

Algoritmo de Filtrado Híbrido-Mixto para Recomendación de Contenidos Audiovisuales a Comunidades Virtuales

Hybrid-Mixed Filtering Algorithm for Audiovisual Content Recommendation to Virtual Communities

Diego Fabián Duran Dorado, José Luis Arciniega Herrera
Ingeniería de sistemas, universidad del Cauca, Popayán, Colombia.
 dduran@unicauca.edu.co
 jlarci@unicauca.edu.co

Resumen— Generalmente, los algoritmos de filtrado son usados en las recomendaciones de contenidos sin tener en cuenta el contexto de aplicación. Como adaptación a un contexto enmarcado por el consumo de contenidos audiovisuales por miembros de comunidades virtuales, se propone un algoritmo de filtrado híbrido-mixto de recomendaciones, que combina los enfoques colaborativo y basado en contenido. La experimentación permite concluir que el algoritmo es preciso con un pequeño vecindario de usuarios, además que reduce los efectos del *Cold-Start*. Lo anterior se consigue porque a diferencia de otros algoritmos, éste trabaja en función de los niveles de interés por las comunidades.

Palabras clave— Algoritmo, Comunidades virtuales, Filtrado híbrido-mixto, Sistema de recomendaciones

Abstract— Regularly, filtering algorithms are used in content recommendation regardless of the context of application. For a context within the bounds of consumption of audiovisual content by the members of virtual communities, a filtering algorithm for hybrid-mixed content recommendation is proposed, combining the collaborative and item-based approaches to context. Experimental results support the conclusion that the propose algorithm is accurate from a small neighborhood of users and get recommendations even with little user information, reducing the effects of Cold-Star problem. This is attained as in this case the process is a function of the degrees of interest in the communities.

Key Word — Algorithm, Hybrid-mixed filtering, Recommender system, Virtual communities.

I. INTRODUCCIÓN

Con el desarrollo de la tecnología IP, se ha popularizado entre las personas el uso de Internet y la Televisión Digital para la generación de conocimiento colaborativo en torno

contenidos audiovisuales a través de la dinámica de comunidades virtuales con intereses similares [1]. Casos específicos son: SocialTV, e-learning y las CAV (Comunidades Académicas Virtuales). Con el tiempo la cantidad de contenidos disponibles ha aumentado, obligando a los usuarios a explorar espacios densos en información, en los que la selección de un contenido de interés es una tarea tediosa y difícil de asumir sin una herramienta de apoyo [2]. Por esta razón, se ha popularizado el uso de los sistemas de recomendaciones [3]. Estos se dividen en dos grandes enfoques: basados en contenido y colaborativos. Los sistemas basados en contenido generan recomendaciones por medio de algoritmos de filtrado, a partir de los contenidos que interesaron en el pasado al usuario – en adelante llamado “usuario activo” – mientras que los colaborativos adoptan algoritmos con técnicas estadísticas para hallar usuarios con preferencias comunes al activo – en adelante “vecinos” – para generar recomendaciones [4]. El uso único de filtrado basado en contenido genera recomendaciones cada vez más similares (problema de sobre-especialización), mientras que el uso del colaborativo recomienda contenidos cada vez más acordes a los gustos de los vecinos, perdiendo de vista las preferencias individuales del usuario activo [5].

Por otro lado, generalmente los algoritmos de filtrado propuestos en la literatura trabajan en función de los niveles de interés (Doi) del usuario activo por los contenidos y sus atributos [6], [1], [7]. Si un contenido nuevo es registrado en el sistema o un usuario nuevo no registra un número suficiente de Doi, podría darse un alto nivel de dispersión de datos, minando la calidad de las recomendaciones y retrasando su generación (problema Cold-Start).

En este artículo se propone un algoritmo de filtrado híbrido-mixto para la recomendación de contenidos audiovisuales a comunidades virtuales en entornos de Internet y Televisión Digital, que combina recomendaciones basadas en contenido y colaborativas de forma complementaria. Además, se adapta a la dinámica de las comunidades, trabajando en función del nivel de interés del usuario activo por ellas para reducir la dispersión de

datos y así, obtener la mejor calidad posible en las recomendaciones mientras se reduce los efectos negativos del Cold-Start. El estudio es validado a través de pruebas experimentales de Error Medio Absoluto y relacionadas con el problema Cold-Start.

Este trabajo es resultado de las investigaciones llevadas a cabo en el proyecto financiado por COLCIENCIAS de nombre ST-CAV (Colombia) [8], [9], [10], cuyo objetivo es el de generar espacios para el fomento de los procesos educativos de las Comunidades Académicas Virtuales en Televisión Digital y al interior del Laboratorio de Televisión Digital de la Universidad del Vigo (España).

El artículo está compuesto de la siguiente manera: en la sección II se presenta y explica un conjunto de conceptos relacionados con la investigación; en la sección III se presentan las consideraciones de diseño del algoritmo, el modelado de los contenidos y los Doi de los usuarios y la descripción del algoritmo; en la sección IV se presentan los resultados y análisis de pruebas de precisión del algoritmo y relacionadas con el Cold-Start; finalmente se presentan las conclusiones del trabajo de investigación.

II. CONCEPTOS BÁSICOS RELACIONADOS CON LA INVESTIGACIÓN

Comunidad Virtual. Grupo o conjunto de individuos que establecen relaciones a partir de temas comunes, donde los vínculos, interacciones y relaciones tienen lugar en un espacio virtual [11].

Problema Cold start. Limitación que se manifiesta hasta no alcanzar un número elevado de usuarios registrados o de información relacionada con sus preferencias, por lo que no se dispone de información suficiente para extraer recomendaciones de calidad [2].

Sobre-especialización. Al sugerir contenidos similares a los que le interesaron al usuario en el pasado, con el tiempo las recomendaciones son cada vez más parecidas, sin dar cabida a la variedad [4].

Dispersión de datos. En el enfoque colaborativo, se manifiesta cuando el aumento de contenidos y la diversidad en temática reduce la probabilidad de que los usuarios compartan los mismos Doi, dificultando el proceso de hallar vecinos. En el enfoque basado en contenido, se manifiesta cuando los contenidos no comparten atributos o características semánticas, dificultando el cálculo de similitudes entre ellos [4].

III. DISEÑO DEL ALGORITMO DE FILTRADO HÍBRIDO

A. Consideraciones de Diseño del Algoritmo de filtrado

Con el fin de obtener la mayor precisión de las recomendaciones, definen las siguientes consideraciones:

- 1) **La dinámica de los usuarios.** En una comunidad virtual existe una gran probabilidad de que entre sus miembros haya gustos similares. Esto puede ser de utilidad en: 1) la producción de información de entrada para el algoritmo y paliar los efectos del problema *Cold-Start*; 2) el cálculo de vecinos con perfiles similares.
- 2) **El entorno de aplicación.** Por lo general, un usuario de televisión e Internet no está dispuesto a someterse a cuestionarios relacionados con sus preferencias. Por lo que el algoritmo debe incluir una forma de hacer eficiente el uso de la poca información que el usuario especifique y registre en el sistema.
- 3) **Los objetivos del sistema.** Para lograr un equilibrio en las recomendaciones, se propone un algoritmo con enfoque híbrido mixto que aproveche la naturaleza complementaria y las ventajas ofrecidas por cada uno de los enfoques basado en contenido y colaborativo.
- 4) **Modelado de la información de los contenidos.** En el enfoque basado en contenido, generalmente se modelan los contenidos a través de matrices que relacionan su título con la ausencia o presencia de atributos. Por lo anterior, puede presentarse dos problemas: 1) los contenidos no comparten un número considerable de palabras clave y exista dispersión de datos en las matrices; 2) las matrices sean demasiado grandes y haya la necesidad de recurrir a técnicas de factorización, aumentando la carga computacional del sistema. En el ámbito de las comunidades virtuales, existen dos hechos importantes que pueden influir en el modelado de la información de los contenidos y reducir la carga computacional del sistema: 1) existe una clasificación de los intereses a la cual asociar los contenidos, obteniendo por lo menos un punto de asociación entre ellos; 2) cada comunidad virtual posee un coordinador experto en las temáticas tratadas, que tiene el criterio de incluir los contenidos subidos al sistema en diferentes comunidades según su pertinencia. A partir de lo anterior, se concluye que es posible modelar los contenidos en función de la temática y clasificación de intereses para las comunidades virtuales. De esta manera, se pasa de modelar los contenidos a partir de una menor cantidad de comunidades, minimizando el tamaño de las matrices y la dispersión de datos en ellas.

B. Condiciones iniciales para el funcionamiento del algoritmo: Modelado de Contenidos y los Niveles de Interés de los Usuarios

Los algoritmos de filtrado requieren información de entrada para generar recomendaciones. Los dos tipos se explican a continuación: 1) información relacionada con el modelado de la información de los contenidos; 2) los Doi (niveles de interés) de los usuarios por los contenidos del sistema.

- 1) **El modelado de la información de los contenidos.** Es posible modelar la información de los contenidos en función de la temática y clasificación de interés en torno a los que giran las comunidades virtuales. Para esto, el coordinador experto debe asociar los contenidos a las comunidades para las cuales podría resultar de gran interés. La asociación

puede ser formalizada utilizando formatos XML, tablas relacionales o taxonomías. A partir de lo anterior, se modela la información de los contenidos en una matriz denominada MCC (Matriz Contenido-Clasificación) (Ver Fig. 1). En ella, el valor de las celdas obedece a (1) [2].

$$A(C_j, A_m) = \begin{cases} 1, & \text{si } C_j \text{ está asociado a la Comunidad } A_m \\ 0, & \text{si } C_j \text{ no está asociado a la Comunidad } A_m \end{cases} \quad (1)$$

		Comunidades			
		A ₁	A ₂	.. A _m	.. A _M
Contenidos	C ₁	1	1	A(C ₁ , A _m)	A(C ₁ , A _M)
	C ₂	0	1	A(C ₂ , A _m)	A(C ₂ , A _M)
	⋮				
	C _j	A(C _j , A ₁)	A(C _j , A ₂)	1	A(C _j , A _M)

Figura 1. Matriz MCC (Matriz Contenido-Clasificación) que modela la información de los contenidos

Debido a que un contenido está asociado a diversas comunidades, se disminuye la dispersión de datos en la matriz MCC, facilitando la aplicación de diferentes métricas del ámbito de las recomendaciones.

- 2) **Los Doi de los usuarios por los contenidos.** Una manera de formalizar el interés despertado en un usuario por un contenido es a través de un valor numérico llamado Doi (nivel de interés). El Doi varía entre valores por debajo de un umbral indicando un interés negativo y por encima indicando un interés positivo. Estos valores son coleccionados por el sistema de forma explícita o implícita. Los Doi son registrados en una matriz denominada MDoi (Matriz de Niveles de Interés). En la Fig. 2 se muestra un ejemplo de matriz MDoi en el que el Doi del usuario U₁ por el contenido C₁ es 5 (cinco). En general DOI_{U_x}(C_j) es el Doi del usuario U_x por el contenido C_j.

	C ₁	C ₂	.. C _j
U ₁	5	1	DOI _{U₁} (C _j)
U ₂	2	5	DOI _{U₂} (C _j)
U ₃	1	3	DOI _{U₃} (C _j)
⋮			
U _x	DOI _{U_x} (C ₁)	DOI _{U_x} (C ₂)	DOI _{U_x} (C _j)

Figura 2. Estructura de la matriz MDoi-Matriz de Niveles de Interés

C. Algoritmo de filtrado Híbrido-Mixto propuesto

En la Fig. 3 se presenta el algoritmo de filtrado híbrido-mixto propuesto. Se observan tres procesos importantes determinados por el algoritmo: 1) El modelado de los perfiles de Usuario; 2) Filtrado basado en Contenido; 3)

Filtrado Colaborativo. La lista de recomendaciones final es una mezcla de las recomendaciones obtenidas en los dos últimos procesos.

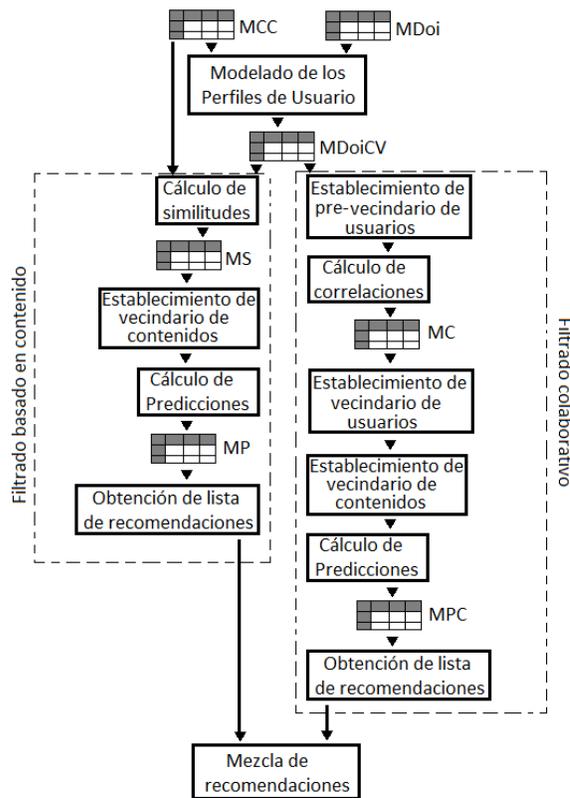


Figura 3. Algoritmo de Filtrado Híbrido-mixto

- 1) **Modelado de los perfiles de Usuario.** La información de los usuarios relacionada con sus preferencias se modela en una matriz denominada MDoiCV (Matriz de Niveles de Interés por las Comunidades Virtuales), que registra los niveles de interés por las Comunidades Virtuales. Los datos de la matriz MDoiCV se obtienen a partir de los registrados en las matrices MDoi y MCC. Para un total de T contenidos asociados a la comunidad académica A_m, el nivel de interés del usuario U_x por la comunidad A_m está dado por (2) [2].

$$DOI_{U_x}(A_m) = \frac{1}{T} \sum_{j=1}^T DOI_{xj} \quad (2)$$

Estos datos se registran en la matriz MDoiCV, tal como se observa en la Fig. 4.

	A ₁	A ₂	.. A _m	.. A _M
U ₁	4.5	1	DOI _{U₁} (A _m)	DOI _{U₁} (A _M)
U ₂	4	5	DOI _{U₂} (A _m)	DOI _{U₂} (A _M)
U ₃	1	3.3	DOI _{U₃} (A _m)	DOI _{U₃} (A _M)
⋮				
U _x	DOI _{U_x} (A ₁)	DOI _{U_x} (A ₂)	DOI _{U_x} (A _m)	DOI _{U_x} (A _M)

Fig. 4. Estructura de la matriz MDoiCV-Matriz de Niveles de Interés por las Comunidades Virtuales

2) **Filtrado basado en contenido.** Los pasos definidos en el algoritmo para este proceso son:

- Cálculo de similitudes. Se calcula la similitud basada en coseno entre todos los contenidos del sistema utilizando los datos registrados en la matriz MCC a través de (3) [3].

$$sim_{cos}(C_j, C_k) = \frac{\vec{C}_j \cdot \vec{C}_k}{|\vec{C}_j| \cdot |\vec{C}_k|} \quad (3)$$

Los vectores \vec{C}_j y \vec{C}_k tienen como componentes el conjunto de valores 1 ó 0 según la (1). Los valores de similitud se registran en la matriz MS (Matriz de Similitudes) presentada en la Fig. 5.

		Contenidos			
		C ₁	C ₂	..C _k	C _K
Contenidos	C ₁	1	$sim_{cos}(C_1, C_2)$	$sim_{cos}(C_1, C_k)$	$sim_{cos}(C_1, C_K)$
	C ₂	$sim_{cos}(C_2, C_1)$	1	$sim_{cos}(C_2, C_k)$	$sim_{cos}(C_2, C_K)$
	⋮	$sim_{cos}(C_j, C_1)$	$sim_{cos}(C_j, C_2)$	1	$sim_{cos}(C_j, C_K)$
	C _j	$sim_{cos}(C_N, C_1)$	$sim_{cos}(C_N, C_2)$	$sim_{cos}(C_N, C_k)$	1

Figura 5. Estructura de la matriz MS-Matriz de Similitudes

- Establecimiento de Vecindario de Contenidos. El vecindario de contenidos es el conjunto de contenidos con mayor probabilidad de ser del gusto de cada usuario activo. Estos son los que registran en la matriz MS un valor de similitud que sobrepasa un umbral establecido con los contenidos de mayor Doi registrado en la matriz MDoi para cada usuario. Las similitudes que sobrepasan el umbral son positivas y las que no, son negativas.
- Cálculo de predicciones. Se calcula el Doi aproximado que podría tener cada usuario activo por cada contenido que hace parte de su vecindario. Suponiendo que el contenido C_j presenta un interés positivo por parte del usuario U_x , y C_k es un contenido del vecindario que presenta una similitud positiva con el contenido C_j , la predicción del Doi del usuario U_x por C_k se calcula a través de (4) [3].

$$Pred(U_x, C_k) = sim_{cos}(C_j, C_k) * DOI_{U_x}(C_j) \quad (4)$$

Donde $DOI_{U_x}(C_j)$ es el nivel de interés del usuario U_x por el contenido C_j . Las predicciones se registran en la matriz MP (Matriz de Predicciones) mostrada en la Fig. 6.

		Vecindario de Contenidos TopN			
		C ₁	C ₂	..C _n	..C _N
Usuario Activo	U _x	$P_{U_x,1}$	$P_{U_x,2}$	$P_{U_x,n}$	$P_{U_x,N}$

Figura 6. Estructura de la matriz MP-Matriz de Predicciones

- Obtención de la lista de recomendaciones basadas en contenido. La lista de recomendaciones se compone de los contenidos con predicciones positivas en la matriz MP (mayores a un umbral establecido).
- 3) **Filtrado Colaborativo.** Los pasos definidos en el algoritmo para este proceso son:
- Establecimiento de pre-vecindario de usuarios. El pre-vecindario de usuarios es el conjunto de usuarios con mayor probabilidad de compartir preferencias con el usuario activo. Estos son los que pertenecen a las comunidades de las que el usuario activo es miembro. Los siguientes pasos se aplican sobre un pre-vecindario para reducir la carga computacional.
 - Cálculo de correlaciones entre el usuario activo y sus vecinos. La correlación es una medida de la fuerza y dirección de la relación entre dos variables estadísticas, en este caso, entre los perfiles de los usuarios. Sea U_a el usuario activo y V_b un usuario de su vecindario, la correlación de Pearson entre ellos se calcula por medio de (5) [12].

$$corr_p(U_a, V_b) = \frac{\frac{\sum_{m=1}^M DOI_{U_a}(Am) \cdot DOI_{V_b}(Am)}{M} - \overline{DOI_{U_a}} \cdot \overline{DOI_{V_b}}}{S_{U_a} \cdot S_{V_b}} \quad (5)$$

Donde $DOI_{U_a}(Am)$ y $DOI_{V_b}(Am)$ es el Doi por la comunidad virtual Am del usuario U_a y V_b respectivamente; M es el total de comunidades del sistema; $\overline{DOI_{U_a}}$ y $\overline{DOI_{V_b}}$ es el Doi promedio por todas las comunidades del sistema de los usuarios U_a y V_b respectivamente; S_{U_a} y S_{V_b} es la desviación estándar de los usuarios U_a y V_b respectivamente a partir de los Doi por las comunidades. La desviación estándar se calcula por medio de (6) [12].

$$S_{U_x} = \sqrt{\frac{\sum_{m=1}^M DOI_{U_x}(Am)^2}{M} - \overline{DOI_{U_x}}^2} \quad (6)$$

Las correlaciones entre el usuario activo y sus vecinos se registran en la matriz MC (Matriz de Correlaciones) de la Fig. 7.

		V ₁	V ₂	..V _b	..V _X
		U _a	$corr_p(U_a, V_1)$	$corr_p(U_a, V_2)$	$corr_p(U_a, V_b)$

Figura 7. Estructura de la matriz MC-Matriz de Correlaciones

- Establecimiento del vecindario de usuarios. El vecindario de usuarios del usuario activo es el conjunto de usuarios del pre-vecindario con las correlaciones que superen un umbral establecido.
- Establecimiento del vecindario de Contenidos. Este vecindario es un conjunto de los contenidos con mayor Doi registrado en la matriz MDoi para los usuarios del vecindario de usuarios. Sobre este conjunto de contenidos se realiza la predicción para obtener la lista de recomendaciones colaborativas.
- Cálculo de Predicciones. Para cada uno de los contenidos del vecindario de contenidos, se calcula la predicción del Doi que podría tener el usuario activo por ellos. La predicción del Doi del usuario activo U_a por el contenido n perteneciente al vecindario de contenidos está dado por (7) [2].

$$P_{U_a,n} = \frac{\sum_{k=1}^V \text{corr}_p(U_a, V_k) \cdot R_{V_k}(n)}{V} \quad (7)$$

Donde V es el tamaño del vecindario de usuarios del usuario activo U_a ; $\text{corr}_p(U_a, V_k)$ es la correlación entre el usuario activo y el k –ésimo usuario vecino; $R_{V_k}(n)$ es el Doi del k –ésimo usuario vecino por el contenido n .

Las predicciones se registran en la matriz MPC (Matriz de Predicciones Colaborativas) presentada en la Fig. 8.

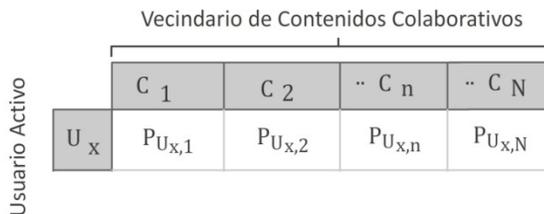


Figura 8. Estructura de la matriz MPC-Matriz de Predicciones Colaborativas

- Obtención de la lista de recomendaciones colaborativas. La lista de recomendaciones se compone de los contenidos con predicciones positivas (mayores a un umbral establecido).

IV. RESULTADOS EXPERIMENTALES Y ANÁLISIS

El algoritmo se implementó a través de componentes Java que cumplen las funciones establecidas en cada paso del filtrado. Las matrices se modelaron en tablas relacionales MySQL.

A. Pruebas del algoritmo

Las pruebas del algoritmo se realizaron en el contexto del proyecto ST-CAV, en el que se registraron 12 (doce) Comunidades Académicas Virtuales, 11 (once) usuarios

miembros de 4 comunidades en promedio, 12 (doce) contenidos asociados a 3 Comunidades en promedio y un promedio de 9 Doi por usuario. El análisis de resultados se realiza para un usuario activo miembro de 8 (ocho) Comunidades que ha registrado 10 (diez) Doi.

Dos tipos de pruebas se desarrollaron: 1) pruebas de precisión sobre el filtrado colaborativo; 2) pruebas relacionadas con el problema Cold-Start sobre el filtrado basado en contenido.

Pruebas de precisión. Un parámetro importante en un algoritmo colaborativo es la cantidad necesaria de vecinos para alcanzar bajos niveles de error en las predicciones. Para la realización de esta prueba, se pone en funcionamiento el algoritmo de filtrado, intentando predecir un conjunto de Doi reales (datos de prueba) que el usuario activo ha registrado previamente. Entre las predicciones y los Doi reales se calcula el Error Medio Absoluto (MAE) a través de (8) [3], variando la cantidad de vecinos del usuario activo.

$$|\overline{\text{MAE}}|_{U_a} = \frac{\sum_{i=1}^N |p_i - r_i|}{N} \quad (8)$$

Donde $|\overline{\text{MAE}}|_{U_a}$ es el Error Medio Absoluto para el usuario activo U_a , N es el total de predicciones tenidas en cuenta para el cálculo, p_i es la predicción entregada por el algoritmo por el i –ésimo contenido, r_i es el Doi real por el i –ésimo contenido.

En la tabla 1 se muestra los valores de MAE del algoritmo propuesto y los del algoritmo tradicional [3], variando la cantidad de vecinos de 1 a 5, y el porcentaje de aciertos en las predicciones.

Cantidad de vecinos	MAE Algoritmo propuesto	Porcentaje de aciertos	MAE Algoritmo tradicional
1	0,7	57%	0,87
2	0,68	57%	0,87
3	0,67	57%	0,87
4	0,65	57%	0,86
5	0,63	57%	0,86

Tabla 1. Tabla comparativa de valores de MAE y porcentaje de aciertos del filtrado colaborativo entre el algoritmo propuesto y el tradicional

Pruebas relacionadas con el problema Cold-Start sobre el filtrado basado en contenido. Tal como se menciona en el capítulo II, el problema denominado Cold-Start consiste en un retraso en la entrega de recomendaciones a un usuario nuevo en el sistema del que no se tiene la información suficiente relacionada con sus preferencias. Esta prueba permite conocer el número de recomendaciones entregadas por el filtrado basado en contenido variando la cantidad de Doi registrados por el usuario activo. Como parámetro inicial, el usuario activo está registrado en dos Comunidades Académicas.

En la tabla 2 se muestra la comparación de la cantidad de recomendaciones entregadas por el algoritmo de filtrado basado en contenido y el algoritmo tradicional en función del número de Doi registrados por el usuario activo.

Cantidad de Doi registrados por el Usuario activo	Cantidad de recomendaciones dadas por el algoritmo propuesto	Cantidad de recomendaciones dadas por el algoritmo tradicional
1	0	0
2	0	0
3	1	0
4	3	0
5	5	0

Tabla 2. Tabla comparativa de la cantidad de recomendaciones dadas por el algoritmo propuesto y el tradicional en función de la cantidad de Doi del usuario activo

B. Análisis de resultados

De la tabla 1 se observa que el filtrado colaborativo del algoritmo propuesto permite alcanzar buenos niveles de MAE con pocos vecinos. Teniendo en cuenta que en el ámbito de las recomendaciones existen ambigüedades debido a la gran cantidad de variables manejadas, el algoritmo propuesto alcanza un valor de MAE de 0,63 con 5 vecinos, valor que es alcanzado por otros algoritmos con entre 30 y 40 vecinos. Los valores de MAE son obtenidos con menor cantidad de vecinos debido a que las métricas (correlación y predicción), se calculan en función de los Doi del usuario activo por las comunidades que comparte con sus vecinos y no por los contenidos directamente. Así, es más probable que los usuarios compartan preferencias por las comunidades que son menores en cantidad, que por los contenidos que representan un mayor número.

En la tabla 2 se observa que el filtrado basado en contenido permite obtener recomendaciones a partir de 2 (dos) Doi registrados en el sistema, reduciendo los efectos negativos del Cold-Start. Esto se debe a que: 1) el algoritmo encuentra similitudes entre los contenidos a partir de la cantidad de comunidades que comparten, y no a partir de los atributos como en otros algoritmos [13]. Así, un Doi registrado por el usuario activo en el sistema obtiene un peso proporcional al número de comunidades a las que está asociado; 2) es más probable que los contenidos compartan comunidades que atributos, por lo que la dispersión de datos es menor y es posible el cálculo de similitudes de manera anticipada.

V. CONCLUSIONES

El algoritmo híbrido-mixto de filtrado colaborativo y basado en contenido, permite la recomendación de contenidos audiovisuales en entornos compuestos por comunidades de personas con intereses similares, basándose en el modelado de datos en matrices y uso de métricas de similitudes, correlaciones y predicciones. Debido a la dinámica colaborativa de las comunidades, el algoritmo genera recomendaciones teniendo en cuenta las preferencias individuales y las de otros usuarios, reduciendo los efectos de la sobre-especialización. A diferencia de otros algoritmos en los que las métricas están en función del nivel de interés del usuario por los

contenidos y sus atributos, en esta propuesta el algoritmo híbrido filtra información en función del nivel de interés por las comunidades, haciendo un uso eficiente de los niveles de interés registrados por cada usuario y reduciendo la dispersión de datos (Ver Tablas 2 y 3). De esta manera, es posible: 1) el cálculo anticipado de vecinos que permitan la reducción del error en las predicciones; 2) la reducción de los efectos negativos del problema Cold-Start, obteniendo recomendaciones aun con una baja cantidad de niveles de interés registrados en el sistema.

Por otro lado, el algoritmo tiene en cuenta la dinámica de las comunidades, adaptando el concepto de vecino tratado en la teoría de los sistemas de recomendaciones y, de esta manera, reducir la carga computacional del proceso de filtrado.

AGRADECIMIENTOS

Especial agradecimiento a los miembros de los grupos de investigación en Televisión Digital Interactiva de la Universidad del Cauca (Colombia) y la Universidad de Vigo (España) [8], [9].

REFERENCIAS

- [1] H. Chan, G. Wang, and X. Yin, "A Cold-start Recommendation Algorithm Based on New User's Implicit Information and Multi-Attribute Rating Matrix," *Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference on*, vol. 2, pp. 353-358, Agosto 2009.
- [2] Y. Blanco, *Propuesta Metodológica para el Razonamiento Semántico en Sistemas de Recomendación Personalizada y Automática*. Vigo, España: Universidad de Vigo, 2007.
- [3] R. Turrin and P. Cremonesi, "Recommender Systems for Interactive TV," *8th European Conference on Interactive TV and Video*, 2010.
- [4] R. Sotelo, *Servicios de Recomendación de Contenidos Audiovisuales para grupos de Individuos*. Vigo, España: Universidad de Vigo, 2010.
- [5] V. Kulkarni, and G.R. Vekariya, "Notice of Violation of IEEE Publication Principles Hybrid Recommender Systems: Content-Boosted Collaborative Filtering for Improved Recommendations," *Communication Systems and Network Technologies (CSNT), 2012 International Conference on*, pp. 649-653, Mayo 2012.
- [6] J. Jiang, "Scaling-Up Item-Based Collaborative Filtering Recommendation Algorithm Based on Hadoop," *Services (SERVICES), 2011 IEEE World Congress on*, pp. 490-497, Julio 2011.
- [7] H. YiBo, "An Item Based Collaborative Filtering Using Item Clustering Prediction," *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, vol. 4, pp. 54-56, Agosto 2009.

- [8] STCAV. [Online]. <http://www.unicauca.edu.co/stcav>
- [9] G. Chanchí, J. Arciniegas, and C. Acevedo, "Surveying Mobile Television," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 3, no. 1, pp. 25-35, 2011.
- [10] A. Rusu, C. Collazos, C. Arciniegas, J. Solano, "Usability Collaborative Evaluation in T-Learning Services: State of the Art," *International Journal on Human-Computer Interaction - e-minds*, vol. 2, no. 8, pp. 19-33, 2012.
- [11] Y. Bai, "Semantic Interaction Model of Knowledge Transfer Process in Virtual Communities of Practice," *Information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on*, vol. 3, pp. 282-285, Septiembre 2011.
- [12] A. Vargas, *Estadística Descriptiva e Inferencial*. España: Universidad de Castilla La Mancha, 1995.
- [13] S. Shiyu, "A creative personalized recommendation algorithm—User-based Slope One algorithm," *Systems and Informatics (ICSAI), 2012 International Conference on*, pp. 2203 - 2207, Mayo 2012.