÉTRICA PROPUESTA

Para medir qué tan personal una lista se sentirá para un usuario, se propone la siguiente métrica:

$$P(u, R) = \frac{\sum_{i \in R} \left(\hat{r}_{u,i} - \frac{\sum_{v \in U} r_{v,i}}{\sum_{v \in U} I_{v,i}} \right)}{|R|}$$

Donde $\hat{r}_{u,i}$ indica la calificación predicha del item i para el usuario u, $r_{v,i}$ indica la calificación dada al item i por el usuario v (se asigna 0 si el usuario no ha visto la película), $I_{v,i}$ indica si el usuario interactuó con el ítem y U es el conjunto de usuarios en el dataset.

Similarmente a la definición establecida para las otras métricas *beyond-accuracy*, se define una versión de la métrica para cada ítem i:

$$P_i = \left(\hat{r}_{u,i} - \frac{\sum_{v \in U} r_{v,i}}{\sum_{v \in U} I_{v,i}} \right)$$

Lo que mide la métrica es la diferencia entre la calificación predicha para una película y la calificación promedio que ha recibido. De esta manera, si el sistema predice que la calificación que le pondrá el usuario es mayor a la calificación promedio, entonces esta recomendación se considerará que es una recomendación personal. Por otra parte, si la predicción es menor al promedio, la métrica *personal* le asignará un valor negativo, ya que indica que al usuario le gustará menos que a una persona promedio.

Es importante comparar estas métricas con las existentes:

- *Diversity*: *Personal* se diferencia a esta, debido a que no busca lograr que los elementos dentro de la lista de recomendación sean distintos entre sí, lo que busca es recomendar elementos que no sean "esperados" al considerar la valoración general de estos elementos. De esta manera, una lista de recomendación que contenga ítems similares entre sí, pero cuyos ítems tengan una predicción para el usuario mucho mayor al promedio de cada ítem, tendrá un valor bajo de *diversity*, pero alto según nuestra métrica.

- *Novelty*: *Personal*es distinta porque no busca realizar recomendaciones que no sean populares en términos de cantidad de usuarios que han interactuado con ella, sino que busca recomendaciones que no sean bien valoradas respecto a su valoración explícita.

- *Serendipity*: Si el usuario vio y le gustó una película que no es altamente valorada, al recomendar una película similar a esta, se tendrá baja *serendipity* (ya que no será una sorpresa), pero alto *personal*.

5. METODOLOGÍA UTILIZADA

5.1. Elección de Modelo

Para los experimentos realizados entorno a la métrica definida anteriormente, es necesario utilizar un modelo de recomendación. Dado que esta investigación no se basa en entregar las recomendaciones con mayor precisión, se consideraron modelos conocidos que pudieran generar recomendaciones aceptables. Además, dado que la base de datos utilizada es bastante grande, se descartaron opciones como algoritmos basados en *Knn*, dado que necesitaban una alta memoria y tiempo para entrenarse.

Con este fin, se realiza un experimento preliminar que permita comparar modelos y evaluar cual de ellos tiene un buen desempeño y logra generar listas de recomendaciones en un tiempo aceptable. También, se espera averiguar que algoritmo presenta una mayor afinidad con la métrica *personal* definida. Para esto, se toma un muestreo aleatorio de usuarios y cada modelo entregará una lista de recomendación para cada usuario con la cual se calcula el valor del *personal* y luego se obtiene un promedio de la muestra. Los modelos que fueron considerados para este experimento son *FunkSVD* , un filtro colaborativo utilizando factorización de matrices no negativas (*NMF*) y un algoritmo basado en *co-clustering*.

5.2. Contraste de *personal* con otras métricas

La existencia de otras métricas *beyond-accuracy*, que buscan describir características similares a las que describe la métrica *personal* definida, requiere comprobar que es necesaria esta nueva forma de evaluar la listas de recomendación para un usuario. Con este fin, se realiza un experimento que busque contrastar la métrica propuesta con la ya existente.

Para lograr lo señalado anteriormente, se seleccionan 100 usuarios aleatoriamente, en donde a cada uno de ellos se les recomienda una lista de películas. Con esta recomendación, se calcula el valor de las métricas por usuario y se registra como una muestra, lo que permitirá que con las 100 muestras se calcule una correlación entre los valores.

Por otra parte, se quiere identificar la relación que existe entre *personal* y alguna métrica de precisión. Para esto, se realiza un experimento similar al detallado anteriormente, pero registrando el valor de

$AP@N$ que tiene cada lista recomendada para los usuarios definida como:

$$AP = \frac{\sum_k P@k * \text{rel}(k)}{\#\text{elementos\_relevantes}}$$

en donde $P@k$ es la precisión en los primeros k elementos de la lista. De esta forma se podrá estudiar gráficamente como se relaciona *personal* con la precisión.

5.3. Propuesta de *Framework*

En esta sección se buscará definir un *framework* que busque generar recomendaciones para un usuario, basado en la importancia que cada usuario le asigne a cada métrica *beyond-accuracy* en cuanto a su satisfacción.

Con este propósito, se propone el siguiente *framework*:

1. En primer lugar, se le entrega a un usuario una serie de listas de recomendaciones de películas y se les pide que a cada lista se le asigne un valor numérico correspondiente a la satisfacción percibida por los usuarios respecto a las listas.

2. Para cada lista entregada al usuario, se almacenan los valores de *personal* (**P**), *serendipity* (**SR**), *diversity* (**D**), *novelty* **N** y el valor de la satisfacción (**S**) entregada por el usuario.

3. Habiéndose registrado una serie de estos valores para un usuario, se procede a realizar un modelo de regresión múltiple que busque predecir la satisfacción que percibirá un usuario respecto a una lista a partir de los valores de las métricas *beyond-accuracy* medidas:

$$\beta_0 + \beta_1 * P + \beta_2 * SR + \beta_3 * N + \beta_4 * D = S$$

4. Los valores $\{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4\}$ representan entonces la importancia estimada que cada usuario le asigna a cada métrica *beyond-accuracy* en la satisfacción final que tendrán con la lista entregada. Se puede predecir entonces la satisfacción estimada $\hat{S}$ que tendrá un usuario con una lista habiendo calculado las 4 métricas *beyond-accuracy* estudiadas de la siguiente manera:

$$\hat{\beta}_0 + \hat{\beta}_1 * P + \hat{\beta}_2 * SR + \hat{\beta}_3 * N + \hat{\beta}_4 * D = \hat{S}$$

   Como se puede observar, el valor de cada parámetro $\hat{\beta}_i$ dependerá de cada usuario, para considerar la importancia personal que le asigna cada usuario a cada una de estas métricas, por lo que se deberá realizar una regresión múltiple individual para cada usuario.

5. Finalmente, se elige un modelo recomendador, el cual entrega una lista *top-n* con las n películas que predice que el usuario le pondrá mejor calificación. Con estas predicciones, se busca generar una lista que aumente el valor predicho de la satisfacción, se realiza una adaptación de la técnica MMR de *rerank* propuesta por Carbonell y Goldstein[7] que define una función objetivo f_{obj} para cada película i como:

$$f_{obj}(i) = \alpha \cdot \hat{S}_i + (1 - \alpha) \cdot \texttt{pred}_i$$

   Donde α es un valor entre 0 y 1, $\texttt{pred}_i$ es la predicción de la calificación entregada por el modelo para el usuario estudiado y $\hat{S}_i$ es la estimación de la satisfacción calculada para la película i, la cual se calcula como:

$$\hat{\beta}_0 + \hat{\beta}_1 * P_i + \hat{\beta}_2 * SR_i + \hat{\beta}_3 * N_i + \hat{\beta}_4 * D_i = \hat{S}_i$$

   Es decir, para una película i, su función objetivo se calcula como una combinación lineal simple entre el valor predicho por el modelo y la satisfacción predicha para el usuario. Finalmente se le entrega al usuario una lista ordenada según este valor con las m películas con un el mayor valor calculado de f_{obj} (con $m << n$).
   De esta manera, la lista final entregada al usuario considera tanto la calificación predicha como la satisfacción estimada que el usuario tendrá con esta. La selección original de las n mejores películas se realiza para realizar un filtro inicial de las películas que son candidatas y evitar películas que generen una *accuracy* muy baja.

5.4. Diseño de Experimento

Para obtener resultados concluyentes respecto a la eficacia del *framework* propuesto y la validez de la métrica *personal* para aumentar la satisfacción, se necesitaría realizar un estudio con una gran cantidad de recursos y tiempo, para obtener un número significativo de datos y a partir de esto, obtener las conclusiones pertinentes.
Sin embargo, a modo de testear la validez y el correcto funcionamiento del *framework* propuesto, además de

obtener resultados que puedan indicar una tendencia, sin que esta sea concluyente, el *framework* fue testeado con 5 voluntarios.

1. Para el testeo del *framework*, cada usuario proporcionó su calificación de al menos 100 películas, las cuales fueron añadidas al dataset de *MovieLens 25M* con el cual fue entrenado el algoritmo de `Funk-SVD`, para obtener predicciones de las calificaciones que el algoritmo estima que los usuarios le pondrían a cada película.

2. Con el modelo entrenado, se genera una lista *top-n* con $n = 500$, la que llamamos $\mathcal{P}$. $\mathcal{P}$ considera tanto películas que el usuario ya ha visto como películas que no ha visto

3. Se genera un subset $\mathbf{p} \subset \mathcal{P}$ a partir de un muestreo aleatorio del 10 % de las películas de $\mathcal{P}$

4. Se selecciona aleatoriamente una métrica $\mathbf{M}_i$ de (P_i, SR_i, N_i, D_i) y un valor de α entre 0,1 y 1 los cuales se usan para realizar un *rerank* de $\mathbf{p}$ de tamaño 7

$$f_{obj}(i) = \alpha \cdot \mathbf{M}_i + (1 - \alpha) \cdot \texttt{pred}_i$$

5. Para esta lista, se calculan y almacenan los valores de *personal*, *diversity*, *serendipity* y *novelty*

6. Se repiten 50 veces los pasos $3 - 6$, generando así, un set con 50 listas, cada una con 7 recomendaciones, el cual se le entrega al usuario, y se le entrega el contexto "Imagina que, a partir de tus calificaciones pasadas, un servicio de recomendación define que estas 7 películas son las más te gustarían a ti, considerando esto, califica numéricamente tu satisfacción para cada una de estas listas".

7. Finalmente, se realiza la regresión múltiple y el *rerank* final de tamaño 10 con el set $\mathcal{P}$ y se le entregan al mismo usuario, en orden aleatorio, las listas obtenidas según la función objetivo para valores de α de 0 (equivalente a realizar el *top-n* del algoritmo `Funk-SVD`), 0,3, 0,5, 0,7 y 1. Y se le pide al usuario que ordene según su preferencia estas 5 listas finales.

El objetivo de seleccionar parámetros aleatorios para el *rerank* sirve para disminuir la probabilidad de generar listas que son muy similares entre sí y obtener una mayor variación para los valores de las métricas.

6. ANÁLISIS DE PARÁMETROS

Para determinar los parámetros utilizados en los algoritmos para recomendar, se realizaron entrenamientos con un conjunto pequeño de usuario con el objetivo de realizar una alta cantidad de recomendaciones y ver a priori que valor de parámetros era mejor.

Dado que se realiza un estudio con un dataset más pequeño que el utilizado en los experimentos, no se evaluarán parámetros como el valor de la tasa de aprendizaje y la cantidad de épocas, pero si se puede determinar la cantidad de factores latentes en el caso de los modelos *SVD* y *NMF* o la cantidad de *clusters* en el modelo basado en *co-clustering*.

Como el objetivo no es obtener las mejores recomendaciones posibles, sino más bien recomendaciones aceptable, se contrasto como variaba el valor de métricas objetivas como el *RMSE*, que representa la raíz del error cuadrático medio y también se analizo cuanto demoraban en entrenarse y realizar una predicción los modelos según la cantidad de factores latentes. En la figura 2 se ven los gráficos que representan estas comparaciones para el algoritmo *SVD*. Es claro que a medida que se agregan factores latentes, disminuye el error pero el tiempo aumenta linealmente. Es por esto que para cada modelo se escogió una cantidad de factores cercana a la convergencia del error y no tan alto para evitar que las recomendaciones realizadas con el dataset más grande fueran muy costosas.

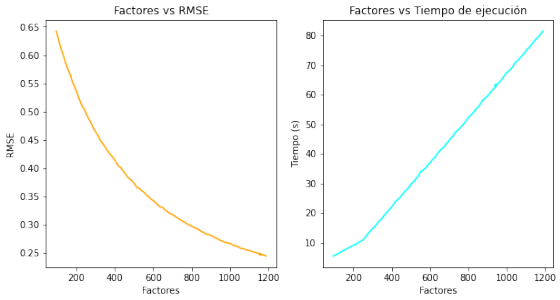

Figura 2: Relación entre la cantidad de factores latentes vs RMSE y el tiempo de entrenamiento y predicción

En el caso del algoritmo *SVD* se utilizaron 400 factores y un estudio análogo se realizó para los otros dos modelos.

7. RESULTADOS OBTENIDOS

7.1. Resultados por modelo

En la tabla 1 se muestran los valores obtenidos, para cada modelo, de *personal* promedio, el *RMSE* y el tiempo promedio que demoraba el algoritmo en realizar una predicción con las 30 mejores película para cada usuario.

Modelo	*SVD*	*NMF*	*Co-Clustering*
Personal	0.711	0.716	0.017
RMSE	0.782	0.873	0.921
Tiempo	32.34	32.15	31.99

Tabla 1: Estadísticas de resultados por modelo

Con estos resultados, se descarto el uso del método con *clusters* dada su bajo valor de *personal*. En cuanto a los otros dos modelos, obtuvieron valores muy similares de *personal* y tiempos de predicción. *SVD* obtuvo un error un poco menor al de *NMF* y dado que es un modelo con mayor documentación, fue el escogido finalmente para los experimentos siguientes.

7.2. Correlaciones

En la tabla 2 se muestra una matriz de correlación entre todas las métricas *beyond-accuracy* definidas, incluyendo a *personal*. En esta, se muestra que las correlaciones entre *personal* y el resto no son valores altos, lo que entrega justificación a la existencia de esta métrica dado que las existente no evalúan lo mismo.

	Per	Div	Ser	Nov
Personal	1	0.027	0.13	-0.25
Diversity	0.027	1	0.028	0.14
Serendipity	0.13	0.028	1	-0.22
Novelty	-0.25	0.14	-0.22	1

Tabla 2: Correlaciones entre métricas *beyond-accuracy*

En cuanto a la relación entre *personal* y la métrica $AP@20$, y se puede apreciar en la figura 3 que no existe una tendencia clara entre ambos valores. Esto tiene un impacto positivo en nuestra métrica dado que a futuro se pueden utilizar modelos que no necesitan sacrificar mucho su desempeño para poder aumentar el valor del *personal*.

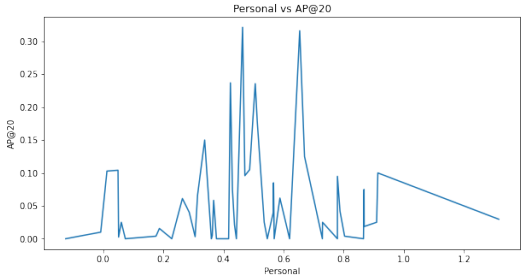

Figura 3: Relación entre *personal* y métrica de precisión

7.3. Framework

Para testear el funcionamiento del *framework*, se realizó el experimento con 5 voluntarios. Cada uno de estos voluntarios, proporcionó una cantidad distinta de calificaciones previas, para añadir a la base de datos *MovieLens 25M* para entrenar el algoritmo. La cantidad de películas que cada usuario proporcionó se detalla en la siguiente tabla:

Usuario	1	2	3	4	5
Películas	1257	391	195	273	348

Tabla 3: Cantidad de calificaciones proporcionadas por voluntarios

Se realizan tests sobre la regresión realizada para cada usuario, para verificar si se cumplen los supuestos necesarios para realizar una regresión múltiple. Los test utilizados son: *Variance Inflation Factor* (VIF) < 10 en cada métrica para la no multicolinealidad, el test de *Durbin-Watson* (DW) para la autocorrelación, el test de Shapiro-Wilkinson (SW) para la normalidad de los residuos y el test de Breusch-Pagan (BP) para la homocedasticidad. Los resultados se muestran en la siguiente tabla. Un valor de 1 indica que la hipótesis nula no fue rechazada:

U	VIF	DW	SW	BP
1	1	1	1	1
2	1	1	1	1
3	1	1	0	1
4	1	1	1	1
5	1	1	1	1

Tabla 4: Test de Hipótesis Supuestos Regresión Múltiple

Podemos ver que el único test que rechaza algún supuesto es el test de normalidad para el usuario 3, el cuál se analizará de manera visual:

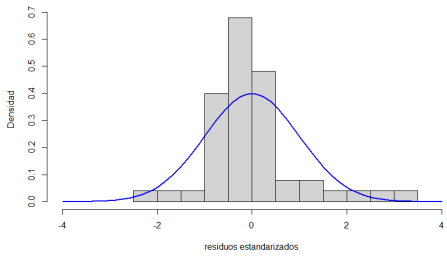

Figura 4: Distribución de densidad de residuos estandarizados

Podemos ver que a pesar de que la distribución no se ajusta correctamente a una distribución normal, visualmente sigue una tendencia que se parece aproximar a una, pero este problema nos puede indicar que quizás sea necesario recopilar más de 50 sets de datos para cada usuario. También puede indicar una menor tendencia del usuario a dar calificaciones diversas a las listas evaluadas, lo cual puede generar un problema en el estudio.

Debido a que solo un test de hipótesis fue rechazado, se procede con el experimento.

A continuación se presenta una tabla con los resultados de las regresiones realizadas:

U	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
1	41,425	4,376	−0,032	−0,026	−5,274
2	44,197	0,395	−0,079	−0,430	−0,389
3	37,729	1,798	−0,142	−0,464	2,666
4	8,677	0,967	−0,016	−0,187	0,870
5	37,258	2,771	−0,105	0,214	−0,116

Tabla 5: Parámetros Estimados Regresión Múltiple

De esta tabla se puede rescatar las diferencias en los valores obtenidos para cada usuario, principalmente en las mediciones de $\hat{\beta}_3$ y $\hat{\beta}_4$, en donde, dependiendo del usuario, puede tener valores negativos o positivos. Esto sugeriría que para un usuario, una lista más diversa o más novedosa puede ser algo positivo, mientras que para otro puede ser algo negativo. A pesar de que la cantidad de usuarios medidos es muy baja para encontrar una tendencia, datos de este estilo reafirmarían la necesidad de generar una regresión individual que se ajuste a las preferencias de cada usuario.

Otro análisis relevante es que las estimaciones del parámetro $\hat{\beta}_1$ son positivas en todas las regresiones, lo cual, realizando un estudio más grande, indicaría una relación positiva entre la métrica *personal* y la satisfacción de los usuarios.

Además, el valor del parámetro $\hat{\beta}_2$ es siempre negativo, lo cual puede sugerir que la *serendipity* sigue una relación inversa con la satisfacción para estos usuarios, contradiciendo en cierta medida las conclusiones de Zhang et al. (2012), en donde se proporcionó un sistema con mayor *serendipity*, al cual los usuarios calificaron como menos *enjoyable*, pero aun así, lo prefirieron al sistema con menor *serendipity*.

Para profundizar en este análisis, se buscó el coeficiente de correlación entre cada métrica y la satisfacción entregada:

U	*personal*	*serendipity*	*novelty*	*diversity*
1	0,726	−0,615	−0,408	−0,615
2	0,594	−0,520	−0,546	0,067
3	0,520	−0,743	−0,798	0,251
4	0,446	−0,308	−0,546	0,017
5	0,221	−0,525	−0,009	−0,408

Tabla 6: Parámetros Estimados Regresión Múltiple

Podemos ver que, la métrica *personal* siempre tuvo un coeficiente de correlación positivo respecto a la satisfacción percibida, mientras que las métricas de *serendipity* y *novelty* muestran una tendencia inversa. Por otra parte, la métrica *diversity* varió entre valores positivos y negativos, señalando las diferencias en las preferencias de cada usuario.

Finalmente, se muestran los resultados del orden según preferencia de las 5 listas L_α para distintos valores de α

Orden descendiente según satisfacción percibida:

- Usuario 1:$L_{0,3} - L_{0,5} - L_{0,7} - L_1 - L_0$
- Usuario 2:$L_1 - L_{0,7} - L_{0,3} - L_{0,5} - L_0$
- Usuario 3: $L_{0,3} - L_{0,7} - L_{0,5} - L_1 - L_0$
- Usuario 4: $L_{0,3} - L_{0,5} - L_1 - L_{0,7} - L_0$
- Usuario 5: $L_1 - L_{0,7} - L_{0,4} - L_{0,5} - L_0$

Podemos ver que los usuarios prefirieron distintos valores de α, pero todos tienen en común que pusieron la lista sin modificaciones en último lugar, mostrando que todos los usuarios prefirieron alguna lista modificada por el *framework*.

8. CONCLUSIONES

- La baja correlación encontrada entre *personal* y las otras métricas *beyond-accuracy* estudiadas, indica que la métrica propuesta es suficientemente diferente a las métricas usuales
- Se propone un *framework* que busca realizar recomendaciones que se adapten a las preferencias individuales de cada usuario, respecto a las distintas métricas *beyond-accuracy* estudiadas.
- El testeo del *framework* se realizó de manera satisfactoria, mostrando la viabilidad de el *framework* para que pueda ser implementado a futuro de manera más masiva.
- Sin ser concluyentes, los testeos realizados, indicaron la posibilidad de que exista una relación positiva entre *personal* y la satisfacción.
- Los resultados de las regresiones reafirmaron que pueden existir diferencias entre las preferencias de los usuarios respecto a distintas métricas *beyond-accuracy*
- En el testeo, los usuarios siempre catalogaron a la lista sin *rerank* como la lista de menor preferencia, indicando una recepción positiva con el *framework*.

9. TRABAJOS A FUTURO

- Principalmente, se requiere testear el *framework* propuesto con una cantidad significativa de usuarios, para así reafirmar las hipótesis y resultados mostrados en el testeo.
- Usar cuestionarios para medir si las listas entregadas por el *framework* propuesto con mayor valor de α efectivamente se sienten más personales y con mayor satisfacción
- Añadir medidas fuera de las *beyond-accuracy* a el proceso de la regresión múltiple, para analizar qué rol desempeña una métrica tradicional en comparación a las métricas *beyond-accuracy*

Referencias

[1] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2016.

[2] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[3] McNee S. Cai-Nicolas, Z. et al. Improving recommendation lists through topic diversification. 01 2005.

[4] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. How serendipity improves user satisfaction with recommendations? a large-scale user evaluation. In *The World Wide Web Conference*, pages 240–250, 2019.

[5] Gediminas Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911, 2011.

[6] Wang N. Yang Y. Yang K. Chen, L. and Q. Yuan. User validation of recommendation serendipity metrics. *arXiv preprint*, 2019.

[7] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

10. Apéndice

10.1. Estadísticas dataset

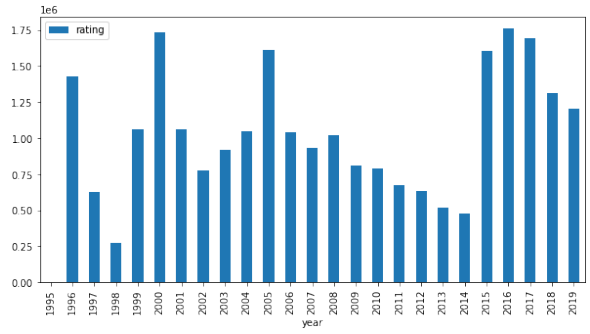

Figura 5: Distribución de años de películas

10.2. Resultados *framework*

Listas finales preferidas por cada usuario:

- Usuario 1:

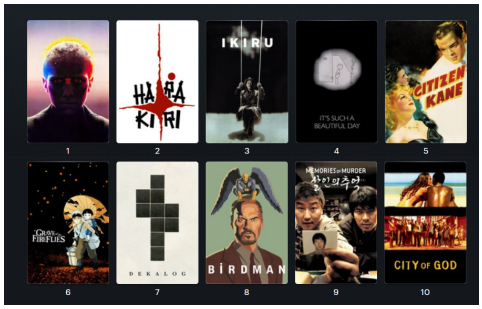

Figura 6: Lista de películas preferidas usuario 1

- Usuario 2:

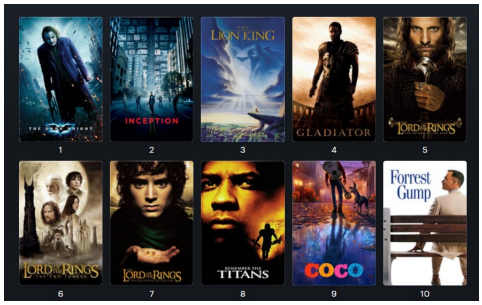

Figura 7: Lista de películas preferidas usuario 2

- Usuario 3:

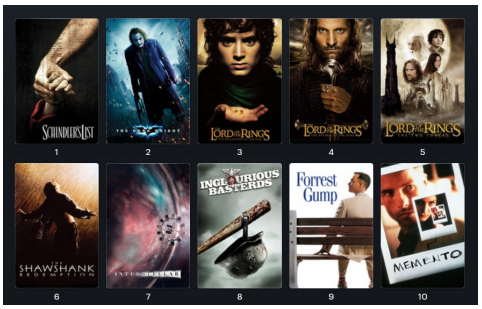

Figura 8: Lista de películas preferidas usuario 3

- Usuario 4:

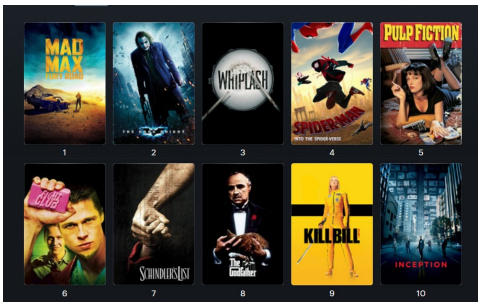

Figura 9: Lista de películas preferidas usuario 4

- Usuario 5:

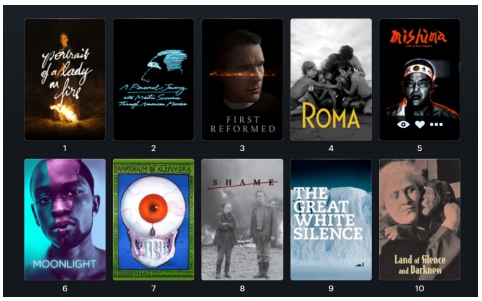

Figura 10: Lista de películas preferidas usuario 5