UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Tecnura

http://revistas.udistrital.edu.co/ojs/index.php/Tecnura/issue/view/680

DOI: http://dx.doi.org/10.14483/udistrital.jour.tecnura.2015.3.a07

INVESTIGACIÓN

Recomendación de productos a partir de perfiles de usuario interpretables

Products recommendation based on interpretable user profiles

Claudia Jeanneth Becerra Cortés* Sergio Gonzalo Jiménez Vargas**, Fabio A. González***, Alexander Gelbukh****

Fecha de recepción: 9 de octubre de 2014 **Fecha de aceptación:** 6 de abril de 2015

Como citar: Becerra Cortés, C. J., Jiménez Vargas, S. G., González, F. A., & Gelbukh, A. (2015). Recomendación de productos a partir de perfiles de usuario interpretables. *Revista Tecnura*, *19*(45), 89-100. doi: 10.14483/udistrital.jour.tecnura.2015.3.a07

RESUMEN

Los sistemas de recomendación automática de productos permiten que los usuarios tengan una visión personalizada de grandes conjuntos de productos, lo cual alivia el problema de la sobrecarga de opciones en los sitios de comercio electrónico. Usualmente las recomendaciones se obtienen usando la técnica denominada "filtrado colaborativo". Esta técnica permite filtrar los productos que el usuario desea de aquellos que no desea, infiriendo las afinidades entre productos, y usuarios, en un espacio de características abstracto. Si bien estas técnicas han mostrado ser de gran valor predictivo, su baja (o nula) interpretabilidad hace que el usuario, al no poder modificar su perfil, quede encerrado en una especie de burbuja, en la cual solo recibe recomendaciones colaborativas condicionadas por su comportamiento histórico. En este trabajo proponemos construir perfiles de usuario definidos en espacios interpretables como el de las etiquetas colaborativas (*tags*) o bien palabras claves extractadas automáticamente de las descripciones de los productos, que al ser interpretables permitan al usuario modificar su propio perfil. Este modelo se basa en la obtención de perfiles usando modelos lineales, cuyos coeficientes, positivos o negativos, reflejan la afinidad del usuario hacia la etiqueta o a la palabra clave. Para probar nuestra hipótesis, utilizamos el conjunto de datos de investigación en sistemas de recomendación de películas de la Universidad de Minnesota, MovieLens; los resultados obtenidos muestran que la capacidad predictiva del modelo es comparable a la de los métodos no interpretables, con el beneficio adicional de la interpretabilidad.

***Palabras clave*:** etiquetado social, filtrado colaborativo, interfaces de usuario, sistemas de etiquetado colaborativo, sistemas de recomendación.

ABSTRACT

Recommender systems allow users to have a personalized view of large sets of products, relieving the overload problem of choice in e-commerce sites. Usually,

* Ingeniera de sistemas, magíster y candidata a doctor en Ingeniería de Sistemas y Computación de la Universidad Nacional de Colombia, Bogotá, Colombia. Contacto: cjbecerrac@unal.edu.co

** Ingeniero de sistemas, magíster y candidato a doctor en Ingeniería de Sistemas y Computación de la Universidad Nacional de Colombia, Bogotá, Colombia. Contacto: sgjimenezv@unal.edu.co

*** Ingeniero de sistemas, magíster en Matemáticas, magíster y doctor en Ciencias de la Computación. Profesor Asociado del Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia. Contacto: fagonzalezo@unal.edu.co

**** Bachelor of Science, magíster en Matemáticas, doctor en Ciencias de la Computación. Profesor-investigador y profesor colegiado del Laboratorio de Lenguaje Natural y Procesamiento de Texto del Centro de Investigación en Computación del Instituto Politécnico Nacional de México, Ciudad de México, México. Contacto: gelbukh@cic.ipn.mx.

recommendations are obtained using the technique called "collaborative filtering". This technique filters the products the users wish, from those they don´t want, inferring affinities between products and users in a space of abstract features, also called a latent space. These techniques have proven to be of great predictive value, but these created profiles are neither understandable, nor editable for users, enclosing users in a *bubble*, in which they only receive collaborative recommendations conditioned by their historical behaviors. In our work we propose a method to build user profiles, defined in interpretable spaces, or defined in terms of collaborative tags or keywords (i.e. words extracted from the descriptions of the product), which can be interpreted and modified by users. The model proposed generate linear profiles, whose coefficients, positive or negative, reflect the user's affinity towards tags or keywords, according to the space selected. To test our hypothesis, we used the dataset of research in movie recommender systems from the University of Minnesota: Movielens. The results show that the predictive ability of the model, based on interpretable user profiles, is comparable to those models based on abstract profiles with the added benefit that these profiles are interpretable.
Keywords: collaborative filtering, collaborative tagging systems, recommender systems, social tagging, user interfaces.

INTRODUCCIÓN

Las técnicas utilizadas en la exploración de grandes colecciones de productos, tales como libros (e.g. librarything.com), películas (e.g. netflix.com), fotos (e.g. flickr.com), artículos científicos (e.g. citeulike.com), web bookmarks (e.g. del.icio.us), etc., se construyen a partir de la investigación en artefactos colaborativos que permitan que el conocimiento, de ciertos individuos acerca de ciertos ítems, se propague a otros, siendo su objetivo generar una inteligencia colaborativa que permita guiar a los usuarios en búsquedas, personalizándolas hacia sus gustos, alejándolo de sus preferencias negativas. De esta forma se logra una experiencia de exploración que elimina la frustración que genera el recorrer inmensas colecciones de productos sin descubrir nada nuevo o interesante.

Los métodos tradicionales de recomendación y también los más precisos construyen perfiles de usuarios, basados en su historial de navegación o basados en evaluaciones explícitas de los productos. A partir de estas evaluaciones, el sistema infiere un conjunto común de características, en un espacio abstracto, que comparten el producto y el usuario, con lo cual se minimiza el error de predicción del conjunto de datos de entrenamiento. En otras palabras, la evaluación es equivalente a la medida de afinidad entre el vector usuario y el vector producto definido en este espacio abstracto. Este método logra capturar así las interrelaciones entre usuarios y productos, de modo que permite predecir la evaluación que un usuario dará a un producto a partir de sus evaluaciones pasadas y de las evaluaciones que usuarios similares han hecho de este producto.

Sin embargo, estos espacios latentes no son interpretables e imposibilitan la interacción directa de los usuarios con sus perfiles. El usuario queda inmerso en una especie de burbuja condicionada a su historial previo, y cambiar su perfil ya solo es posible de manera indirecta, esto es, cambiando sus evaluaciones pasadas o encontrando nuevos productos que modifiquen su perfil de manera manual. Por tal razón, en este trabajo queremos proponer un método de obtención de perfiles de usuarios interpretables, construidos en espacios ya no abstractos, sino espacios que tengan sentido para el usuario y con los cuales pueda interactuar. Ejemplos de estos espacios pueden ser: espacios de palabras claves extractadas de las descripciones textuales de los productos (Lops, Gemmis, & Semeraro, 2011), o espacios de etiquetas colaborativas, las cuales denominaremos en adelante *tags*,

dado el extendido uso de ese término en aplicaciones en internet (Lops *et al*., 2011). No obstante, estos espacios no necesariamente quedan limitados a los anteriores. Este modelo puede ser utilizado para calcular los perfiles de usuario en cualquier espacio, como por ejemplo espacios de características de los productos como géneros, actores, o en combinaciones de varios espacios.

Para ilustrar gráficamente la idea de los perfiles interpretables de usuario que pretendemos calcular, y su potencial como artefacto de interacción con usuarios, las figuras 1 y 2 muestran un ejemplo de interacción con un perfil de usuario basado en *tags*. Para esto se utilizó el conjunto de películas MovieLens (movielens.org, de la Universidad de Minnesota) y el subconjunto de *tags* correspondiente al trabajo The Tag Genome (Vig, Jesse, Sen, & Riedl, 2012). La visualización corresponde al prototipo de recomendación que se ha estado desarrollando con el proyecto financiado por Colciencias y la Universidad Nacional de Colombia sede Bogotá, código 1101-521-28465, denominado "Sistema de recomendación basado en conocimiento extractado de manera automática para ambientes de comercio electrónico".

En la figura 1 se observan las preferencias de un usuario divididas verticalmente en dos zonas. En la parte izquierda se muestran los *tags* y las evaluaciones de las películas que le agradan al usuario, y en la parte derecha los que le desagradan. El objetivo de este trabajo es obtener un modelo lineal donde los coeficientes de afinidad del usuario para cada uno de los *tags* se obtienen a partir de las evaluaciones que el usuario realiza. Estos coeficientes pueden ser positivos, como los mostrados en el lado izquierdo de la figura 1, o negativos, como los que están a la derecha. Al leer el perfil del usuario del ejemplo, vemos que sus *tags* preferidos, en orden de mayor a menor preferencia, son: "miyasaki, superhero, boarding school, sequel, franchise, fantasy, adventure, pixar…", y los tags de mayor preferencia negativa son: "talking animals, heartwarming, alternate reality, hilarious, underdog…". Estos *tags* nos podrían dar una

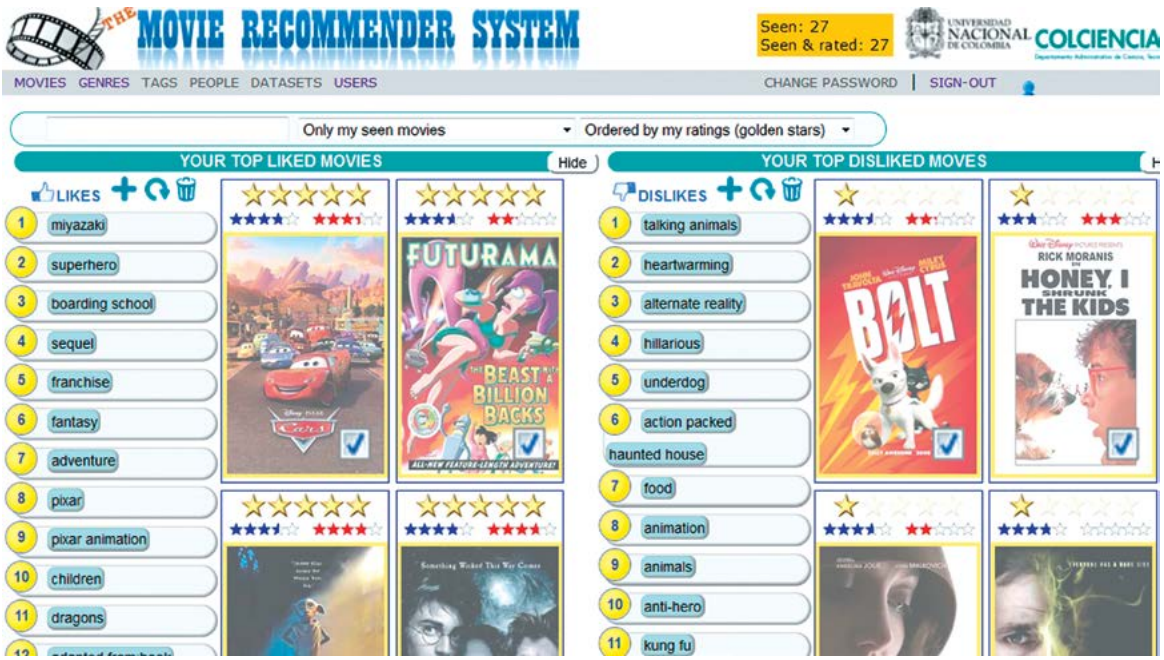

Figura 1. Ejemplo de perfil de usuario basado en *tags*.

Fuente: elaboración propia.

idea de las preferencias y del filtrado colaborativo que el sistema de recomendación utiliza para formular sus recomendaciones.

Ahora bien, para ilustrar la idea del potencial de interacción con los perfiles, en la figura 2 se muestran las nuevas recomendaciones que el usuario recibe después de ajustar su perfil insertando en primer lugar de preferencia los *tags* relacionados con las diferentes categorías de los premios Oscar. Como se puede observar, esta es una manera bastante rápida de escapar de la burbuja e inmediatamente recibir recomendaciones colaborativas de usuarios ya familiarizados con las películas y que las han etiquetado, sin dejar atrás su perfil.

En este trabajo se experimentó con la construcción de perfiles de usuario interpretables basados en *keywords* y en *tags*. Su alcance se centra en dos objetivos principales, a saber: 1) proponer un método para obtener perfiles interpretables y 2) comparar la calidad de las recomendaciones generadas con perfiles interpretables con aquellas obtenidas con perfiles no interpretables utilizando el método de factorización de matrices. Como plataforma de pruebas se trabajó con los datos del sistema de recomendación de películas movielens.org de la Universidad de Minnesota.

La organización de este artículo es la siguiente: en primer lugar se mostrará el trabajo relacionado en el cálculo de perfiles de usuario; posteriormente se explicará el modelo de cálculo propuesto; se describirá el marco experimental y las medidas de desempeño a utilizar; luego se mostrarán los resultados y, para terminar, se propondrá una visualización de los perfiles obtenidos.

TRABAJO RELACIONADO: EL MÉTODO DE FACTORIZACIÓN DE MATRICES

Probablemente, el método más preciso y popular utilizado para recomendar productos es la factorización de matrices (Bell R.M., Koren Y. & C, 2007; Koren, Bell, & Volinsky, 2009). En este modelo el

Figura 2. Ejemplo de interacción con perfiles de usuarios basados en *tags*.

Fuente: elaboración propia.

valor estimado de la evaluación (*rating*) $\tilde{r}_{um}$ que daría el usuario u al ítem m, es modelado como la suma de los sesgos del usuario y del ítem, más una medida de afinidad $\hat{r}_{um}$ entre el usuario y el ítem. Para determinar esta afinidad, se encuentra la caracterización tanto de los usuarios como de los ítems, en un espacio abstracto de dimensión preestablecida f, que da lugar a la medida de afinidad observada en los datos de entrenamiento. De esta manera, la parte de la evaluación explicada por la afinidad entre el usuario y el ítem podría escribirse como se plantea en la ecuación (1).

$$\hat{\boldsymbol{r}}_{\boldsymbol{um}} = \vec{\boldsymbol{U}}_{\boldsymbol{u}\to\mathcal{R}^f} \cdot \left(\vec{\boldsymbol{M}}_{\boldsymbol{m}\to\mathcal{R}^f}\right) \tag{1}$$

Donde $\vec{U}_{u\to\mathcal{R}^f}$ y $\vec{M}_{m\to\mathcal{R}^f}$ denotan la caracterización del usuario u y el ítem m en el espacio latente $\mathcal{R}^f$ respectivamente. Aquí, la medida de afinidad utilizada es el producto punto. A su vez, los vectores de caracterización del usuario $\vec{U}_{u\to\mathcal{R}^f} = [U_{u1}, \dots, U_{ui}, \dots, U_{uf}]$ y del ítem $\vec{M}_{m\to\mathcal{R}^f} = [M_{m1}, \dots, M_{ui}, \dots, M_{uf}]$ estan compuestos bien por los *coeficientes individuales de afinidad* en el caso de los usuarios, o por los *coeficientes de relevancia* M_{ui} en el caso de los ítems. El subíndice i define la i-ésima dimensión del espacio latente, donde $1 \leq i \leq f$. Por tal razón la evaluación del producto puede rescribirse como la ecuación (2).

$$\hat{r}_{um} = \sum_{i=1}^{f} (U_{ui} \cdot M_{mi}) \tag{2}$$

En la cual los coeficientes de caracterización U_{ui} y M_{mi} se encuentran minimizando el error de predicción e_{um}, el cual se calcula con expresión de la ecuación (3).

$$e_{um} = \left(r_{um} - \sum_{i=1}^{f} (U_{ui} \cdot M_{mi}) \right)^2 \tag{3}$$

Para evitar el sobreajuste, es común introducir un coeficiente de regularización β que penaliza la norma de las caracterizaciones de los usuarios y los ítems. De esta manera el error regularizado de predicción $\ddot{e}_{um}$ se define como la ecuación (4).

$$\ddot{e}_{um} = e_{um} + \beta \left(\left\|\vec{U}_{u\to\mathcal{R}^f}\right\|^2 + \left\|\vec{M}_{m\to\mathcal{R}^f}\right\|^2 \right) \tag{4}$$

Finalmente, los vectores de caracterización de usuarios e ítems son encontrados minimizando el error de estimación en el conjunto de entrenamiento en la ecuación (5).

$$\min_{\vec{U}_u, \vec{M}_u} \sum_{r_{um} \in \mathbb{R}_{U\times M} \wedge r_{um} \neq 0} \left(r_{um} - \sum_{i=1}^{f} (U_{ui} \cdot M_{mi}) \right)^2 + \beta \left(\left\|\vec{U}_{u\to\mathcal{R}^f}\right\|^2 + \left\|\vec{M}_{m\to\mathcal{R}^f}\right\|^2 \right) \tag{5}$$

Una vez encontrados los vectores $\vec{U}_u$ y $\vec{M}_m$, se calcula la matriz de estimación de evaluaciones $\widehat{\mathbb{R}}_{U\times M}$. Esta matriz se utilizará como base para nuestro método.

MÉTODO PROPUESTO

La explicación del modelo propuesto se hará en tres partes. En la primera se presentará el modelo general de cálculo de perfiles en espacios interpretables propuesto; en la segunda se explicará su aplicación particular a espacios de palabras claves (*keywords*) extractadas directamente de las descripciones textuales de los productos, y en la tercera se explicará su aplicación a espacios de *tags* colaborativos.

Un método general para extracción de perfiles de usuario lineales

Una vez obtenida la matriz de estimación de ratings $\widehat{\mathbb{R}}_{U\times M}$, con el método de factorización de

matrices, podemos asumir ahora que esta matriz es una consecuencia de la afinidad de los usuarios a los productos, pero ya no caracterizados en un espacio latente, sino en un espacio de características interpretables de tamaño , como se muestra en la ecuación (6).

$$\widehat{\mathbb{R}}_{U\times M} = \mathbb{U}_{U\times X} \cdot (\mathbb{M}_{M\times X})^{\mathrm{T}} \tag{6}$$

Donde $\mathbb{U}_{U\times X}$ es la matriz de perfiles de usuarios en $\mathcal{R}^X$, y $\mathbb{M}_{M\times X}$ es la matriz de representación de los ítems en el espacio $\mathcal{R}^X$, corresponde al número de usuarios, y M corresponde al número de ítems. La matriz $\mathbb{U}_{U\times X}$, de tamaño $U \times X$ también puede ser expresada como en la ecuación (7).

$$\mathbb{U}_{U\times X} = \begin{bmatrix} U_{11} & \cdots & U_{1X} \\ \vdots & \ddots & \vdots \\ U_{U1} & \cdots & U_{UX} \end{bmatrix} = \begin{bmatrix} \vec{U}_{1\to\mathcal{R}^X} \\ \vdots \\ \vec{U}_{U\to\mathcal{R}^X} \end{bmatrix} \tag{7}$$

Donde U_{ux} representan los coeficientes de afinidad entre el u-ésimo usuario y la x-ésima dimensión en el espacio $\mathcal{R}^X$, para valores de u $\{1, \dots, U\}$ en y valores de x en $\{1, \dots, X\}$. El vector $\vec{U}_{u\to\mathcal{R}^X}$ denota el perfil del usuario en el espacio $\mathcal{R}^X$.

De la misma manera, la matriz de perfiles de ítems caracterizados en el espacio $\mathcal{R}^X$, $\mathbb{M}_{M\times X}$ puede también expresarse como la ecuación (8).

$$\mathbb{M}_{M\times X} = \begin{bmatrix} M_{11} & \cdots & M_{1X} \\ \vdots & \ddots & \vdots \\ M_{M1} & \cdots & M_{MX} \end{bmatrix} = \begin{bmatrix} \vec{M}_{1\to\mathcal{R}^X} \\ \vdots \\ \vec{M}_{M\to\mathcal{R}^X} \end{bmatrix} \tag{8}$$

Donde M_{mx} denota el coeficiente de relevancia del ítem m-ésimo en la x-ésima dimensión del espacio $\mathcal{R}^X$, para valores de m en $\{1, \dots, M\}$ y x en $\{1, \dots, X\}$. $\vec{M}_{m\to\mathcal{R}^X}$ representa el perfil del ítem m en el espacio $\mathcal{R}^X$.

Ahora es posible escoger un espacio interpretable $\mathcal{R}^X$ en el cual la matriz de perfiles de producto $\mathbb{M}_{M\times X}$ sea conocida. Así, los perfiles de usuario pueden ser directamente calculados utilizando la ecuación (9).

$$\mathrm{U}_{\mathrm{U\times X}} = \widehat{\mathbb{R}}_{U\times M} \cdot \left((\mathrm{M}_{\mathrm{M\times X}})^{\mathrm{T}}\right)^{\dagger} \tag{9}$$

Donde $((\mathbb{M}_{M\times X})^{\mathrm{T}})^{\dagger}$ denota la seudoinversa (Penrose & Todd, n.d.) de la transpuesta de la matriz de la representación de los ítems en $\mathcal{R}^X$, y $\widehat{\mathbb{R}}_{U\times M}$ es la matriz de evaluaciones conocidas.

Perfiles de usuario basados en palabras claves

La mayoría de enfoques de recomendación basados en descripciones textuales de producto (Lops *et al.*, 2011) utilizan el *modelo de espacio vectorial* (Salton, Wong, & Yang, 1975) para convertir las descripciones textuales de cada ítem m en vectores $\vec{M}_{m\to\mathcal{R}^W}$, cuyos componentes individuales, denotados por M_{mw}, cuantifican la relevancia de la palabra w para el ítem m.

Los coeficientes de relevancia pueden ser inferidos a partir de las ocurrencias de las palabras en la colección total de descripciones textuales de los ítems. La práctica más común es utilizar un esquema de pesado como *tf-idf* (Jones, 1972; Salton *et al.*, 1975) u Okapi BM-25 (Robertson, 2005). Estas técnicas hacen un balance entre la heurística de asignar pesos pequeños a palabras muy comunes en la colección y la de asignar pesos altos a las palabras frecuentes en una descripción de un ítem. Esta combinación de heurísticas busca encontrar las palabras más representativas para cada ítem.

Con las descripciones textuales de cada ítem, ya representadas como vectores $\vec{M}_{m\to\mathcal{R}^W}$, se forma la matriz de relevancia ítems-palabras $\mathbb{M}_{M\times W}$ a partir de la cual se calculan los perfiles de usuario $\vec{U}_{U\times W}$ evaluando la ecuación (10).

$$\mathbb{U}_{U\times W} = \widehat{\mathbb{R}}_{U\times M} \cdot \left((\mathbb{M}_{M\times W})^{\mathrm{T}}\right)^{\dagger} \tag{10}$$

Perfiles de usuario basados en tags

De la misma manera que en los perfiles de usuario basados en palabras, la matriz $\mathbb{U}_{U\times T}$ de perfiles de usuarios basados en *tags* es calculada utilizando la representación de los ítems en el espacio de *tags* $\mathbb{M}_{M\times T}$, evaluando la ecuación (11).

$$\mathbb{U}_{U\times T} = \hat{\mathbb{R}}_{U\times M} \cdot \left((\mathbb{M}_{M\times T})^{\mathrm{T}}\right)^{\dagger} \quad (11)$$

Los perfiles individuales de los productos $\vec{M}_{m\to\mathcal{R}^T}$, que componen la matriz $\mathbb{M}_{M\times T}$ pueden ser obtenidos utilizando una gran variedad de técnicas (Lops et al., 2011; Zhang, Zhou, & Zhang, 2011). El más simple de los enfoques consiste en una aplicación booleana, que asigna $M_{mt} = 1$ cuando el *tag t* ha sido asignado al ítem *m* y asigna $M_{mt} = 0$ en caso contrario. Es importante notar que en el método propuesto se pueden esperar mejores resultados cuando se utilizan *tags* independientes entre ellos. La independencia puede ser lograda agrupando *tags* que están morfológicamente relacionados utilizando *stemmers* (reducir las palabras a sus raíces) o lematizadores (reducir las palabras a sus lemas). Lops *et al.* (2009) fueron más allá agrupando los *tags* por similitud semántica utilizando *synsets* de WordNet (Fellbaum, 1998).

Los perfiles de ítems con coeficientes continuos también pueden ser obtenidos con métodos que combinan componentes manuales y automáticos. Por ejemplo, en el trabajo de Vig *et al.* (2012) se obtiene la relevancia de los *tags* a las películas realizando una encuesta y preguntando directamente a los usuarios la relevancia de los *tags* a los filmes. Con estos resultados se entrena un modelo de regresión basado en vectores de soporte (Smola & Schölkopf, 1998), que utiliza las descripciones textuales de los ítems para predecir los coeficientes de relevancia del vector de caracterización del ítem en el espacio de *tags*.

Como se observa, la propuesta plantea un modelo que independiza el método de caracterización de los ítems del método de caracterización de los perfiles de usuario.

METODOLOGÍA

Los experimentos realizados buscan evaluar la calidad de las recomendaciones producidas con los perfiles de usuario basados en palabras y en *tags* usando el método propuesto. Posteriormente se comparan estos resultados con los obtenidos utilizando factorización de matrices. Además, esta sección contiene una descripción comprensiva de los datos y la medida utilizada para medir el desempeño de los métodos a comparar.

Datos utilizados

Los datos utilizados corresponden a la unión de los datos de movielens.org, con las descripciones de los productos extractados de la API de Netflix y los *tags* sociales empleados en el trabajo The Tag Genome (Vig, Jesse *et al.*, 2012). A continuación se describe detalladamente el conjunto de datos conformado.

Información colaborativa en el dominio de filmes

El conjunto de usuarios, películas y evaluaciones (*ratings*) fue obtenido de la base de datos de producción del sistema MovieLens en abril de 2012. Este conjunto de datos se filtró extractando aquellos usuarios y películas con más de 1.000 evaluaciones. Así, este conjunto filtrado genera un subconjunto de 200 usuarios, 1.462 películas y 150.915 evaluaciones. La escala de evaluaciones es {1,2,…,5}, donde 5 corresponde al máximo grado de preferencia y 1 el mínimo. La distribución de evaluaciones en el conjunto de datos se muestra en la figura 3. Distribución de evaluaciones (*ratings*) en el subconjunto de datos de MovieLens. El número promedio de evaluaciones por película es 101.6 ($\sigma = 37.5$), y por usuario es 742.5 ($\sigma = 188.5$).

Figura 3. Distribución de evaluaciones (*ratings*) en el subconjunto de datos de MovieLens.

Fuente: elaboración propia.

Descripciones textuales de películas

Las descripciones textuales de las películas se obtuvieron utilizando el campo sinopsis de la API de Netflix (http://developer.netflix.com) durante el año 2012, y la correspondencia con los filmes de MovieLens se obtuvo mediante una cooperación de investigación con el grupo de investigación GroupLens (http://www.grouplens.org).

Estas descripciones textuales se representan como un modelo vectorial de bolsa de palabras. La dimensionalidad de esta representación se reduce con el ánimo de obtener un vocabulario basado en popularidad y relevancia. De esta manera se obtuvo un vocabulario de 5.848 palabras aplicando la siguiente serie de acciones de preprocesamiento: (1) todos los caracteres fueron convertidos a caracteres en minúsculas; (2) los nombres y apellidos de personas fueron concatenados con el símbolo "_"; (3) los números fueron removidos; (4) 334 palabras tomadas de la lista de palabras nulas (*stop words*) de *gensim* (http://radimrehurek.com/gensim) fueron removidas; (5) palabras que ocurren en menos de 10 sinopsis y en más del 95% de las sinopsis fueron removidas; y finalmente (6) todas las marcas de puntuación fueron retiradas.

Los pesos asignados a las palabras, con el fin de establecer la relevancia de la palabra en el vector de caracterización del ítem, fueron obtenidos con la fórmula de recuperación Okapi BM-25 (Robertson, 2005) utilizando el método propuesto por Vanegas, Caicedo, Camargo, & Ramos-Pollán (2012), en el cual el peso $w(p,d)$ de una palabra en un documento (sinopsis) *d* está dada por la ecuación (12).

$$w(p,d) = log\left(\frac{M - df(p)}{M}\right)\frac{(k_1+1)tf(p,d)}{K + tf(p,d)}$$

$$K = k_1\left((1-b) + b\frac{dl(d)}{avdl}\right) \qquad (12)$$

Donde, $df(p)$ es el número de documentos (sinopsis) en el cual aparece la palabra *p*, $M = 1.462$ es el número de películas, $tf(p,d)$ es el número de ocurrencias de la palabra *p* en el documento *d*, y $avdl = 33$ es el largo promedio de los documentos. Los parámetros adicionales k_1 y *b* se asignan en 1,2 y 0,75, respectivamente (Robertson, 2005). En la tabla 1 se muestran dos ejemplos de representación vectorial basada en palabras de las películas *Bewitched* (2005) y *Rocky V* (1190). La agregación de los vectores obtenidos produce la matriz de perfiles de producto $\mathbb{M}_{M\times W}$, cuyas dimensiones son $M = 1.462$ películas y $W = 5.848$ palabras. Esta matriz es dispersa con solo 0,518% de entradas diferentes a cero.

Tabla 1. Ejemplos de palabras clave (keywords) en las descripciones de Netflix

Filme: *Bewitched* (2005)
will_ferrell (0,237), *jack* (0,147), *update* (0,142), *samantha* (0,131), *sitcom* (0,131), *witch* (0,119), *nicole_kidman* (0,119), *convinced* (0,116), *michael_caine* (0,114), *right* (0,107), *hoping* (0,105), *know* (0,103), *career* (0,099), *perfect* (0,098), *doesn't* (0,097), *actor* (0,092), *make* (0,068), *film* (0,045)
Filme: *Rocky V* (1990)
burt_young (0,249), *talia_shire* (0,242), *broke* (0,15), *upandcoming* (0,150), *shots* (0,150), *boxer* (0,150), *crooked* (0,142), *trainer* (0,136), *glory* (0,136), *accountant* (0,131), *ended* (0,131), *lifetime* (0,128), *memory* (0,124), *training* (0,124), *rocky* (0,121), *inspired* (0,107), *taking* (0,101), *career* (0,099), *left* (0,092), *series* (0,071), *takes* (0,063), *finds* (0,058)

Fuente: elaboración propia.

Tags sociales

El conjunto de *tags* utilizados para caracterizar las películas es la selección de *tags* propuesta por Vig *et al.* en The Tag Genome (2012), en la cual se escogen 1.128 *tags* de los cerca de 30.000 *tags* libremente aplicados en el sistema MovieLens. Para su obtención se removieron: *tags* con menos de 10 aplicaciones, *tags* con errores de digitación, nombres de personas y duplicados cercanos. Por último se seleccionaron el 95% de ellos ordenados de acuerdo con la métrica de calidad basada en entropía propuesta por Sen, Harper, LaPitz, & Riedl (2007). Solo 1.081 *tags* de la selección de Vig *et al.* (2012) ocurrieron en los 1.462 filmes. De esta manera las dimensiones finales de la matriz de perfiles de ítems $\mathbb{M}_{M\times T}$ son 1.462 x 1.081.

En el conjunto de datos existen 13.332 aplicaciones de *tags* a las películas. Esto corresponde a 1.370 filmes que tienen al menos un *tag* asociado, para un promedio de 9,7 aplicaciones por película ($\sigma = 8{,}5$). En consecuencia, la matriz de perfiles de ítems es dispersa, con una densidad de 0,844% de entradas diferentes a 0. La distribución de *tags* es considerablemente más uniforme que la distribución Zipf (1950). Así, los 108 *tags* más frecuentes (10%) representan solo el 42% de las aplicaciones de *tags*. Esto se puede ver en la tabla 2, en la cual se muestran ejemplos de *tags* seleccionados de rangos de frecuencia uniformemente separados (e.g. "based on a book" es el *tag* más frecuente).

Tabla 2. Ejemplos de *tags* del conjunto seleccionado de 1.128 *tags* de MovieLens*

Rango	Ejemplos de *tags*
1-3	*based on a book* (194), *comedy* (182), *classic* (143)
9-12	*boring* (107), *70mm* (193), *romance* (98), *quirky* (91)
17-19	*sci fi* (78), *stylized* (64), *adventure*(62), *humorous*(62)
25-26	*crime* (53), *sequel*, *tense*, *violence*, *remake* (52)
34-35	*animation* (42), *politics*, *satirical*, *war*, *hilarious* (41)
42	*bittersweet* (34), *gay*, *historical*, *musical*, *suspense*
50	*forceful* (26), *military*, *satire*, *small town*, *very good*
59	*cult classic* (17), *dark humor*, *earnest*, *epic*, *japan* {17}
67	*action packed*(9), *alien*, *aviation*, *based on comic* {41}
75	*3d*(1), *adoption*, *airplane*, *alcatraz*, *arms dealer*: {80}

* Se coloca en paréntesis el número de películas asociadas al *tag*; si no se coloca, es el mismo número del *tag* precedente. El número de *tags* en el mismo rango es mostrado en corchetes; si no aparece es porque se mostraron todos los *tags* en el mismo rango.

Fuente: elaboración propia.

Medición del desempeño

Con el objetivo de evaluar la calidad de las recomendaciones producidas por los métodos propuestos, se proveen dos escenarios de validación en 10 hojas: validación cruzada y *arranque en frío de producto* (Schein, Pennock, & Ungar, 2002). En el escenario de validación cruzada las evaluaciones se permutan aleatoriamente y se dividen en 10 hojas. En cada hoja, 90% de las evaluaciones son utilizadas para entrenamiento y el restante 10% para pruebas. En el escenario de arranque en frío el procedimiento es el mismo, pero todas las evaluaciones de los ítems incluidos en el conjunto de evaluación son eliminadas.

La medida de rendimiento utilizada para establecer la calidad de las recomendaciones es la raíz del error medio cuadrático (RMSE, por sus iniciales en inglés), definido como se muestra en la ecuación (13).

$$RMSE = \sqrt[2]{\frac{\sum_{\{r_{um}\}\in test}(\hat{r}_{um} - r_{um})^2}{|test|}} \qquad (13)$$

Donde test es el conjunto de datos de evaluación (usuario, íitem, evaluación) y *test* es su cardinalidad.

Métodos de línea de base

En el escenario de validación cruzada los resultados se comparan contra los resultados obtenidos con el método de factorización de matrices, expuesta en la primera sección. El número de factores latentes utilizado es 30 y el parámetro de regularización se establece en 0.07. La minimización de la función de error se realiza utilizando el método de optimización LBFGSB (Byrd, Lu, Nocedal, & Zhu, 1995).

En el escenario de arranque en frío de ítems, el método de línea de base empleado es la evaluación promedio en el sistema más la desviación promedio del usuario. Esto en razón a que los ítems sin evaluaciones no pueden ser caracterizados en el espacio latente.

RESULTADOS

Los resultados de los experimentos realizados se presentan en la tabla 3. Los errores mostrados corresponden al error promedio en 10 hojas, en cada uno de los escenarios de prueba, y la columna σ muestra la desviación estándar del error. Las primeras dos filas muestran los resultados para los métodos de línea de base propuestos para cada

Tabla 3. Resultados obtenidos

MÉTODO	Arranque en frío		Validación cruzada	
	RMSE	σ	RMSE	σ
Promedio del sistema más sesgo de usuario	1,065	0,022	-	-
LBFGSB regularizado (Factorizacion con 30 factores)	-	-	0,995	0,010
Perfiles basados en palabras (Factors+Norm[-1,1])	1,052	0,015	0,939	0,016
Perfiles basados en *tags* (Factors +Norm[-1,1])	1,062	0,021	0,985	0,012

Fuente: elaboración propia.

uno de los escenarios de prueba propuestos, y las siguientes filas presentan los errores de predicción obtenidos con los método de perfilado de usuarios basados en palabras y en *tags*, respectivamente.

Respecto al escenario de validación cruzada, los resultados muestran que los métodos propuestos superan al método de línea de base. Particularmente, la mejora reportada por el método de perfilado basado en palabras es significativa, al estar la medida de rendimiento (error) más de tres desviaciones estándar por debajo.

El escenario de validación de arranque de productos en frío es claramente más desafiante. Como se observa, los métodos de recomendación propuestos escasamente superan el promedio. Sin embargo nuestros métodos muestran ser coherentes en ambos escenarios de prueba, manteniendo la interpretabilidad y la capacidad explicativa del modelo.

CONCLUSIONES

A partir de los resultados obtenidos se pueden sacar dos conclusiones importantes: *i*) el método propuesto permite generar recomendaciones de la misma calidad que el método de factorización de matrices, y *ii*) el método para obtener los perfiles permite aprovechar las diferentes representaciones de los productos, en los diferentes espacios. Ejemplo de esto fue el uso del método de representación de las películas utilizando el modelo de bolsa de palabras con pesaje Okapi BM-25, el cual influye en los buenos resultados obtenidos con el perfil basado en palabras claves. Asimismo, muy probablemente el uso de la representación en términos de *tags* con pesaje booleano interfirió negativamente en los resultados.

Ahora pasaremos a observar, y a comparar, los perfiles obtenidos. El propósito de esta comparación es discutir el potencial de uso y la interpretabilidad de los perfiles obtenidos basados en palabras claves y en *tags*. Con este objetivo en mente, se selecciona un usuario cuyos gustos son muy populares entre los usuarios del sistema, pero también son rechazados por un gran número de usuarios. Para esto se calcula la matriz de correlación de Pearson usuario-usuario, utilizando los perfiles basados en palabras $\mathbb{U}_{U\times W}$ encontrados, y se escoge el usuario con mayor número e intensidad de correlaciones tanto positivas como negativas. Dado que los usuarios son anónimos en el conjunto de datos, nos referiremos a este usuario como el *usuario 156*. En la figura 4 se muestra una visualización de los perfiles basados en palabras claves y en *tags*.

En la figura 4 se muestra en color negro el perfil del usuario 156 y en colores los perfiles de los 10 usuarios más correlacionados con él. La parte izquierda de la figura (figura 4a) muestra el perfil basado en palabras, y a la derecha (figura 4b) se muestra el perfil basado en *tags*. A su vez, para cada perfil se muestran los dos extremos. En el extremo izquierdo se muestran las 40 palabras, o *tags*, con mayores coeficientes de afinidad, que más inciden en la evaluación positiva del ítem; y en el extremo derecho se muestran las palabras o *tags* con menores coeficientes de afinidad y que más inciden en la baja evaluación de los ítems que los contienen.

Ahora, se analizará la legibilidad del perfil. Si observamos el perfil basado en palabras, no es

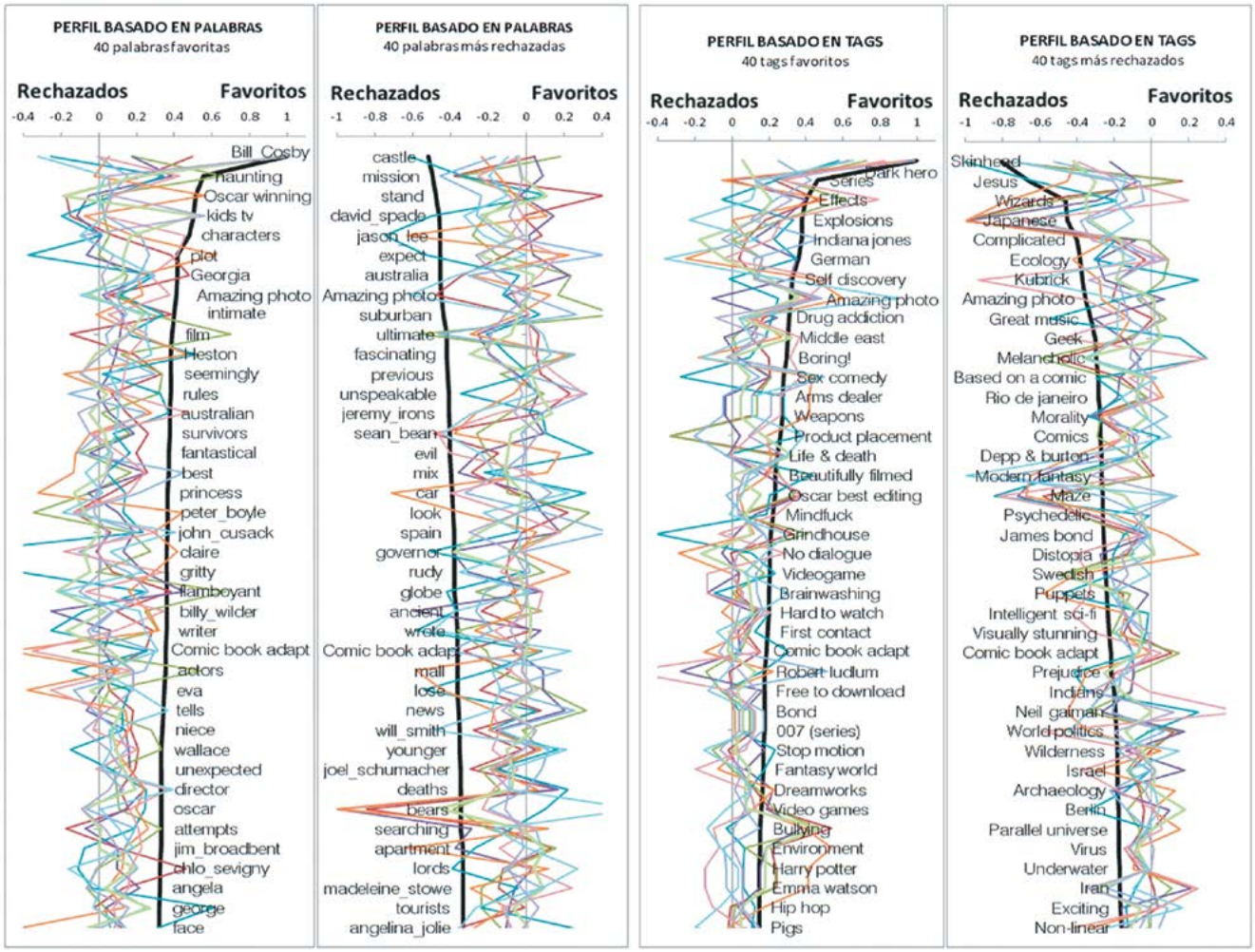

Figura 4. Perfiles basados en palabras claves y en *tags* del usuario 156 (en negro) y de los 10 usuarios más similares a él (en colores). Para cada perfil se muestra a la izquierda las 40 palabras o *tags* con mayores coeficientes de afinidad, o los "favoritos", y a la derecha las 40 palabras o *tags* con menores coeficientes de afinidad o los "más rechazados".

Fuente: elaboración propia.

posible observar un patrón claro. A diferencia de este, en el perfil basado en *tags* se observa que 20 de los 40 *tags* favoritos del usuario se relacionan con películas de acción e infantiles y para adolescentes. Estos *tags* son: *Dark hero*, *Effects*, *Explosions*, *Indiana jones*, *German*, *Drug addiction*, *Arms dealer*, *Weapons*, *Life & death*, *Videogame*, *First contact*, *Comic book adapt*, *Bond, 007 series*, *Stop motion*, *Fantasy world*, *Dreamworks*, *Video games*, *Harry potter*, *Emma Watson*. Esta observación particular nos permite obtener nuestra tercera y última conclusión: los perfiles de usuarios basados en *tags* se plantean como un interesante artefacto de interacción con usuarios, dada su interpretabilidad y coherencia.

AGRADECIMIENTOS

Especiales gracias al profesor John Riedl (q.e.p.d.) de la Universidad de Minnesota, y al profesor Shilad Sen del Macalester College, quienes inspiraron este artículo.

Financiamiento

Este trabajo fue financiado por Colciencias y la Universidad Nacional de Colombia, proyecto 1101-521-28465. El profesor Alexander Gelbukh agradece al Gobierno Mexicano (SNI, COFAA-IPN, SIP 20131702, CONACYT 50206-H) y CONACYT–DST.

REFERENCIAS

Bell R.M., Koren Y., & C, V. (2007). The BellKor solution to the Net Flix Prize. *Technical Report, AT&T Labs Research*. doi:http://www.netflixprize.com/assets/ProgressPrize2007_KorBell.pdf

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, *16*(5), 1190–1208. doi:10.1137/0916069

Fellbaum, C. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press. Retrieved from http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=8106

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11-21.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30-37. doi:10.1109/MC.2009.263

Lops, P., Gemmis, M., & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 73-105). Springer US. Retrieved from http://dx.doi.org/10.1007/978-0-387-85820-3_3

Lops, P., Gemmis, M., Semeraro, G., Musto, C., Narducci, F., & Bux, M. (2009). A Semantic Content-Based Recommender System Integrating Folksonomies for Personalized Access. In G. Castellano, L. Jain, & A. Fanelli (Eds.), *Web Personalization in Intelligent Environments* (Vol. 229, pp. 27–47). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-02794-9_2

Penrose, R., & Todd, J. A. (n.d.). On best approximate solutions of linear matrix equations. *Mathematical Proceedings of the Cambridge Philosophical Society*, *null*(01), 17–19. doi:10.1017/S0305004100030929

Robertson, S. (2005). How Okapi Came to TREC. In E. M. Voorhees & D. K. Harman, *TREC: Experiment in Information Retrieval* (pp. 287–300). MIT Press.

Salton, G., Wong, A. K. C., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Commun. ACM*, *18(11)*, 613–620.

Schein, A., Pennock, D., & Ungar. (2002). Methods and metrics for cold-start recommendations. In *SIGIR* (pp 253-260). New York, NY, USA: ACM

Sen, S., Harper, F. M., LaPitz, A., & Riedl, J. (2007). The quest for quality tags. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (pp. 361–370). New York, NY, USA: ACM.

Smola, A. J., & Schölkopf, B. (1998). A Tutorial on Support Vector Regression. *Royal Holloway College, London, U.K., NeuroCOLT Tech. Rep. TR 1998-030, 1998*.

Vanegas, J. A., Caicedo, J. C., Camargo, J. E., & Ramos-Pollán, R. (2012). Bioingenium at Image. CLEF 2012: Textual and Visual Indexing for Medical Images. In *CLEF (Online Working Notes/Labs/Workshop)*. Rome, Italy.

Vig, Jesse, Sen, S., & Riedl, J. (2012). The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Transactions on Interactive Inteligent Systems*, *2*(3). (pp 13:1-13:44)

Zhang, Z.-K., Zhou, T., & Zhang, Y.-C. (2011). Tag-Aware Recommender Systems: A State-of-the-Art Survey. *Journal of Computer Science and Technology*, *26*(5), 767–777. doi:10.1007/s11390-011-0176-1

Zipf, G. K. (1950). Human behavior and the principle of least effort. *Journal of Clinical Psychology, Adisson Wesley*, *6*(3). doi:10.1002/1097-4679(195007)6:3<306::AID-JCLP2270060331>3.0.CO;2-7