

Sistema de Recomendación de Artistas Musicales Mediante Minería de Reglas de Asociación

Ignacio Maltagliatti

Minería de Datos, U.T.N. Facultad Regional Paraná,
Av. Almafuerte 1033, (E3102SLK) Paraná, Entre Ríos
ignaciomaltagliatti@gmail.com

Resumen El objetivo del trabajo es crear un sistema de recomendación de artistas musicales basado en los hábitos de escucha de la comunidad de Last.fm empleando técnicas de minería de datos mediante el descubrimiento de reglas de asociación. Se eligió al método de filtrado colaborativo como eje del sistema propuesto. En primer lugar, se establecieron grupos de usuarios con gustos de música similares sobre el conjunto de datos estudiado a través de un proceso de iteración que correlacionó las etiquetas que describían a los artistas escuchados con las que colocó cada usuario. Luego, en los clústeres se buscaron reglas de asociación que permitieron descubrir nuevos artistas para recomendar entre los integrantes del grupo. Los resultados se compararon con el sistema recomendador de la plataforma Last.fm y se concluyó que fueron satisfactorios cuando se trabajó con usuarios frecuentes del servicio.

1. Introducción

Uno de los mayores problemas que afrontan las personas cuando navegan por Internet es la gran cantidad de información que encuentran. Los sistemas de recomendación ayudan a filtrar esa información y seleccionar las mejores opciones teniendo en cuenta las preferencias del usuario respecto a los temas de su interés (películas, música, libros, noticias, imágenes, páginas web, etc.).

En la actualidad existen diversos métodos y enfoques de filtrado de la información que componen un recomendador. Los sistemas más destacables se pueden clasificar en:

- *Sistemas con filtrado colaborativo:* detectan usuarios con intereses similares y luego crean recomendaciones sobre esa base.
- *Sistemas con filtrado de contenido:* recomiendan ítems que son similares a los ítems que previamente valoraron los usuarios.
- *Sistemas con filtrado demográfico:* recomiendan en función de atributos personales de los usuarios (edad, sexo, situación geográfica, profesión, etc.).
- *Sistemas con filtrado de conocimiento:* recomiendan en base al conocimiento de las necesidades de los usuarios y a las características de los productos.

Todos ellos se pueden combinar para generar sistemas híbridos e intentar cubrir algunas deficiencias que presentan los sistemas individualmente.

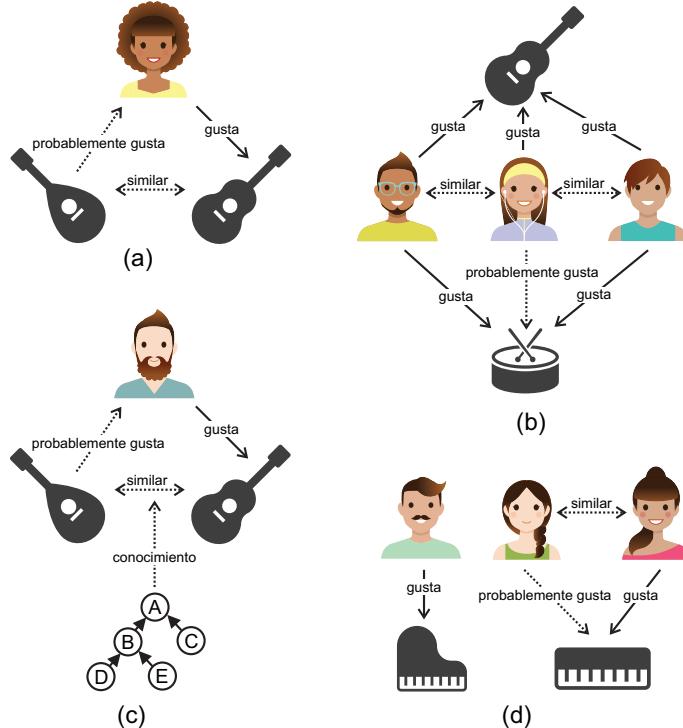


Figura 1. Enfoques de sistemas de recomendación: (a) filtrado de contenido, (b) filtrado colaborativo, (c) filtrado de conocimiento (d) filtrado demográfico.

El Filtrado colaborativo se basa en que si una persona tiene la misma opinión que otra sobre un tema es más probable que ambas tengan la misma opinión en otro tema diferente con respecto a la de una persona elegida azar. Esta técnica tiene la ventaja de ser agnóstica a la naturaleza de los elementos ya que simplemente aprovecha el supuesto de que habrá usuarios similares en el sistema. En la práctica conduce a buenos resultados, pero necesita superar varias dificultades. Uno de los problemas del filtrado por colaboración se conoce como “cold start” o arranque en frío que se manifiesta cuando se incorpora un usuario o elemento nuevo, los cuales carecen de información suficiente para ser comparados con otros usuarios o elementos. Por otro lado, son sistemas con una gran dispersión de la información ya que ésta se crea en un entorno no controlado; por ejemplo, cuando los usuarios de un servicio de música clasifican libremente el género de un artista (“rock”, “pop”, etc.), es decir, la clase musical de un artista escuchado varía de acuerdo a la apreciación del oyente.

Last.fm [1] es una plataforma social que integra una radio online y un sistema de recomendación de música basado en técnicas de filtrado colaborativo. Construye perfiles y patrones sobre los gustos musicales de los usuarios registrados mediante "scrobbler". Al reproducir canciones en sitios webs, aplicaciones, o reproductores de música que utilizan scrobbling el servicio agrega esa información a su base de datos (título de la canción, artista, etc.) para crear estadísticas de los hábitos de escucha de una persona registrada en la red que se comparan con las preferencias musicales de otros usuarios y si coinciden comparten recomendaciones de canciones, cantantes, etc. Dicho servicio, cuenta con una API pública que permite acceder a la información recolectada en su base de datos.

Este trabajo propone un sistema de recomendación colaborativo para sugerir artistas de música en base a datos provistos por la API de Last.fm. Para superar los inconvenientes de dispersión de datos se plantea agrupar a los usuarios estudiados por sus gustos musicales mediante un proceso iterativo que asocia la forma en que ellos etiquetaron a los artistas escuchados con el conjunto de etiquetas que recibió cada artista de los diferentes oyentes; luego, se plantea aplicar en cada clúster reglas de asociación que permitan descubrir relaciones entre los artistas escuchados por el grupo a partir de las cuales se realizan las recomendaciones. El sistema se limita a ser aplicado en usuarios frecuentes del sitio web, es decir, que hayan colocado una cantidad suficiente de etiquetas para ser incluidos en un clúster.

2. Metodología

Se utilizaron conjuntos de datos públicos recopilados a través de la API de Last.fm que contienen algunos de los hábitos de escucha de 1892 usuarios. Los mismos se dividen en cuatro archivos (tabla 1) publicados en formato *.dat* en la 5^a Conferencia ACM sobre Sistemas de Recomendación (RecSys 2011) [2] en el marco del segundo taller internacional sobre Heterogeneidad de la Información y Fusión en los Sistemas de Recomendación (HetRec 2011).

Tabla 1. Archivos de datos.

Archivos	Datos	Cantidad de datos
artists.dat	artistas musicales escuchados y etiquetados por usuarios	17632
tags.dat	etiquetas disponibles en el dataset	11946
user_artists.dat	artistas escuchados por cada usuario y conteo de escucha para cada par usuario-artista	92834
user_taggedartists.dat	asignaciones de etiquetas de artistas proporcionadas por cada usuario y marcas de tiempo cuando se realizaron	186479

En la tabla 2 se resumen las características de los atributos dispuestos en cada archivo.

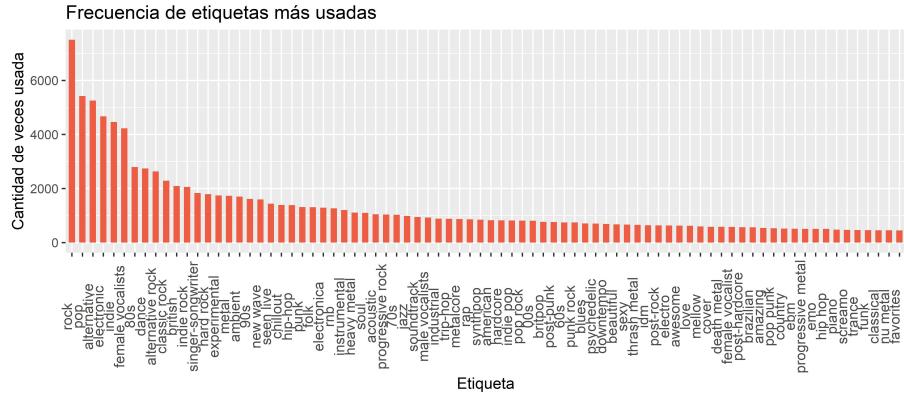
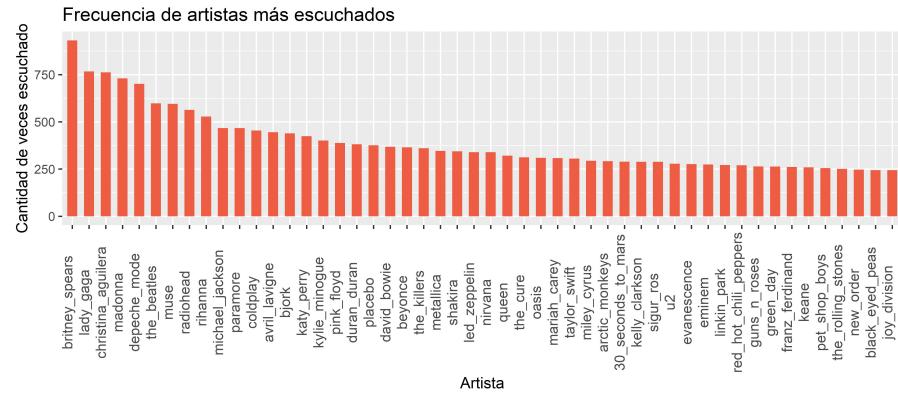
Tabla 2. Características de los datasets.

Archivo	Atributo	Escala	Valores nulos
artists	Id	Nominal	No
	name	Nominal	No
	url	Nominal	No
	pictureURL	Nominal	No
tags	tagID	Nominal	No
	tagValue	Nominal	No
user_artists	userID	Nominal	No
	artistID	Nominal	No
	weight	Razón	No
user_taggedartists	userID	Nominal	No
	artistID	Nominal	No
	tagID	Nominal	No
	day	Intervalo	No
	month	Intervalo	No
	year	Intervalo	No

Los conjuntos de datos no contenían valores nulos, los principales problemas que se encontraron son atributos nominales con palabras o frases que referían a un mismo artista o etiqueta, pero fueron ingresados con alguna diferencia de escritura. Además, se observaron algunos datos inconsistentes como por ejemplo la etiqueta “happygoodfunandhandclaps”.

La figura 2 muestra las 80 etiquetas más usadas para describir a los artistas musicales en el conjunto de datos original. Se destaca la importancia de la etiqueta “rock” (7503 veces usada) por encima de las demás, incluso se aprecian otras etiquetas que hacen referencia a subgéneros del estilo rock. En menor cantidad, 4200 a 5500 etiquetas, se distingue un segundo grupo de palabras usadas (“pop”, “alternative”, “electronic”, “indie” y “female vocalists”) y el resto va descendiendo gradualmente desde una frecuencia aproximada a 2500 etiquetas. Por otro lado, los 50 artistas más escuchados se representan en la figura 3 donde sobresale la cantante Britney Spears con 931 escuchas.

Se combinaron los datasets en una única tabla y se determinó que el conjunto evaluado de 1892 usuarios colocó 9749 etiquetas diferentes de las 11946 existentes

**Figura 2.** Etiquetas más usadas.**Figura 3.** Artistas más escuchados.

en el archivo *tags*, mientras que el número de artistas escuchados por éstos asciende a 12523, sobre un total de 17632 artistas encontrados en el archivo *artists*.

2.1. Limpieza de datos

Para mejorar la calidad de los datasets *artists* y *tags* se utilizó la herramienta OpenRefine [3] con la cual se transformó todo el texto a minúsculas, se eliminaron espacios en blancos al principio y final de las palabras y se reemplazaron espacios entre palabras por guiones bajos. También se unificaron palabras similares mediante técnicas de agrupación como colisión de llaves o vecinos más cercanos (figura 4).

Posteriormente, se trabajó en Excel exclusivamente con el conjunto *tags* para tratar de aumentar en forma “manual” el agrupamiento de etiquetas. Con la

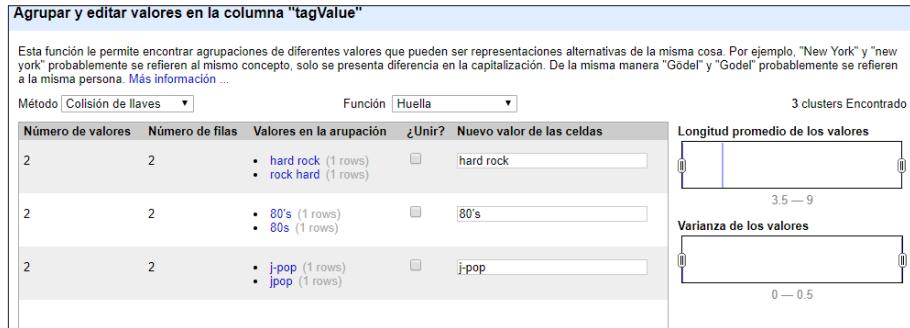


Figura 4. Detección de palabras similares en la herramienta OpenRefine.

herramienta de filtro se buscaron palabras similares (no detectadas en OpenRefine) y se evaluó si se relacionaban; por ejemplo, la etiqueta “1980” se unificó con la etiqueta “80s” (ambas refieren a música de los años 80s). Asimismo, se estudiaron las etiquetas que aludían explícitamente a un género musical; cuando se creyó conveniente se agruparon géneros con subgéneros de música con el fin de crear un grupo mayor y tratar de reducir la dispersión de etiquetas asignadas. Puede mencionarse el caso de las etiquetas “black metal” (subgénero musical) que se renombraron y agregaron al grupo “death metal”, teniendo en cuenta la clasificación de géneros musicales provista en sitios webs como Musicmap¹, Wikipedia², etc. (figura 5).

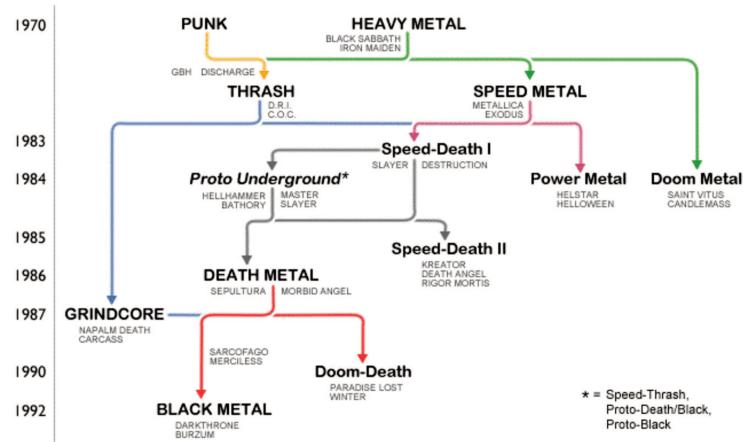


Figura 5. Subgéneros de música Metal.

¹ <https://musicmap.info>

² <https://wikipedia.org>

Como consecuencia de la limpieza se redujeron en el dataset *tags* la cantidad de etiquetas diferentes a 8171 (de 11946) de las cuales 6637 representan el etiquetaje de los 1892 usuarios que se usó para describir a los artistas. Mientras que el dataset *artists* disminuyó su cantidad a 16814 (de 17632) artistas de los cuales 11714 fueron escuchados por los usuarios en estudio. En la figura 6 se compara en forma visual el conjunto de etiquetas originales en el archivo *tags* con el obtenido luego del tratamiento de los datos.

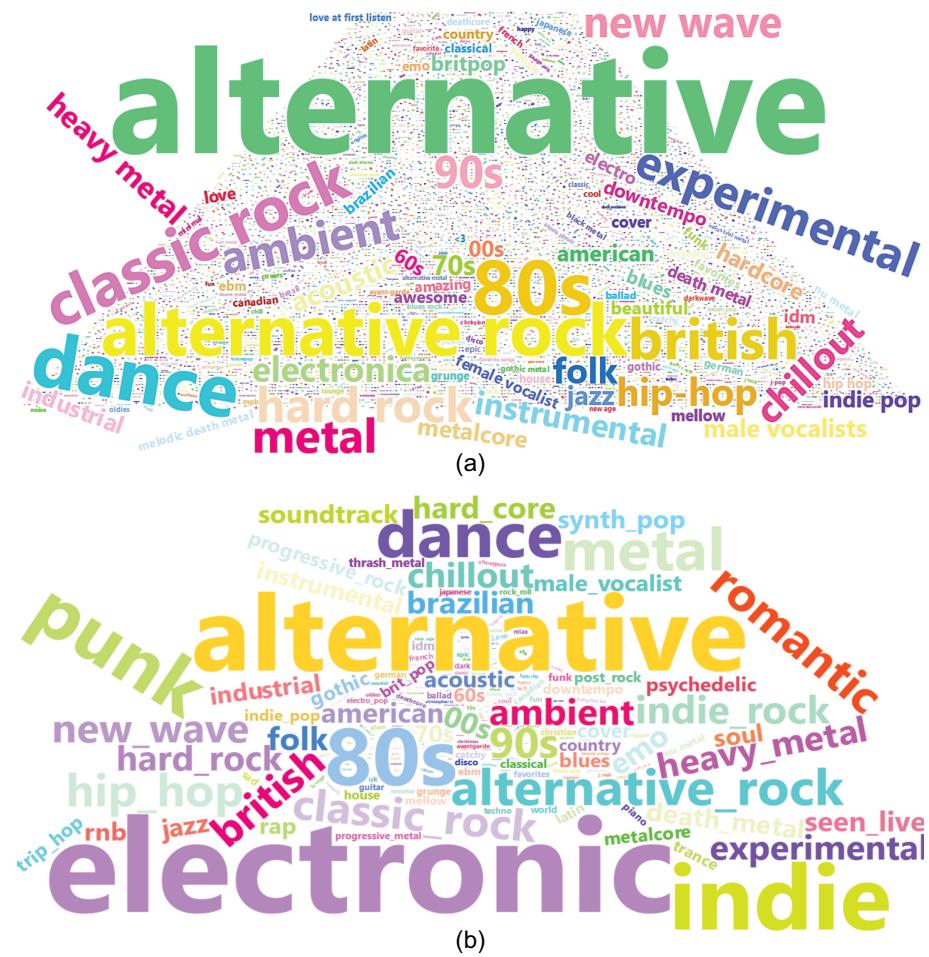


Figura 6. Nube de palabras de etiquetas más usadas: (a) dataset *tags* original (b) dataset *tags* tratado.

Se observa en la figura 7 que 5190 etiquetas distintas fueron asignadas por un grupo de usuarios 1 a 4 veces y 1207 etiquetas 5 a 49 veces. Esto quiere decir que la mayor cantidad de asignaciones (mayor o igual a 50) se centra en 240

etiquetas, sobre el total de 6637. Visto de otro modo, 240 etiquetas representan el 86,34 % de las asignaciones.

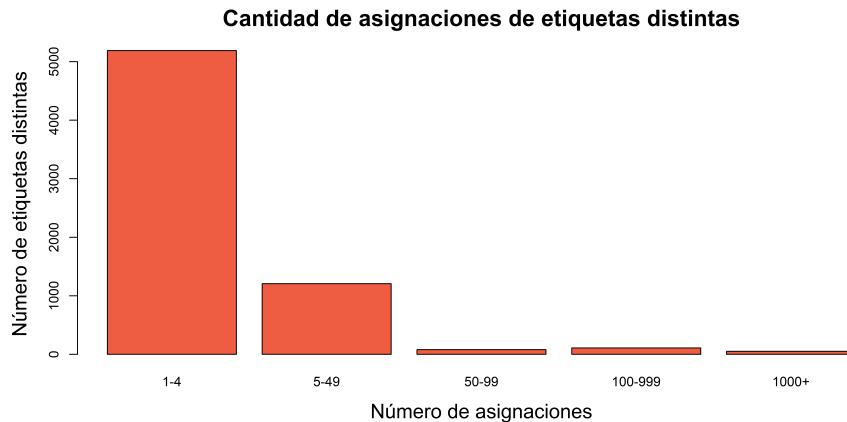


Figura 7. Frecuencia de uso de etiquetas distintas.

En base a lo analizado se adoptó como criterio encarar el trabajo con las 240 etiquetas más utilizadas, o sea, las que se asignaron más de 50 veces quedando de esta forma 161009 etiquetas asignadas por 1822 usuarios. Ello implica que 70 usuarios han etiquetado a artistas usando únicamente palabras poco frecuentes y se considera que no aportan suficiente información para incluirlos en el sistema de recomendación de filtrado colaborativo propuesto. Los 1822 usuarios escucharon en total 11213 artistas distintos. En la tabla 3 se resume lo mencionado.

Tabla 3. Resultados de la limpieza de datos.

Datos	Conjunto original (ud)	Conjunto limpio (ud)	Conjunto de trabajo del original (ud)	Porcentaje (%)
Usuarios	1892	1892	1822	96,30
Artistas diferentes disponibles	17632	16814	-	-
Artistas diferentes escuchados	12523	11714	11213	89,54
Etiquetas diferentes disponibles	11946	8171	-	-
Etiquetas diferentes usadas	9749	6637	240	2,46
Asignaciones de etiquetas	186479	186479	161009	86,34

2.2. Sistema de recomendación propuesto

La muestra de usuarios se dividió en clústeres para acotar la dispersión de los datos en cuanto al etiquetaje realizado y mejorar el rendimiento de las recomendaciones. Como hipótesis se consideró que los usuarios que usaron etiquetas similares comparten semejantes gustos musicales o tipos de artistas escuchados. Por ejemplo, en la tabla 4 se piensa que los usuarios 1 y 2 tienen gustos parecidos, mientras que el 3 prefiere otros estilos de música y, por lo tanto, desea escuchar otros artistas.

Tabla 4. Ejemplo de etiquetas colocadas por usuarios.

Usuario 1	{pop, rock, pop_rock, alternative, 80s, amazing}
Usuario 2	{pop, rock, classical_rock, amazing, indie_rock}
Usuario 3	{electro, ebm, electronic, hip_hop}

No obstante, se debe tener en cuenta que en la plataforma Last.fm los artistas son etiquetados de diversas formas por los diferentes usuarios. Así, se encontraron usuarios que escucharon los mismos cantantes o grupos de música, pero los clasificaron distinto como se aprecia en la figura 8. En consecuencia, se decidió agrupar a los usuarios a partir de sus etiquetas, pero tomando en consideración también la relación que éstas tienen con los artistas, o sea, comprobando todas las palabras o frases que pueden referirse al mismo músico o banda. Esto se llevó adelante con ayuda de reglas de asociación.

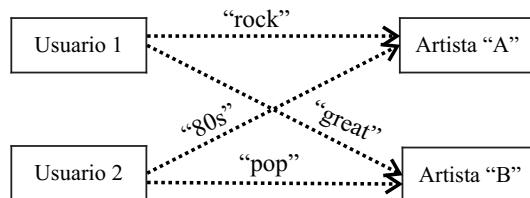


Figura 8. Etiquetado diferente entre usuarios.

Se generaron en el software R Studio con la librería Arules [4] y el conjunto de datos de trabajo transacciones de etiquetas por artista. Con el algoritmo Apriori se buscaron reglas de asociación con valores de soporte 0,01 y de confianza 0,70 bajo el criterio de generar suficientes asociaciones para formar posteriormente una adecuada cantidad de clústeres de usuarios sin afectar los recursos computacionales disponibles. Además, se limitó a que las reglas resultantes tengan un solo ítem como consecuente a los fines de simplificar el proceso de agrupamiento. Se consiguieron 1813 reglas con un valor de lift mínimo de 2,94, lo cual asegura que hay una relación positiva entre todas las asociaciones. Con la función

“is.redundant” de Arules se eliminaron reglas específicas redundantes contenidas en otras más generales de mayor o igual valor de confianza, y quedaron finalmente 1659 reglas. Las relaciones encontradas se pueden interpretar como, por ejemplo, si un artista fue clasificado por un usuario con el texto “electro_pop” pudo ser etiquetado por otros usuarios como “electro”, o tal vez “pop”.

En base a lo anterior se llevó adelante, también en R Studio, un proceso iterativo donde se comparó cada etiqueta definida por cada usuario con las reglas de asociación obtenidas. Concretamente, para cada usuario se buscaron los antecedentes de las reglas que coincidían con sus etiquetas y se armó una tabla con todos los consecuentes correspondientes. Se adoptó que el consecuente más repetido representa el grupo del usuario. En la figura 9 se muestra parte del proceso de iteración donde la mayor cantidad de consecuentes refieren a la palabra “pop”, entonces para ese caso se considera que el Usuario 1 se encuentra en el grupo de usuarios cuya mayoría de consecuentes obtenidos también refieren a la misma palabra (“pop”).

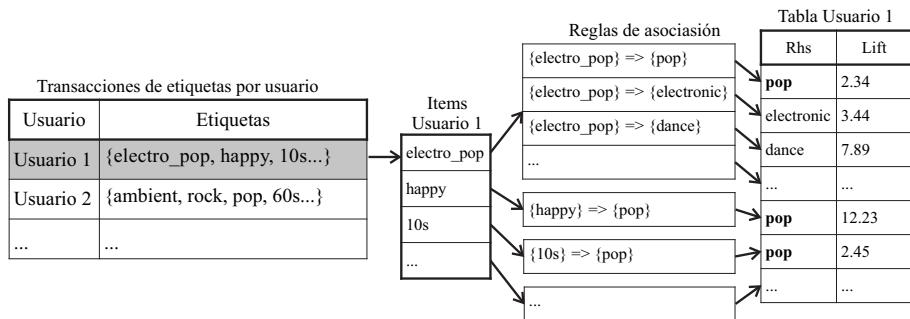


Figura 9. Proceso de iteración para crear clústeres de usuarios.

Cuando hubo un “empate” en la cantidad de palabras o frases consecuentes, se sumaron los valores de lift de cada una y el mayor alcanzado definió la clase del usuario.

Se constituyeron 16 grupos con un total de 1409 usuarios. El hecho de que hayan quedado relegados en el camino usuarios (413) se debe a que no presentaron suficiente información para incluirlos en un clúster aplicando esta metodología. Se menciona el caso del usuario “userID 5” que sólo usó como etiqueta la palabra “summer” y no se pudo ligar a ninguna de las reglas calculadas. La tabla 5 expone los resultados conseguidos en donde se observa que algunos grupos cuentan con pocos usuarios lo cual es razonable teniendo en cuenta el tamaño del dataset de prueba con el que se trabajó.

Tabla 5. Clústeres de usuarios.

N	Clúster	Cantidad usuarios
1	rock	659
2	alternative	197
3	pop	191
4	pop_rock	97
5	electronic	83
6	romantic	58
7	hip_hop	29
8	metal	28
9	80s	19
10	indie	14
11	chillout	11
12	british	7
13	classic_rock	7
14	industrial	5
15	dance	3
16	ambient	1

Por último, en cada clúster se armaron transacciones de los artistas escuchados por los integrantes y con el algoritmo Apriori, en R Studio, se buscaron reglas de asociación para descubrir recomendaciones de músicos entre las personas del conjunto. Se hicieron varias pruebas con valores de soporte relativamente chicos, entre 0,05 y 0,20, para tratar de incluir la mayor cantidad de artistas posibles y valores de confianza en un rango de 0,70 a 0,90. Los resultados se contrastaron con las métricas lift e hyperlift [5] incluidas en la librería Arules.

3. Resultados y discusión

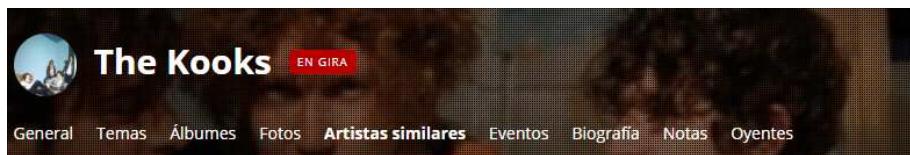
A modo de ejemplo se evaluaron los artistas escuchados por la clase “alternative”. Con un soporte de 0,10 y confianza de 0,70 se obtuvieron 10 asociaciones entre artistas; no se halló ninguna regla redundante. En la tabla 6 se visualizan las reglas descubiertas que cuentan con un solo antecedente junto con los valores de las métricas soporte, confianza, lift e hyperlift.

Last.fm posee una sección que sugiere artistas similares a un músico o banda determinados. Se asume que la posición de las recomendaciones tienen un orden de importancia de acuerdo a los criterios del sistema. Escogiendo al grupo The

Tabla 6. Reglas de asociación en clúster “alternative”.

N	Lhs	Rhs	Sop.	Conf.	Lift	Hlift	Posición Last.fm
1	{rihanna}	{lady_gaga}	0,102	0,909	4,165	2,222	6
2	{the_kooks}	{arctic_monkeys}	0,102	0,741	2,861	1,667	1
3	{katy_perry}	{lady_gaga}	0,107	0,724	3,318	1,909	6
4	{britney_spears}	{lady_gaga}	0,122	0,828	3,792	2,182	19
5	{franz_ferdinand}	{arctic_monkeys}	0,107	0,724	2,797	1,615	4
6	{oasis}	{arctic_monkeys}	0,107	0,700	2,704	1,615	20
7	{the_strokes}	{arctic_monkeys}	0,142	0,757	2,923	1,877	3

Kooks, figura 10, se puede comprobar que la banda Arctic Monkeys aparece primera entre las recomendaciones; en la tabla 6 se coloca en la última columna la posición que ocupan los consecuentes (Rhs) respecto a los antecedentes (Lhs) en la sección mencionada del sitio online.



Artistas similares



Arctic Monkeys

Arctic Monkeys es una banda inglesa de rock alternativo e indie



Miles Kane

Miles Paul Kane, nacido el 17 de marzo de 1986, es un músico inglés.



The Fratellis

The Fratellis son una banda de rock alternativo originaria de Glasgow,

Figura 10. Sección de artistas similares en el servicio Last.fm.

Las recomendaciones obtenidas usando el método propuesto, que se desprenden de las reglas de asociación de los artistas de un clúster (como Lhs), coinci-

dieron con alguna de las 20 primeras sugerencias de Last.fm. Se han realizado otras pruebas con otras clases y en general los resultados fueron coherentes, aunque hubo asociaciones calculadas que no aparecieron como recomendaciones en la plataforma web; en esos casos no se puede asegurar que las reglas condujeron a resultados negativos (malas recomendaciones) ya que no se conoce la precisión del servicio online y, por otro lado, se observó que los artistas consecuentes derivados de las asociaciones se relacionaban con el estilo de música de sus antecedentes. Dicho de otra forma, el sistema propuesto pudo haber descubierto algún artista para recomendar consistente con los hábitos de escucha de un usuario que el sistema de recomendación de Last.fm no encontró.

En muchas pruebas se notó que si se consideraban solo las reglas de asociación halladas con valores de hyperlift mayor a 1,5 crecía el porcentaje de coincidencias de las recomendaciones propuestas con las de la plataforma.

4. Conclusiones

El sistema desarrollado demostró ser un método válido y sencillo para usar como recomendador cuando se aplica sobre usuarios frecuentes de la comunidad Last.fm bajo un nivel determinado de etiquetas asignadas, ya que en general las recomendaciones fueron coherentes con las que provee el sistema de dicha comunidad. Hubo casos específicos donde la propuesta arrojó asociaciones que no se encontraron entre los artistas recomendados por la plataforma, pero no se puede concluir que son resultados inexactos ya que no se conoce el nivel de precisión de sus recomendaciones.

Los datasets *tags* y *artists* todavía pueden ser mejorados para aumentar la calidad y la cantidad de sus datos, aunque inevitablemente al tratarse Last.fm de una red colaborativa el ingreso de los datos de etiquetas y nombres de artistas no está controlado, entonces se produce información dispersa y duplicada.

La obtención de reglas de asociación usando el algoritmo Apriori de la librería Arules en el software R Studio, tanto para realizar la clasificación de usuarios como para encontrar recomendaciones de artistas, no demandó excesiva cantidad de recursos computacionales, sin embargo, el proceso de iteración para crear los clústeres de usuarios fue más exigente en este sentido en la medida que se aumentó la cantidad de reglas.

En el trabajo se distinguieron cuatro tipos de usuarios; los nuevos que no tienen registros de escuchas de música en el sitio web (no se incluyeron en éste desarrollo), usuarios que etiquetaron sólo con palabras o frases “poco frecuentes” de acuerdo a un umbral definido, aquellos que usaron al menos una de las etiquetas más frecuentes para describir a un artista pero en una cantidad insuficiente para asociarlos a otros usuarios y, por último, los usuarios que etiquetaron en un nivel suficiente para poder dividirlos en clústeres. Estos últimos son los únicos aptos para ser probados en el sistema de recomendación propuesto. En cambio, los tres primeros casos mencionados representan un problema de “cold start” y si se desea agrupar los usuarios para reducir la dispersión de datos deben

utilizarse otras técnicas de filtrado, por ejemplo, empleando las características demográficas (edad, sexo, país, etc.) provistas al registrarse en el servicio.

Por último, se debe tener en cuenta que el agrupamiento de usuarios similares conlleva al descubrimiento de artistas de estilos musicales parecidos, entonces podría ser deseable complementar las recomendaciones con artistas de otros géneros aplicando reglas de asociación directamente sobre el conjunto global.

Referencias

1. Servicio de Música Online Last.fm, <https://www.last.fm>
2. 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011) held in conjunction ith the 5th ACM Conference on Recommender Systems (RecSys 2011) . Chicago, IL, USA (2011).
<http://ir.ii.uam.es/hetrec2011/datasets.html>
3. Herramienta OpenRefine, <http://openrefine.org>
4. Hahsler, M., Buchta, C., Gruen, B. and Hornik, K. Arules: Mining Association Rules and Frequent Itemsets. R package version 1.6-1 (2018).
<https://CRAN.R-project.org/package=arules>
5. Hahsler, H., Hornik, K.: New Probabilistic Interest Measures for Association Rules. Vienna University of Economics and Business Administration, Augasse 2–6, A-1090. Vienna, Austria (2007)