

Análisis de la Influencia de las Propiedades Semánticas en los Sistemas de Recomendación

Ávila J.*; Riofrío X.*; Palacio K.**; Albán H.**; Espinoza M.***; Saquicela V.***

*Universidad de Cuenca, Facultad de Ingeniería, Escuela de Informática
e-mail: johnny.avilam@ucuenca.ec, xavier.riofriom@ucuenca.ec

**Universidad de Cuenca, Departamento de Ingeniería Eléctrica, Electrónica y Telecomunicaciones
e-mail: kenneth.palacio@ucuenca.edu.ec, humberto.alban@ucuenca.edu.ec

***Universidad de Cuenca, Departamento de Ciencias de la Computación
e-mail: mauricio.espinoza@ucuenca.edu.ec, victor.saquicela@ucuenca.edu.ec

Resumen: Este artículo propone un procedimiento para la evaluación del impacto que tiene la inclusión de determinadas propiedades semánticas y sus combinaciones en la estimación de recomendaciones de contenidos audiovisuales, generadas por un sistema de recomendación semántico para usuarios en el dominio de la Televisión Digital. Este trabajo parte de la hipótesis de que el incremento moderado del número de propiedades involucradas en el cálculo de las predicciones mejora paulatinamente su precisión y que cada propiedad semántica tiene una influencia específica. Los resultados experimentales demuestran que el uso de diferentes combinaciones de propiedades semánticas, tiende a reducir el error promedio en distinta proporción.

Palabras clave: Sistemas de recomendación, Web semántica, Televisión digital, Ontologías, Conceptos, Propiedades Semánticas

Abstract: This paper presents an evaluation procedure focused on analyzing the impact of the inclusion of determined semantic properties and their combinations, on the audiovisual content recommendation estimation performed by a semantic recommender system to users in the domain of Digital Television. This work is based on the assumption that by increasing the number of semantic properties involved in the prediction estimation, its accuracy will be improved given that each semantic property has a determined influence. The obtained results show that the use of different semantic properties combinations leads, in general, to a specific reduction of the average estimation error.

Keywords: Recommender systems, Semantic web, Digital television, Ontologies, Concepts, Semantic Properties

1. INTRODUCCION

El cambio de era protagonizado por el advenimiento de la Televisión Digital, evidencia las profundas transformaciones que esta nueva tecnología incorpora a la experiencia del usuario, ya que además de entretenerse con los contenidos que se difunden por las estaciones de televisión, éste tiene la posibilidad de interactuar con este medio de comunicación a través de diferentes aplicaciones y modalidades, revolucionando así, el modelo tradicional de mirar televisión. Actualmente, los estándares de transmisión de TV Digital adoptados a nivel mundial permiten el transporte ágil de grandes volúmenes de información, lo que propicia la disponibilidad de excesiva oferta televisiva en los hogares. Esta sobreoferta de alternativas de entretenimiento originará un escenario en el que el usuario se enfrenta a una amplísima cantidad de programación disponible, que a pesar

de contar con una extensa diversidad y entornos de interactividad, podría llegar a abrumar incluso al televidente más entusiasta. Estas condiciones pueden originar que la experiencia televisiva se vea degradada paulatinamente y evolucione en una tediosa e interminable búsqueda de programas de entre cientos de opciones. Este contexto enmarca un evidente problema que ha favorecido la aplicación de Sistemas de Recomendación de Contenido en el campo de la Televisión Digital, que surgen como un servicio que persigue asistir al usuario en la búsqueda y selección de la programación que mejor se ajuste a sus gustos e intereses [1].

En este artículo se parte de una breve reseña sobre los sistemas de recomendación, presentada en la sección 2, en la que se menciona su origen y características principales, para posteriormente, estudiar los diferentes enfoques que se han propuesto para generar recomendaciones de ítems para distintos usuarios según sus preferencias particulares, como se describe en [22]. A partir de estos enfoques, en la literatura se ha establecido la clasificación de los sistemas de recomendación en dos grandes tipos, como se describe en la

Artículo recibido el 15 de Abril de 2014; revisado 7 de Septiembre de 2014. Este artículo forma parte del proyecto: Aplicación de Tecnologías Semánticas para disminuir la sobrecarga de información en usuarios de TV Digital, financiado por la Dirección de Investigación de la Universidad de Cuenca DIUC. Autor para correspondencia: kenneth.palacio@ucuenca.edu.ec, 074051000 ext. 2311, Universidad de Cuenca, Ecuador.

sección 2.1. Particularmente, en la sección 2.2 se analiza los sistemas de recomendación semánticos con el objetivo de establecer la necesidad de creación de un entorno de evaluación de la influencia de las relaciones entre conceptos o propiedades semánticas utilizadas en este tipo de algoritmos de recomendación y con ello la pertinencia de esta propuesta. En este sentido, a través de este trabajo se pretende analizar cómo el uso de diferentes propiedades semánticas influye sobre las recomendaciones generadas por un SRS. En la sección 3, se describe el procedimiento planteado, en el que se extiende el trabajo realizado en [20] para crear un mecanismo que posibilite la evaluación individual y colectiva de cada propiedad semántica y la cuantificación de su influencia sobre la reducción del error promedio en las recomendaciones generadas por el sistema. Los resultados obtenidos se presentan en la sección 4, mientras que la sección 5 muestra las conclusiones y planteamientos a futuro que han surgido de los resultados obtenidos del presente artículo.

2. SISTEMAS DE RECOMENDACIÓN

Como se mencionó en la sección 1, una vez establecida la necesidad de contar con sistemas de recomendación enmarcados en el ámbito de la televisión digital, es pertinente mencionar que éstos sistemas se originaron a inicios de los 90, cuando algunas bibliotecas norteamericanas emprendieron un mecanismo de generación de servicios de difusión selectiva de contenido. Esta selección se realizaba en base a segmentos diferenciados por valores, preferencias o atributos, y de acuerdo al perfil de cada usuario suscrito al servicio [2]. De aquí en adelante, su uso en otros dominios se ha ido ampliando paulatinamente. El surgimiento de la web y un manejo de la información con características muy similares al empleado en las bibliotecas, propició el surgimiento de sistemas de filtrado de información, como herramientas que analizan los recursos disponibles en la web, generalmente en formato XML (lenguaje de marcas extensible) y que facilitan a los usuarios la recuperación de información exclusivamente de su interés. Los primeros sistemas de recomendación evolucionaron de los grupos de noticias o newsgroups y su uso se ha convertido en una necesidad que se robustece con el transcurso del tiempo y el incremento de las fuentes de información disponibles a nivel global. Un ejemplo claro de ello se evidencia en las grandes corporaciones, que emplean sistemas de recomendación como instrumentos de mejoramiento de sus mecanismos de difusión aplicados en múltiples campos y contextos, con el objetivo de personalizar la información transmitida a sus clientes y así conservar su interés en un tipo de programación específica. Los autores Luigi Ceccaroni y Xavier Verdaguere [3] describen un sistema de recomendación basado en contenido, con un modelo de dominio con semántica formal de televisión digital orientado principalmente a televidentes y a usuarios de computadores personales y móviles en el que se analiza la motivación del empleo de estos sistemas desde la perspectiva comercial.

Hoy en día, los sistemas de recomendación son ampliamente utilizados en múltiples disciplinas, incluyendo ámbitos relacionados al entretenimiento, el turismo y hasta el comercio electrónico. Un ejemplo particular es el caso de Amazon1, que emplea un SISTEMA DE RECOMENDACIÓN-SDR- basado en el historial de compras, búsquedas y navegación de cada usuario. Esta es una herramienta fundamental para los proveedores en línea, como se menciona en [4].

Un sistema de recomendación, busca pre-seleccionar o inferir el interés que un usuario podría tener en algún ítem en particular, para lo cual, predicen la valoración o calificación que un usuario otorgaría a dicho ítem antes de que el usuario lo haga [1]. Esto le permite crear listas ordenadas de preferencias basadas en los intereses del propio usuario, que le son presentadas a manera de recomendaciones a través de una plataforma de comunicación.

2.1 Tipos de Sistemas de Recomendación

Paul Kantor [1] distingue los sistemas de recomendación tanto en aplicaciones prácticas como en la literatura según dos tipos fundamentales:

- Sistemas de Recomendación de Filtrado Colaborativo o sociales.
- Sistemas de Recomendación basados en Contenido.

Adicionalmente, la literatura suele contemplar la sinergia de dos de estas disciplinas para definir enfoques híbridos que combinan las mejores características de cada tipo de sistema, según se describe en [5].

2.1.1 Sistemas de Recomendación de Filtrado Colaborativo

Para realizar las recomendaciones, estos sistemas buscan relaciones de similitud agrupando a usuarios con preferencias similares, y así deducir sus preferencias [11, 5]. La importancia de un ítem es calculada basándose en calificaciones provistas por otros miembros pertenecientes al grupo del usuario activo, por ende, utilizando información de usuarios con características similares y dándole muy poca importancia al contenido de un ítem como tal [6]. Los ítems con valoraciones altas serán recomendados a usuarios similares también conocidos como vecinos cercanos. Este es el enfoque conocido como filtrado colaborativo basado en usuarios: user-based collaborative filtering. Dentro de este tipo de sistemas, también se cuenta con el enfoque basado en los ítems: item-based collaborative filtering, en el que el filtrado colaborativo se basa en las similitudes entre los patrones de clasificación de elementos, y se recomienda al usuario ítems similares a aquellos sobre los que demostró interés en el pasado [12].

Limitaciones

1. Problema de las valoraciones dispersas:

También conocido como *sparsity problem*, ocurre cuando no se tiene información o ésta es insuficiente para poder estimar las relaciones necesarias entre ítems o usuarios en los sistemas de filtrado colaborativo. Este problema suele presentarse en forma de ítems que rara vez son recomendados a los usuarios [6].

2. **Arranque en frío (nuevos usuarios y nuevos ítems):**
Cuando se registra un nuevo usuario, éste tiene que calificar muchos ítems antes de que el sistema pueda hacer recomendaciones acertadas, al principio el usuario recibirá recomendaciones pobres debido a la poca valoración de los ítems [5] [7]. Para identificar este problema se utiliza el término *cold start*.
3. **Problema de la oveja gris:**
Que se caracteriza por la dificultad de proveer buenas recomendaciones para personas que estén en la frontera de dos tipos de preferencias o que tengan gustos muy dispersos, por lo que no se pueden beneficiar del filtrado colaborativo. En términos de los sistemas de recomendación estos usuarios reciben el apelativo de *grey sheep*.
4. **Efecto portafolio (Poca diversidad):**
Esta limitación se conoce globalmente como *portfolio effect*. En este caso el sistema sólo recomendará los ítems más populares. Por lo tanto, los ítems nuevos que ingresen tendrán muchos problemas para ser recomendados, así sea la mejor opción para el usuario. En [19], se establece que este problema conlleva que los mismos ítems o ítems idénticos se recomiendan aun mismo usuario.

2.1.2 Sistemas de Recomendación Basados en el Contenido

Estos sistemas clasifican los ítems de acuerdo a sus características y a las preferencias de cada usuario, tomando en cuenta las preferencias de otros usuarios del sistema. Por ejemplo: un usuario que denote interés por películas de acción tiene una probabilidad mayor de recibir recomendaciones de películas dentro de ese mismo género.

Para la evaluación de un ítem, el sistema se basa en calificaciones anteriores realizadas por el usuario sobre ítems con características similares, cada una de ellas responde a un esquema de ponderación y tiene asignada un peso respectivo, proporcional a su importancia relativa. Para la descripción de un ítem generalmente se usan etiquetas (*tags*) o palabras claves. En la estimación de la predicción en la valoración de un ítem, los sistemas de recomendación basados en el contenido usan mayoritariamente modelos heurísticos, técnicas de aprendizaje de máquina (*machine learning*), estadística, modelos de aprendizaje de redes neuronales [8].

Limitaciones

1. **Análisis restringido de contenidos:**

Los sistemas de recomendación basados en contenidos están restringidos a las características explícitamente asociadas con los ítems a ser recomendados, por lo tanto, la efectividad de estos algoritmos depende de que tan bien estén descritos los ítems en el sistema. Es decir, que para lograr una buena recomendación es necesario tener muchos metadatos sobre los ítems [9].

2. **Sobre-especialización del contenido:**
Estos sistemas únicamente recomiendan ítems similares a los ya valorados por el usuario, así por ejemplo un usuario sin experiencia en películas de drama, nunca obtendrá una recomendación de una película de drama [10]. Este problema se conoce como *over specialization*.
3. **Arranque en frío (problemas con los nuevos usuarios):**
Se produce un caso muy similar a aquel de los sistemas colaborativos. En este caso, se requiere de un mínimo número de ítems valorados para que el sistema pueda establecer relaciones entre ítems.

2.2 Sistemas de Recomendación Semánticos

Los sistemas de recomendación semánticos (SRS) se enmarcan dentro de los sistemas de recomendación basados en contenido, y como se menciona en [12], basan su desempeño en una base de conocimiento definida a través de un mapa conceptual, una ontología –una conceptualización formal de un dominio determinado– [24].

Comúnmente, los SRS requieren del modelamiento de los perfiles de usuario para obtener información y crear datos que serán usados para la recomendación, como se menciona en [23]. Así, los perfiles de usuario se convierten en un pilar fundamental de un SRS con el fin de obtener un filtrado más preciso. Con este objetivo en mente, los perfiles de usuario son enriquecidos mediante herramientas de la Web Semántica y el uso de ontologías como un riguroso esquema conceptual basado en cierto tipo de etiquetas, dentro de uno o varios dominios; esto, con la finalidad específica de facilitar la comunicación de información entre diferentes sistemas y entidades. Existen varios lenguajes ontológicos que se emplean en el desarrollo de estos sistemas, entre los que se pueden mencionar: Marco de Descripción de Recursos (*Resource Description Framework - RDF*) y Lenguaje de Ontologías Web (*Web Ontology Language - OWL*). Adicionalmente en el dominio de la televisión digital, las guías de programación televisiva pueden también enriquecerse a través de procedimientos como el propuesto en [21] y alimentar algoritmos de recomendación de inferencias semánticas.

Alrededor del mundo se han desarrollado múltiples enfoques y plataformas de recomendación basados en tecnologías semánticas, entre los cuales se puede mencionar el Modelo de Wang y Kong [13], que es un sistema de recomendación

que utiliza un modelo ontológico para calcular la correlación entre ítems. En [25] se presenta un algoritmo muy similar al utilizado en [13], en el que se destaca el uso de dos ontologías: una para describir los ítems a recomendar y otra que describe el perfil del usuario, con el objetivo de inferir información implícita que ayude a obtener recomendaciones más precisas. Por otra parte cabe mencionar el sistema de recomendación desarrollado por Victor Codina [14], quien propone una arquitectura que mantiene el mecanismo de recomendación separado del dominio de la aplicación para así lograr independencia del dominio. Con este enfoque, la aplicación en un dominio específico presenta un listado de ítems que deben personalizarse y que se anotan en base a una estructura conceptual jerárquica de la ontología del dominio, que al mismo tiempo se comparte con el mecanismo de recomendación. Esto le permite recolectar información del usuario mientras éste interactúa con aplicaciones web.

Un sistema de recomendación basado en tecnologías semánticas ampliamente analizado en el estado del arte, es AVATAR, desarrollado por Yolanda Blanco Fernández y otros, documentado ampliamente en [15], [16] y [17]. Este modelo, recomienda al usuario ítems semánticamente relacionados a aquellos que fueron de su agrado en el pasado. Entre las principales características de esta implementación se pueden citar, como se describe en [17]:

- Explora la base de conocimiento buscando las relaciones escondidas entre las clases e instancias definidas más allá de las métricas de similitud tradicionales de naturaleza sintáctica.
- Emplea una metodología que asegura la relevancia de las relaciones encontradas acorde al perfil del usuario, asegurando su evolución en el tiempo y así garantizando la diversidad de las recomendaciones.
- Es independiente del dominio de la aplicación.

Finalmente, se analiza brevemente el SRS desarrollado por Sotelo R., Juayek M y Scuoteguazza A., cuya documentación se encuentra en [12, 20] y cuyo algoritmo ha sido utilizado como base para la implementación del SRS de este proyecto, y que se ha extendido agregándole funciones que permiten la evaluación de la influencia de las propiedades semánticas. En su propuesta, los autores implementan un SRS que sigue los lineamientos del sistema AVATAR [15] y en [12] comparan los resultados obtenidos a través del análisis semántico con aquellos obtenidos a través de sistemas de filtrado colaborativo basados en usuarios e ítems, concluyendo que el modelo semántico provee mejores resultados, particularmente para ítems con calificaciones entre 4 y 5 en la escala 1-5.

Este sistema se basa en la deducción de asociaciones semánticas resultantes del análisis de las propiedades formalizadas en la ontología considerada, identificando así, secuencias de propiedades establecidas como clases que se interconectan a través de propiedades comunes. Por otra parte,

las preferencias de cada usuario se modelan empleando una ontología que incluye las películas que el usuario ha visto, su descripción semántica y la calificación otorgada, nombrada como Nivel de Interés (Degree of Interest - DOI) para este contexto. El DOI asignado a un ítem, varía entre 1 y 5 para reflejar un menor o mayor interés en una película respectivamente. Dado que las características semánticas de una película, permiten por su parte encontrar relaciones entre diferentes ítems, ésta implementación incorpora un mecanismo de propagación del DOI de un ítem particular hacia sus propiedades semánticas (este mecanismo de propagación es el mismo que se ha propuesto en AVATAR), permitiendo así obtener el DOI de nuevos ítems a partir de los DOI de sus propiedades semánticas compartidas con ítems previamente evaluados por el usuario.

1. PROCEDIMIENTO DE EVALUACIÓN

El sistema propuesto, apunta a crear un entorno para la evaluación del impacto de la inclusión de diversas propiedades semánticas y sus combinaciones, en la estimación de las predicciones generadas por un sistema de recomendación semántico en el contexto de la televisión digital. Para el efecto, y como se mencionó en la sección 2.2, se ha utilizado como algoritmo núcleo el sistema de recomendación semántico desarrollado por Sotelo, Juayek, y Scuoteguazza [12, 20].

3.1 Preparación del Algoritmo

El algoritmo base del proyecto, se modificó para que sea capaz de recibir como entrada una cadena de texto conteniendo las propiedades semánticas involucradas en la estimación de la predicción. Las propiedades de entrada para las cuales se adaptó el algoritmo son: Género (Genre), Actor (Actor), Escritor (Writer), Director (Director) y adicionalmente (por las razones expuestas en la sección 3.3), la calificación de la película obtenida en IMDB (Internet Movie Data Base).

El algoritmo analiza las propiedades ingresadas y calcula la predicción para la película buscando solamente el DOI del concepto deseado utilizando la ontología de los usuarios. Un ejemplo de cadena de entrada puede ser el siguiente: "Genre, Actor, Director, Writer, IMDB", para la cual, el sistema analiza e identifica las propiedades ingresadas para encontrarlas en la ontología del usuario analizado. Si por ejemplo la cadena indica la presencia de la propiedad "Actor" la estimación del DOI se realiza para todos los actores de la película que se requiere calificar y el algoritmo devuelve el valor promedio de todos los DOI encontrados.

Todas las propiedades semánticas son procesadas de la misma manera, se emplea el DOI de cada una para calcular un promedio y se obtiene la calificación deseada para la película. El modelo matemático que describe detalladamente esta operación se presenta en la sección 3.4.

3.2 Arquitectura del Sistema

La figura 1 muestra la arquitectura del sistema de recomendación semántico (SRS) propuesto; en ella se distingue el algoritmo núcleo (core algorithm: desarrollado en [12] y cuyas adaptaciones se describen en la sección 3.1) como elemento central de una arquitectura modular a la que se han agregado una serie de componentes orientados a mejorar la precisión de las predicciones generadas por el algoritmo y al mismo tiempo, establecer una plataforma flexible de evaluación. Además puede observarse que el sistema recibe como entradas principales, las ontologías del usuario y de películas (ítems en general y orientados al contexto de televisión digital) para producir una recomendación a la salida, representada como el DOI del usuario en un ítem específico.

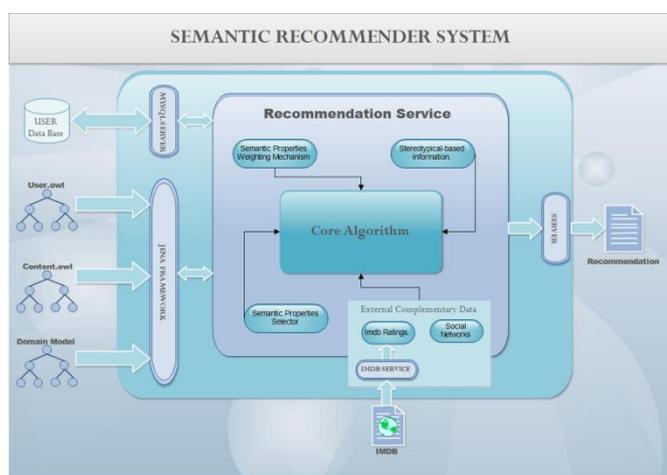


Figura 1. Arquitectura del Sistema

El SRS incluye cuatro módulos que incorporan funciones e información adicionales al algoritmo núcleo:

1. Información de estereotipos: *Stereotypical-based Information*
2. Mecanismo de ponderación para cada propiedad semántica: *Semantic Properties weighting mechanism*
3. Selector de propiedades semánticas: *Semantic Properties selector*
4. Componentes externos no-semánticos: *External complementary data:*

Las funciones de cada uno de estos módulos se detallan en las siguientes secciones.

3.3 Módulos del Sistema

3.3.1 Información de Estereotipos

Este módulo busca solventar el problema del Arranque en frío expuesto en la sección 2.1.2, en el que los usuarios recientes del sistema o aquellos que cuentan con un número bajo de calificaciones se enmarcan dentro de un “estereotipo”, para el cual, se generan recomendaciones dirigidas a todos los miembros del grupo en función de sus características comunes reales tales como su género y rango de

edad. El módulo contempla catorce estereotipos diferentes: un grupo para cada género, y éstos a su vez, se subdividen en siete rangos de edades (menor a 18, 18-24, 25-34, 35-44, 45-49, 50-55, 56+). Este esquema de estereotipo busca evaluar el comportamiento del SRS tras la suposición que individuos del mismo género y con edades cercanas tienen preferencias televisivas relativamente similares.

3.3.2 Mecanismo de Asignación de Pesos para cada Propiedad Semántica

En la obtención de la predicción de un ítem, el resultado de cada propiedad semántica (Género, Actores, Directores, etc.) se ajusta mediante una determinada ponderación, orientada a incorporar un mecanismo de calibración y evaluación que permita determinar la relevancia de las propiedades en el cálculo de la predicción.

3.3.3 Selector de Propiedades Semánticas

Los contenidos audiovisuales pueden contener tantas propiedades como se desee, sin embargo, el utilizar todas las propiedades disponibles en un SRS no necesariamente causará que las predicciones obtenidas sean más precisas incluso, podrían reflejar una degradación de los resultados. Con el objetivo de analizar la influencia de la inclusión de las propiedades semánticas más comunes presentes en este tipo de contenidos en el SRS, se propone la incorporación de un módulo que permita elegir desde hasta n propiedades semánticas como entradas del algoritmo núcleo y así, obtener resultados con cualquier tipo de combinación que se desee evaluar.

3.3.4 Componentes Externos No-semánticos

Para la incorporación de este módulo, se parte de la premisa de que una predicción obtenida mediante un SRS, puede mejorarse mediante información externa no necesariamente de naturaleza semántica. Para ello, se plantea incrementar la precisión de las predicciones generadas por el SRS al combinar sus resultados con información externa, tal como aquella extraída de la base de datos IMDB (Internet Movie Data Base) u otras fuentes similares. Por ejemplo, una película catalogada en IMDB generalmente tiene una calificación promedio o (average rating) obtenida de entre las valoraciones otorgadas por miles de usuarios registrados, por lo que puede asumirse como un indicativo estable que potencialmente, aportaría positivamente a la precisión del sistema de recomendación.

El SRS se desarrolló en JAVA, el cual está basado en la programación orientada a objetos. Para la gestión de datos se utilizó el Sistema Gestor de Base de Datos (SGBD) de uso libre MYSQL. Para la interacción con la web semántica se utilizó JENA SEMANTIC WEB FRAMEWORK, que cuenta con librerías para java y herramientas para interactuar con los datos de la web semántica (RDF, OWL, SPARQL) bajo una licencia de software libre APACHE LICENSE.

3.4 Cálculo de las Predicciones

La ecuación (1) refleja el modelo matemático que describe el cálculo de la predicción P dado un usuario u para una película m: P(u,m)

$$P(u, m) = \frac{1}{n} \cdot \sum_{g=1}^K \frac{DOI(Generos)_{u,m} (n-1-b)}{K} + \frac{\sum_{p=1}^L \frac{DOI(Propiedad)_{u,m}}{L}}{n} + \frac{1}{n} \cdot b \cdot \left[\frac{R_{IMDB}(m)}{2} \right] \quad (1)$$

Donde:

- n representa el número total de las propiedades usadas en el algoritmo, donde únicamente se toma en cuenta las propiedades que existen, es decir aquellas con un DOI ≠ 0.
- K corresponde al número de géneros asignados al ítem analizado.
- L representa el número de propiedades semánticas con un DOI disponibles para el ítem analizado.
- RIMDB representa la calificación del ítem obtenida desde IMDB.
- b corresponde a una constante relacionada a la existencia de una calificación para el ítem en IMDB: b = 1, si el ítem tiene una calificación en IMDB; b = 0, en caso contrario.

La ecuación se divide en tres términos: el primero corresponde al promedio del DOI de la propiedad géneros, ya que cada ítem puede tener uno o más (K) valores de género, por ejemplo un programa o película puede simultáneamente pertenecer a romance y comedia. El segundo término equivale al promedio de los DOI de todas las propiedades restantes que tenga la película (Actor, Director, etc.); como se mencionó anteriormente si alguna propiedad tiene un DOI = 0, la contribución de dicha propiedad es nula. Finalmente, el tercer término corresponde a la valoración o rating de la película m obtenido desde IMDB: RIMDB(m); factor que se divide para dos ya que IMDB maneja una escala de 1- 10, mientras que el SRS propuesto lo hace de 1 - 5. Los tres términos incluyen un factor del tipo ln que posibilita distribuir uniformemente la aportación de cada propiedad semántica y no semántica al cálculo de la calificación. En el segundo término, el numerador (n - 1 - b) permite obtener el aporte de las propiedades semánticas diferentes del género y de la valoración obtenida mediante IMDB (si es que existe: lo que queda determinado por la constante b).

4. EXPERIMENTACIÓN Y RESULTADOS PRELIMINARES

En esta sección se detallan los experimentos realizados y la naturaleza de los datos involucrados en cada uno.

4.1 Origen de los Datos

Para la realización de las pruebas se utilizó la base de datos de MovieLens [18], que contiene 6040 usuarios,

3952 películas y más de un millón de ratings, en la que cada usuario tiene al menos 20 películas calificadas, llegando a haber usuarios con un historial de calificaciones de más de 1000 películas. La base de datos consta de 3 archivos, el primero contiene la información concerniente a las 3952 películas, cada película contiene un número, un código de película, su título y sus géneros (acción, comedia, drama, etc.). El segundo archivo contiene la información de los usuarios, cada usuario está descrito por un código, su sexo, su edad, su ocupación y un código zip; para este estudio se tomará en cuenta únicamente el código del usuario. Las demás características serán tomadas en cuenta para experimentos futuros de clasificación y recomendación por estereotipos. El tercer archivo contiene la información de las calificaciones de los usuarios sobre las películas, por cada registro se tiene el código del usuario, el número de la película y la calificación que obtuvo esa película por dicho usuario. Todas las calificaciones están en un rango de 1 a 5 siendo 5 la máxima calificación.

4.2 Preparación de los Datos

En primer lugar se seleccionó un grupo de usuarios con el mayor número de calificaciones, asegurando que al menos 100 usuarios igualen o superen ese número; el número base encontrado fue de 825 calificaciones. Finalmente se obtuvo un conjunto de 102 usuarios, cada uno con más de 825 películas calificadas. Se identificaron dos conjuntos disjuntos de datos: un conjunto de entrenamiento y otro de prueba. Por cada usuario, del conjunto de películas calificadas, se eligieron películas para cada set en base a dos criterios: 1) las películas del set de prueba no pueden estar presentes en el set de entrenamiento y 2) las películas del set de prueba deben tener al menos una propiedad semántica en común (a manera de una función heurística) con las películas del set de entrenamiento. Estas dos condiciones aseguran que en el momento de realizar la prueba no ingresen datos no vinculados con las preferencias del usuario representadas en la ontología, y además, que no ingresen datos que se hayan sometido al entrenamiento y que puedan afectar de una manera u otra en los resultados.

Finalizado este proceso, se obtuvieron dos archivos: uno destinado para entrenamiento, que contiene 102 usuarios con un promedio de 580 películas calificadas cada uno, y otro archivo para test, con los mismos usuarios pero con un promedio de 255 películas calificadas por cada uno. Ninguna combinación de usuario-película-calificación están repetidas dentro de los archivos ni incluidas en ambos.

4.3 Creación de las Ontologías de Usuarios

A partir del archivo de usuarios, se utiliza el algoritmo original de [12, 20] para el entrenamiento y la creación de las ontologías que describen los intereses del usuario. Dentro de cada archivo se puede encontrar en formato RDF, la descripción semántica de todos los recursos que son de interés para el usuario, como actores, directores, escritores o las películas completas. Cada recurso

cuenta con su respectivo DOI el cual ha sido obtenido mediante la propagación de las calificaciones de las películas. El sistema obtiene la información semántica correspondiente a cada película que el usuario calificó y la anota en una ontología preliminar, agregando como una propiedad adicional la calificación que el usuario otorgó para esa película. Con esta ontología, se procede a la propagación del DOI según se describe en [15], fase en la que el sistema completa la ontología incluyendo el nivel de interés para cada concepto (Actores, Directores, Escritores, Géneros). El DOI es calculado a partir de la calificación que el usuario asignó a la película.

4.4 Experimentación

En esta etapa, se introdujeron al sistema 8 combinaciones diferentes de propiedades semánticas en igual número de pruebas, la propiedad Género (Genre) se incluye en todas las pruebas ya que es la única propiedad que tiene un DOI en las ontologías de todos los usuarios. Los resultados obtenidos se exponen a continuación:

Al introducir la cadena "Genre", el sistema busca únicamente la propiedad del género en las ontologías. Así, se realizaron pruebas con las combinaciones de Género, Actor; Género, IMDB; Género, IMDB, Actor; Género, Actor, Director; Género, Actor, Director, Escritor; Género, Actor, Escritor; Género, Actor, Director, Escritor, IMDB. La figura 2 muestra los resultados para cada una de los experimentos realizados, concretamente graficando el error promedio por todos los usuarios o MAE (siglas en inglés de Mean Average Error). Este error mide la diferencia entre la calificación obtenida por el sistema y aquella registrada por cada usuario en el set de prueba, según se define en la ecuación (2):

$$MAE = \frac{1}{n} \cdot \sum_{u,i} |P_{u,i} - R_{u,i}| \quad (2)$$

Donde:

- U corresponde al conjunto de los usuarios examinados.
- I corresponde al conjunto de ítems evaluados.
- N corresponde al total de ratings comparados
- P corresponde a la predicción obtenida por el sistema.
- R corresponde a la calificación otorgada por el usuario.
- En la figura 2 se muestra un gráfico del MAE por cada usuario, en cada una de las pruebas efectuadas.

4.5 Análisis de Resultados

Como se puede observar en la figura 3, la prueba que involucra únicamente la propiedad género, refleja un error promedio del 21.6% en la estimación de las predicciones, que en una escala del 1 al 5 representa un error promedio de 1.08 puntos; Al introducir la propiedad actor, el error se reduce en

un aproximado del 4% llegando así a obtener un error del 17.5 %, que representa 0.87 puntos en la escala del 1 al 5. A medida que se siguen agregando propiedades semánticas, el error tiende a reducir aunque en menor proporción.

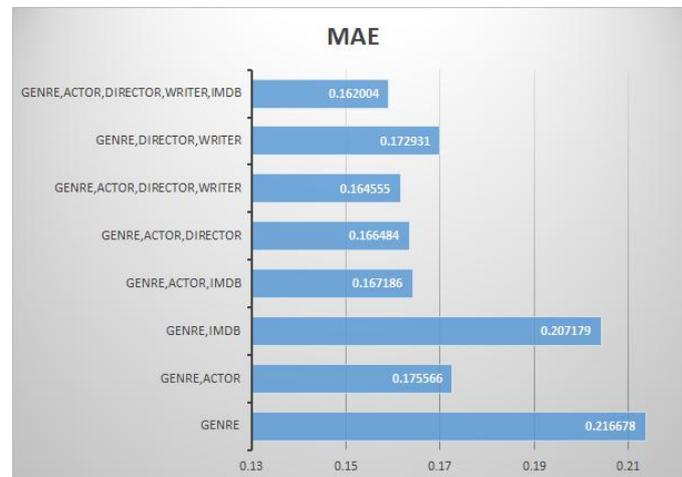


Figura 2. Promedio de error de todos los usuarios en cada una de las pruebas.

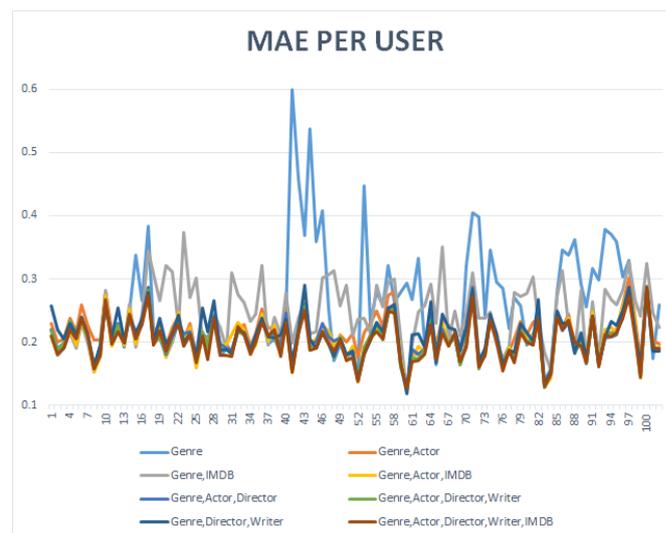


Figura 3. Error promedio de cada usuario en cada una de las pruebas.

Se puede observar también que cuando se utiliza tres propiedades semánticas, incluyendo la propiedad actor, se obtiene un error menor que al ejecutar el experimento con igual número de propiedades semánticas sin incluir la propiedad actor. Los resultados sugieren que esta propiedad aporta en mayor medida a la reducción del error que las otras propiedades del ítem analizado, sin tomar en cuenta la propiedad género ya que como se mencionó anteriormente, se utiliza para todas las pruebas. Por otra parte cuando se introduce la propiedad IMDB el error tiende a reducir ligeramente en todos los casos, aunque este cambio parece ser mínimo en comparación con la reducción originada por las otras propiedades. De ello se puede concluir que el aporte de información de carácter no-semántico en la estimación de

las predicciones, no es tan significativo como lo es el aporte de las propiedades semánticas.

Los picos y singularidades en la figura 3, reflejan casos particulares de usuarios con gustos dispersos (problema de la oveja gris, según se comentó en la sección 2).

5. CONCLUSIONES Y TRABAJO A FUTURO

Este trabajo describe la evaluación de la influencia de las propiedades semánticas involucradas en la estimación de predicciones generadas por un sistema de recomendación semántico SRS. Debido a este análisis es posible determinar qué propiedades semánticas o combinaciones de ellas influyen mayormente a mejorar la precisión de las recomendaciones, y con ello, establecer la necesidad de incluirlas como información asociada a los ítems a recomendar, que en el contexto de este proyecto corresponden a programas de televisión.

El procedimiento propuesto toma como sistema base el algoritmo de recomendación documentado en [12, 20]. Dados los resultados preliminares se puede concluir experimentalmente que un SRS genera predicciones más precisas a medida que se incrementa el número de propiedades semánticas consideradas en el algoritmo núcleo: particularmente, el menor MAE se obtiene cuando se emplea todas las propiedades disponibles (Género, Actores, Directores y Escritores) incluyendo una propiedad no semántica (ratings-IMDB). Los primeros descubrimientos muestran que al combinar la propiedad Género con cualquier otra propiedad semántica, el error tiende a disminuir alrededor de un 19 %, lo que indica preliminarmente, que puede conseguirse una mejora sustancial sí al menos se utilizan dos propiedades semánticas en el algoritmo.

Para todas las pruebas realizadas se utilizó un esquema en el que cada propiedad semántica involucrada en el cálculo de las predicciones aporta exactamente el mismo peso, logrando con ello un mecanismo que posibilita evaluar diferentes combinaciones de propiedades en el que matemáticamente todas tienen la misma importancia. Se espera que una asignación no uniforme de pesos a las propiedades semánticas utilizadas en la estimación de las predicciones generadas por el SRS, influya en la precisión que se puede obtener en el SRS. Se requiere de investigación adicional que permita determinar mecanismos automáticos de ponderación que posibiliten minimizar el MAE y encontrar la combinación óptima de coeficientes en la ecuación (1), los mismos que corresponden a los pesos asignados para el aporte de cada propiedad semántica.

El trabajo a futuro se enfocará en la implementación de mecanismos similares de selección de propiedades en otros algoritmos de recomendación presentes en el estado del arte y que utilicen tecnologías semánticas, con el fin de comparar los resultados y establecer con más precisión la influencia de cada concepto o propiedad semántica en las predicciones

generadas por un SRS sobre las calificaciones o ratings otorgadas por los usuarios a un ítem determinado.

Una futura versión del sistema, prevé incluir un enriquecimiento de perfiles de usuario por medio de las redes sociales (Facebook, Twitter, etc.) siguiendo los lineamientos descritos en [23], para incluir información adicional de carácter explícito en las ontologías de usuario tal como gustos sobre los actores, escritores, etc, que puede ponderarse con un peso mayor en la estimación de las predicciones dada su condición de información explícita provista por el usuario.

6. RECONOCIMIENTOS

Esta investigación forma parte del proyecto: “Aplicación de Tecnologías Semánticas para Disminuir la Sobrecarga de Información en Usuarios de TV digital”, financiado por la Dirección de Investigación de la Universidad de Cuenca.

REFERENCIAS

- [1] Kantor Paul B., “Recommender systems handbook”, Springer, 2009.
- [2] Foltz Peter W. and Dumais Susan T., “Personalized Information Delivery: An Analysis of Information Filtering Methods”, Communications of the ACM - Special issue on information filtering, vol. 35, no. 12, pp. 51-60, 1992.
- [3] Ceccaroni Luigi and Verdaguer Xavier, “TV FINDER: una aproximación semántica a la televisión interactiva”, BCN Media, 2003.
- [4] Schafer J. Ben, Konstan Joseph A. and Riedl John, “E-Commerce Recommendation Applications”, Ecommerce Recommendation Applications, vol. 5, no. 1-2, pp. 115-153, 2001.
- [5] Adomavicius G. and Tuzhilin A., “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions”, Knowledge and Data Engineering, IEEE Transactions, vol. 17, no. 6, pp. 734-749, 2005.
- [6] Balabanović Marko and Shoham Yoav, “Fab: content-based, collaborative recommendation”, Communications of the ACM, vol. 40, no. 3, pp. 66-72, 1997.
- [7] Yu K., Schwaighofer, A., Tresp V., Xiaowei Xu and Kriegel H., “Probabilistic memory-based collaborative filtering”, Knowledge and Data Engineering, IEEE Transactions, vol. 16, no. 1, pp. 56-69, 2004.
- [8] Belkin Nicholas J. and Croft W. Bruce, “Information filtering and information retrieval: two sides of the same coin?”, Communications on the ACM, vol. 35, no. 12, pp. 29-38, 1992.

- [9] Shardanand Upendra and Maes Pattie, "Social information filtering: algorithms for automating word of mouth", In CHI '95 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 210-217, 1995.
- [10] Sheth B. and Maes Pattie, "Evolving agents for personalized information filtering", In Proceedings of the Ninth Conference on Artificial Intelligence for Applications, pp. 345-352, 1993.
- [11] Riedl J. T., Konstan J. A. and Ekstrand M. D., "Collaborative Filtering Recommender Systems", Foundations and Trends in Human-Computer Interaction, pp. 175-243, 2011.
- [12] Sotelo, R., Juayek, M. and Scuoteguazza A., "A comparison of audiovisual content recommenders systems performance: Collaborative vs. semantic approaches", Broadband Multimedia Systems and Broadcasting (BMSB), IEEE International Symposium, pp. 1-5, 2013.
- [13] Wang R.Q. and Kong, F.S., "Semantic-Enhanced Personalized Recommender System", In International Conference on Machine Learning and Cybernetics, vol. 7, pp. 4069-4074, 2007.
- [14] Codina Victor and Ceccaroni Luigi. "Taking Advantage of Semantics in Recommendation Systems", Proceedings of the 2010 conference on Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence, pp. 163-172, 2010.
- [15] Fernandez Y.B., Pazos-Arias J.J., Nores M.L., Solla A.G. and Cabrer, M.R., "AVATAR: an improved solution for personalized TV based on semantic inference", Consumer Electronics, IEEE Transactions, vol. 52, no. 1, pp. 223-231, 2006.
- [16] Blanco-Fernández Yolanda, Pazos-Arias José J., Gil-Solla Alberto, Ramos-Cabrer Manuel, Barragáns-Martínez Belén, López-Nores Martín, García-Duque Jorge, Fernández-Vilas Ana and Díaz-Redondo Rebeca P., "AVATAR: An Advanced Multi-agent Recommender System of Personalized TV Contents by Semantic Reasoning", Lecture Notes Computer Science, no. 3306, pp. 415-421, 2004.
- [17] Blanco-Fernández Y., Pazos-Arias J.J., Gil-Solla A., Ramos-Cabrer M., López-Nores M., García-Duque J., Fernández-Vilas A., Díaz-Redondo R.P. and Bermejo-Muñoz J., "A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems", Knowledge-Based Systems, vol. 2, no. 4, pp. 305-320, 2008.
- [18] MovieLens Dataset, disponible en <http://www.movielens.org/> 2014.
- [19] Seyerlehner K., "Content-Based Music Recommender Systems: Beyond simple Frame-Level Audio Similarity", PhD thesis, Johannes Kepler Universität, Linz, December, 2010.
- [20] Juayek M. and Scuoteguazza A., "Evaluación de sistemas recomendadores de contenidos audiovisuales basados en técnicas inteligentes", Facultad de Ingeniería, Universidad de Montevideo, Uruguay, 2012.
- [21] Saquicela V., Espinoza-Mejía M, Palacio-Baus K, and Albán H., "Enriching Electronic Program Guides Using Semantic Technologies and External Resources", Proceedings of the XL Latin American Computing Conference (CLEI 2014), Oct. 2014.
- [22] Saquicela V., Espinoza-Mejía M., Mejía J. and Villazón-Terrazas B., "Reduciendo la sobrecarga de información en usuarios de televisión digital", in Proceedings of the Workshop on Semantic Web and Linked Data, 2014.
- [23] Espinoza-Mejía M., Saquicela V., Palacio-Baus K, and Albán H., "Extracción de Preferencias Televisivas desde los Perfiles de Redes Sociales", Revista Politécnica, Escuela Politécnica Nacional - EPN, vol. 34, 2014.
- [24] Gruber, T.R., "Towards principles for the design of ontologies used for knowledge sharing", In The International Journal of Human-Computer studies, 43(5/6) : pp. 907-928, 1995.
- [25] Ruiz-Montiel M. and Montes J. F. A., "Semantically Enhanced Recommender Systems", in Robert Meersman; Pilar Herrero and Tharam S. Dillon, ed., 'OTM Workshops', Springer, pp. 604-609, 2009.

