Extracción de Preferencias Televisivas desde los Perfiles de Redes Sociales

Espinoza M.*; Saquicela V.*; Palacio K.; Albán H.****

**Universidad de Cuenca, Departamento de Ciencias de la Computación, Cuenca, Ecuador*
e-mail: {mauricio.espinoza; victor.saquicela} @ucuenca.edu.ec
*** Universidad de Cuenca, Departamento de Ingeniería Eléctrica, Electrónica y Telecomunicaciones, Cuenca, Ecuador*
e-mail: {kenneth.palacio; humberto.alban} @ucuenca.edu.ec

Resumen: *El objetivo final de un sistema de recomendación es satisfacer las necesidades de información de un usuario. En el caso de la TV digital los sistemas de recomendación han mostrado ser una excelente alternativa para hacer frente a la sobrecarga de información y sugerir la selección de los programas más interesantes para ver. Sin embargo, para que esto ocurra se requiere que estos sistemas consideren en su diseño los intereses y preferencias del usuario.*

Con el fin de proporcionar un enfoque más robusto para la modelación del perfil de un usuario, se propone el uso de tecnologías de laWeb semántica junto con laWeb social. La idea de juntar ambos campos es convertir las declaraciones implícitas de las preferencias televisivas de un usuario en la Web social, a expresiones de la forma sujeto-predicado- objeto (conocidas en términos semánticos como tripletes). Esta combinación permitirá añadir significado a los datos capturados con el fin de que el perfil pueda ser aprovechado de forma sencilla por una máquina. El sistema propuesto confía en un grupo de ontologías de diferentes dominios y se basa en una arquitectura genérica que permite capturar manipular, y serializar el perfil de un usuario de TV digital.

Palabras clave: *Web Semántica, Web Social, Ontologías, Sistemas de Recomendación, Perfil de Usuario.*

Abstract: *The final goal of a recommender system is to satisfy the user's information needs. In the case of digital TV, the recommender systems have proven to be an excellent alternative to address the information overload and facilitate the selection of the most interesting programs to watch. However, for this to occur it is required that these systems consider at their design the user interests and preferences.*

In order to provide a more robust approach for modeling the user profile, we propose the use of the Semantic Web technologies together with the Social Web. The idea of merge both fields is to convert the implicit declarations of the user's television preferences on the social Web, to expressions of the shape: subject-predicate-object (known in semantic terms as triplets). This combination will add meaning to the captured data so that the profile can easily be exploited by a machine. The proposed system relies on a set of ontologies for different domains and a generic architecture to capture, manipulate, and serialize the profile of a digital TV user.

Keywords: *Semantic Web, Social Web, Ontologies, Recommendation Systems, User Profile.*

1. INTRODUCCION

Con el fin de gestionar la enorme cantidad de información disponible en la actualidad, la gente necesita apoyo específico de sistemas de recomendación. Esta situación es una necesidad en diferentes campos de aplicación como: medios digitales, comercio electrónico, aprendizaje en línea, TV digital, etc. En todos estos campos de aplicación, la información destino depende de las necesidades y preferencias del usuario. Así, uno de los factores clave en los sistemas de recomendación es la captura y representación del perfil de un usuario.

En un trabajo previo [1] se identificó que la representación formal del perfil de un usuario es uno de los elementos clave para filtrar la gran cantidad de información a la que están sometidos actualmente los usuarios de TV digital. Mientras que en [1] se describió la arquitectura y los procesos requeridos por el sistema de recomendación, este trabajo describe el proceso seguido para i) identificar qué información relativa al usuario requiere ser modelada, y ii) cómo capturar esta información desde diferentes fuentes.

El modelado del perfil conlleva un proceso de aprendizaje continuo de la información proporcionada por el usuario de

Artículo recibido el XX, 2013; revisado XX julio de 2013. (Escriba la fecha en que presentó su documento para su revisión).
Esta sección puede ser utilizada para colocar información adicional de los autores. Esta obra fue financiada en parte por la SENESCYT el marco del Proyecto BS123456 (patrocinador y reconocimiento el apoyo financiero va aquí). Los títulos de los trabajos deberán ser escritos en letras mayúsculas y minúsculas, no solo en mayúsculas. Evite escribir fórmulas largas con subíndices en el título; fórmulas cortas que identifican los elementos están bien (por ejemplo, "Nd-Fe-B"). No escriba "(invitado)" en el título. Los nombres completos de los autores se prefieren en el campo de autor, pero no son necesarios. Ponga un espacio entre las iniciales de los autores. El formato es Apellido N.
Autor para correspondencia: Dirección de Autor FA correos electrónicos, teléfono y dirección institucional.

manera implícita o explícita. Los enfoques actuales de representación del perfil (ver por ejemplo [2]) recomiendan una extracción implícita de las preferencias del usuario, derivadas de su actividad en aplicaciones disponibles en la Web y su modelamiento a un formato accesible por máquinas que permita ejecutar procesos de inferencia automáticos sobre el contenido relevante.

En este artículo se propone el uso de la Web social [3] como fuente principal para extraer las preferencias de un usuario de TV digital. El término Web social hace relación al cambio de paradigma que se ha producido en la Web, pasando de un enfoque centrado en la máquina hacia un enfoque centrado en el usuario y en la comunidad. En esta nueva cultura de participación en la Web, la gente comparte y comunica sus ideas, opiniones y preferencias a través de diferentes aplicativos (ej. Facebook[1] o Twitter[2]). La hipótesis es que éste gran depósito de datos sociales puede ser usado para extraer las preferencias televisivas de los usuarios.

Un objetivo adicional de este trabajo, es hacer uso de la Web semántica [4] para ayudar con la representación e interpretación de los datos que definen el perfil de un usuario. Este enfoque permitirá reducir el problema del arranque en frío, el cual es un problema común en la mayoría de sistemas de búsqueda y recomendación.

La solución planteada consiste de tres pasos: en primer lugar se construirá una ontología para capturar las características y preferencias televisivas de los usuarios de TV digital. En segundo lugar, se tomará ventaja del ambiente dinámico de las redes sociales y otras fuentes similares para recolectar los intereses de los usuarios en múltiples contextos. Por último, se poblará la ontología con instancias que cubren los aspectos de interés televisivo descubiertos en el paso anterior.

La contribución de este trabajo puede ser resumido en los siguientes aspectos:

- Se modela las características y preferencias de los usuarios de TV digital usando ontologías.
- Se captura el perfil de usuario desde diferentes redes sociales.
- Se transforman los datos ambiguos y no estructurados de las redes sociales a un formato semántico accesible por una máquina.

El resto de este artículo tiene la siguiente estructura. La sección 2, describe la motivación de juntar la Web semántica y social para crear un modelo general, amplio y extensible del perfil de usuario. En la sección 3, se describen algunos conceptos relacionados con el modelado de las características y preferencias televisivas de un usuario. La sección 4, identifica los componentes de la arquitectura y los procesos requeridos para capturar y serializar el perfil del usuario. En la sección 5 se describen brevemente algunos trabajos relacionados con esta propuesta. Finalmente, la sección 6 presenta las conclusiones y trabajos futuros.

[1]https://www.facebook.com/
[2]https://twitter.com/

2. MOTIVACIÓN

A pesar de que la Web semántica y la Web social son entidades muy diferentes, la vinculación de los dos dominios promete un entorno más productivo, personalizado e intuitivo de la gran cantidad de información poco estructurada y altamente subjetiva presente en las redes sociales.

El resultado de la convergencia entre los elementos de las redes sociales y las tecnologías de la Web Semántica, es una Web innovadora de conocimiento interrelacionado y semánticamente rica denominada Web Semántica Social [5]. Esta visión de la Web se compone de documentos relacionados entre sí, datos, e incluso aplicaciones creadas por los propios usuarios como consecuencia de todo tipo de interacciones sociales, y está accesible en formatos legibles por máquinas [6]. La Web Semántica Social ofrece una serie de posibilidades en términos de aumento de la automatización y la difusión de información, ofreciendo nuevas e innovadoras formas de personalizar el contenido de los sistemas de recomendación, basados en la información semántica sobre usuarios, sus intereses y relaciones con otras entidades.

De hecho, algunos estudios [7, 8] muestran que las actividades efectuadas por un individuo dentro de una red social llevan consigo información interesante sobre sus intereses y por tanto pueden desempeñar un papel vital en el desarrollo de sistemas de recomendación.

Es necesario hacer notar que incluso la recolección de la gran cantidad de información proveniente de las diferentes redes sociales no resulta un problema en la actualidad, gracias a la introducción de interfaces de programación de aplicaciones (APIs). Estas interfaces de programación ofrecen acceso a diferentes servicios de las redes sociales y representan un mecanismo de acceso fácil para otros componentes de software. Esta situación ha provocado que los investigadores de laWeb Semántica tengan acceso a un entorno donde se pueda probar futuras tecnologías semánticas y sociales con mayor facilidad.

En definitiva, la vinculación de las diferentes identidades sociales de un usuario sobre la Web y el enriquecimiento semántico de toda la información embebida en éstas, permite alcanzar modelos más ricos y dinámicos de los usuarios. Para el caso particular de esta investigación, éstos modelos pueden ayudar a mejorar los procesos de recomendación de programación televisiva.

3. MODELADO DE USUARIO USANDO ONTOLOGÍAS

En esta sección se describen algunos conceptos generales que participan en el proceso de modelado de usuario. El modelo de usuario (también llamado perfil de usuario en el contexto de recomendadores) es una estructura que contiene información sobre el usuario en un formato uniforme y comprensible por una máquina, empleado por un sistema para apoyar la entrega de contenido inteligente y proveer recomendaciones al usuario [2].

Con el objetivo de describir de manera uniforme y no ambigua la caracterización de un usuario, en este trabajo se abordará una interpretación semántica de las preferencias del

usuario. El término "semántica", se refiere al paradigma que emplea conocimiento ontológico formal como base para la construcción de descripciones de modelos de usuario semánticamente estructuradas. Una ontología formal consiste de: a) conceptos y sus propiedades (que pueden subdividirse en atributos escalares y relaciones no escalares), e instancias que representan entidades que pertenecen a los conceptos, y b) axiomas y predicados que representan las reglas específicas que representan el conocimiento del dominio [9].

Algunas cuestiones deben ser consideradas en el proceso de captura y representación semántica del perfil de un usuario: i) la(s) ontología(s) que puedan proporcionar una base de conocimiento significativa para capturar la semántica pertinente al usuario y dominio, ii) la expresividad del lenguaje ontológico más adecuado para articular este conocimiento, y iii) los recursos y medios para capturar y representar discretamente el comportamiento del usuario. En las siguientes secciones se abordan estos asuntos.

3.1 Modelos Ontológicos para Representar el Perfil de Usuario

Los requisitos para decidir sobre el modelo ontológico más adecuado para representar el perfil de los usuarios de TV digital incluyen determinar el nivel de granularidad, la precisión semántica, y la expresividad de la ontología. El nivel de granularidad en la representación viene determinado por las potenciales aplicaciones de la ontología que modele el perfil de un usuario. Para determinar el nivel de granularidad o grado de especificidad de la representación que se desea introducir en la ontología, es posible usar como técnica, las preguntas de competencia [10], que consiste en escribir preguntas en lenguaje natural que la ontología a ser construida debe ser capaz de responder.

La precisión semántica viene determinada por la forma en cómo modelar la ontología. En este sentido, el interés de los autores de este trabajo, es utilizar como técnica de modelación una propuesta *top-down*, en lugar de una *bottom-up*. Es decir, modelar el perfil de un televidente partiendo de ontologías de alto nivel que puedan ser especializadas y provean una potente base de conocimiento para capturar la semántica del dominio y usuario.

Respecto a la expresividad de la ontología, un gran número de estándares han sido propuestos por el Consorcio Word Wide Web[3] (World Wide Web Consortium – W3C) para ejecutar el procesamiento semántico de la información. De manera concisa, se puede mencionar el modelo de datos RDF[4], el cual cubre las relaciones semánticas básicas usando tripletas. RDFs[5] (RDF Schema), una extensión semántica de RDF, proporciona mecanismos para la descripción de grupos de recursos relacionados y las relaciones entre estos recursos. Para la representación formal del conocimiento, el lenguaje OWL[6] se ha convertido en la opción imperante. Actualmente, OWL tiene tres variantes (OWLLite, OWL-DL, OWL-Full) que pueden ser usadas para lograr el equilibrio deseado entre alta expresividad/completitud y complejidad.

En este trabajo, el lenguaje OWL ha sido usado para describir a ontología, con su expresividad limitada a OWLDL, a fin de mantener un razonamiento escalable sin sacrificar demasiado poder expresivo.

A pesar del gran esfuerzo desarrollado en la comunidad para la creación de ontologías que permitan modelar el perfil de un usuario de TV digital, todavía no es posible hablar de una ontología estándar para estos propósitos. La idea en esta propuesta es reutilizar el trabajo existente en la literatura para crear una ontología propia que permita describir el conocimiento relevante de un usuario de TV digital. El modelo propuesto reutiliza modelos ontológicos existentes que han sido fusionados y modificados mediante un proceso de re-ingeniería, tomando en consideración su importancia, disponibilidad y potencial de explotación. La figura 1, muestra el esquema conceptual de esta red de ontologías.

La ontología está compuesta de nueve modelos ontológicos que incluye al menos cinco categorías de datos: datos personales (basado en datos biográficos y demográficos), preferencias, actividades, tiempo y lugares. La identificación, agrupación y validación del conjunto de requisitos que debe satisfacer la ontología, está fuera del alcance de este artículo. Sin embargo, aquellos lectores interesados en la metodología usada para establecer los requerimientos y el modelo conceptual de la ontología a partir de recursos ontológicos existentes pueden consultar [11].

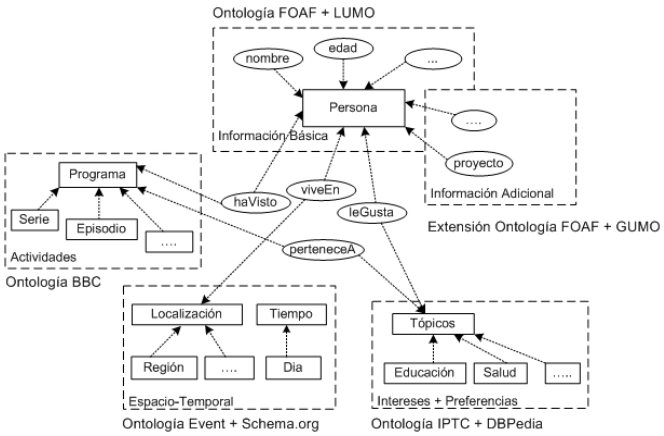


Figura 1. Metamodelo de la Ontología del Perfil de Usuario

3.2 Fuentes de Información

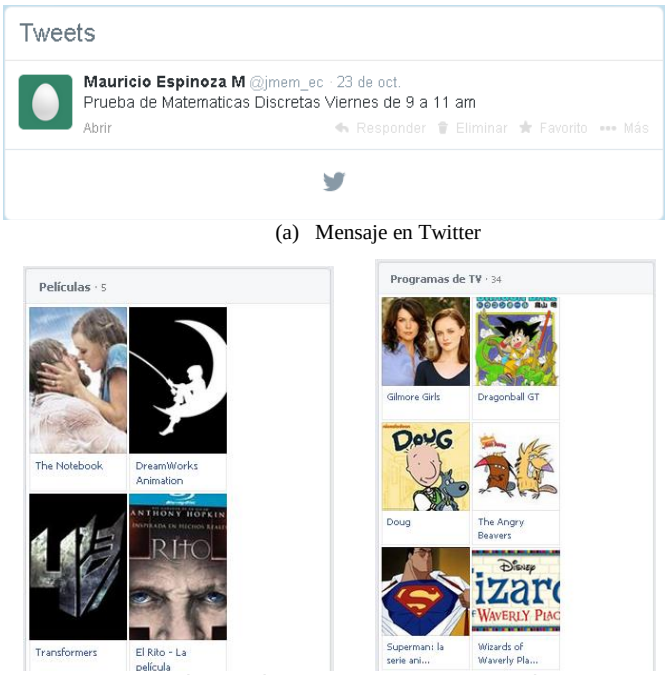
Otro de los factores a considerar en el proceso de captura y representación semántica del perfil de un usuario, es la selección de las fuentes de información más apropiadas para identificar implícitamente el comportamiento del usuario. Como ya se mencionó en la sección 1, en este trabajo se propone el uso de la Web social como medio para descubrir esta información.

La abrumadora popularidad de las redes sociales crea una oportunidad de mostrar los aspectos dados de uno mismo [12]. De hecho, la información del perfil de usuario

[3]http://www.w3.org/
[4]http://www.w3.org/RDF/
[5]http://www.w3.org/TR/rdf-schema/
[6]http://www.w3.org/2001/sw/wiki/OWL

publicado en redes sociales como Facebook o Google+ provee un rico repositorio de información personal en un formato de datos estructurado, que hace que esta información sea susceptible de ser automatizada. Aunque la información del perfil es publicada normalmente en un formato de fácil acceso, la mayoría de veces el contenido del perfil es raramente ingresado en su totalidad por los usuarios. Además, no todas las redes sociales proveen un mecanismo de acceso a la información en un formato estructurado, lo que hace más difícil obtener un perfil de usuario comprensible. A pesar de que muchos usuarios no hacen explícita la información personal en sus redes sociales, la interacción y generación de contenido en estos medios, hace que sea posible sugerir algunos tipos de atributos del usuario y sus preferencias. En la Figura 2, se muestra un ejemplo de un mensaje en Twitter e información de las secciones películas y programas en Facebook que permiten determinar las actividades de una persona, sus preferencias de películas y programas de televisión, respectivamente.

Es necesario hacer notar que aunque la información mostrada en la figura 2(a), no guarda relación directa con los hábitos de consumo televisivo de un usuario, esta información es de suma importancia para establecer posibles preferencias televisivas. En realidad, es muy probable que el creador del mensaje en Twitter esté interesado en reportajes de ciencia y tecnología que involucre el uso de diferentes algoritmos matemáticos. Para ello será necesario establecer primero que Matemáticas Discretas es un área de las Matemáticas y esta última una ciencia; todo esto es factible con el uso de tecnología semántica.



(a) Mensaje en Twitter

(b) Ejemplo de sección de Películas en Facebook

(c) Ejemplo de sección de Programas en Facebook

Figura 2. Ejemplos de Twitter y Facebook para inferir el perfil de un usuario

Para hacer uso de los servicios y funcionalidades proporcionadas por el gran número de redes sociales existentes y estar al día con los demás miembros, los usuarios tienden a crear varias cuentas en diferentes sitios. Este fenómeno requiere una recuperación efectiva de todos los datos disponibles sobre una persona de varias redes sociales y la integración de estos datos en un único perfil.

Hoy en día, existen varios sitios como Snitch Name[7] y Pipl[8], los cuales permiten la búsqueda de usuarios por nombre, entre múltiples redes sociales. Sin embargo, obtener las referencias cruzadas entre los usuarios de diferentes sitios no es una tarea fácil, ya que, en muchos casos, los usuarios de las diferentes redes sociales rellenan detalles diferentes, utilizan diferentes nombres de usuario, y tienen diferentes listas de amigos. Para complicar aún más el problema, muchos usuarios diferentes tienen nombres y datos personales similares. Por ejemplo, una simple consulta del usuario Juan Pérez en Facebook retorna cientos de perfiles. Como resultado, incluso si dos perfiles de usuario tienen el mismo nombre y apellido, no es suficiente para confirmar que estos dos perfiles de usuario pertenecen a la misma persona.

En este trabajo, se evita un proceso de identificación automático de los diferentes perfiles que pertenecen a un mismo individuo (conocido como resolución de entidades[9]), ejecutando un proceso manual de identificación.

Durante el proceso de autenticación, el usuario del sistema es redirigido a la página de confirmación de identidad en la red social, para ingresar su nombre de usuario y contraseña. Una vez que el usuario se ha autenticado, el sistema solicita los permisos respectivos para leer la información de la cuenta de usuario (perfil, lista de contactos, etc.). Este proceso elimina la necesidad de implementar un algoritmo para resolver el problema de reconocimiento de entidades para los perfiles de diferentes redes sociales. Sin embargo, todavía queda abierta la pregunta sobre qué redes sociales deben ser accedidas para descubrir la información del perfil.

Al respecto es necesario indicar que otros trabajos han estudiado previamente acerca de las redes sociales más populares en Ecuador, algunas de las más populares son: Facebook, YouTube, Hi5 y Twitter [14]. El estudio efectuado como parte de este trabajo [15] confirmó estos datos, mostrando además que los tres grupos etarios mayoritarios que hacen uso de las redes sociales en el país (ver figura 3) está conformado por las personas de entre 15 a 44 años.

Para la selección de las redes sociales a utilizar dentro del sistema, se tomó en consideración algunos aspectos como: popularidad, tipo de información que proveen, métodos de acceso, y disponibilidad de documentación. Las tres redes con más alta valoración fueron seleccionadas: Facebook, Twitter, y Google+. De las redes sociales seleccionadas, se intenta recuperar datos como el perfil, contactos, likes y actividades del usuario.

[7]http://snitch.name/

[8]https://pipl.com/

[9]La resolución de entidades (entity resolution en inglés), es un proceso que permite la identificación y vinculación/agrupación de diferentes manifestaciones de un mismo objeto del mundo real [13].

Figura 3. Distribución de Usuarios de Facebook en Ecuador por grupos de edad

4. ARQUITECTURA PROPUESTA

La arquitectura propuesta en este trabajo tiene la intención de capturar, actualizar y representar el perfil de un usuario tomando como base principal las anotaciones efectuadas en diferentes redes sociales. El modelo propuesto será usado para la recomendación de programas televisivos que puedan ser de interés para el usuario [16]. Los datos sobre el perfil de un usuario son ligados a conceptos ontológicos a través de una secuencia que comprende cuatro pasos de procesamiento (ver figura 4):

- *Extracción de Datos:* Este componente implementa un proceso de recolección de información del usuario *explícito e implícito.* El proceso explícito hace uso de formularios Web que permiten recolectar los datos básicos del usuario. Dado que no se puede asumir que la descripción sobre las preferencias televisivas serán proporcionadas de forma explícita por el usuario, entonces, dos fuentes de información han sido consideradas para explotar la información del perfil: la interacción del usuario con el televisor y las anotaciones efectuadas en distintas redes sociales.
- *Pre-Procesamiento de Anotaciones Sociales:* Para facilitar el proceso de emparejamiento de los datos no-estructurados provenientes de las redes sociales con los conceptos de la ontología, es necesario un pre-procesamiento que ejecute transformaciones morfológicas y semánticas de los datos. Este tarea se lleva a cabo en el sistema, únicamente en aquellos casos en los que los datos recuperados contienen descripciones textuales (ej. mensajes en Twitter) que necesitan ser analizadas.
- *Categorización de Entidades en Conceptos Ontológicos:* Una vez que los datos recuperados desde las diferentes fuentes de información han sido procesados, estos datos son automáticamente convertidos en instancias de conceptos en la ontología. Para ejecutar este proceso, se define un mapeo entre las etiquetas que definen el perfil del usuario y los conceptos en la ontología.
- *Enriquecimiento del Perfil Semántico:* El proceso de enriquecimiento toma como entrada un conjunto de conceptos en la ontología y aplica un algoritmo de expansión de conceptos para hacer esta información más adecuada para el sistema de recomendación.

Estos pasos son explicados con más detalle en las siguientes subsecciones.

4.1 Extracción de Datos

El modulo de extracción de datos se encarga de recopilar información relacionada con el usuario a través de dos mecanismos. El primer método usa un procedimiento explícito que solicita al usuario datos básicos como su nombre, edad, sexo, correo electrónico, ubicación y ocupación. De estos seis campos, sólo los cuatro primeros son obligatorios y tienen como fin categorizar de forma inicial a un usuario mediante el uso de estereotipos.

Los avances recientes en el desarrollo de ontologías para perfiles de usuario, abren la posibilidad de generar modelos estereotipados como instancias independientes de la ontología [2]. Es decir, cada vez que el sistema recibe el registro de un nuevo usuario, los datos ingresados explícitamente son usados para caracterizar al usuario mediante un estereotipo. Todo el modelo de usuario se actualiza con la información del estereotipo seleccionado. Esta información se mantiene, mientras no sea posible recolectar datos adicionales sobre el usuario en cuestión.

Para completar la información del perfil de usuario, otros dos mecanismos implícitos de captura han sido implementados en el sistema. El primero, captura la interacción del usuario con la televisión. Este proceso confía en una aplicación implementada en LUA[10] que tiene dos objetivos: i) obtener los datos del nombre del programa y una breve descripción del mismo, y ii) enviar la información sobre la programación vista por un usuario a un servidor remoto. Esta información es capturada por el sistema y almacenada automáticamente en la ontología. Para la ejecución de esta actividad fue necesario implementar un mapeo previo entre las entidades capturadas desde el setop-box y los conceptos que modelan las actividades de un usuario en la ontología.

El segundo mecanismo implementado para recolectar implícitamente información sobre el usuario y sus preferencias, es la conexión con redes sociales. Como se mencionó en la sección 3.2, las redes sociales Facebook, Twitter y Google+ fueron seleccionadas para adquirir este conocimiento de los usuarios. Aunque las redes sociales actuales proporcionan un acceso fácil a los datos de perfil de usuario mediante APIs dedicadas, estas funciones no proporcionan información precisa sobre los esquemas de respuesta, agravando la integración de perfiles de usuario. Para aliviar este problema, se utilizó la librería HybridAuth[11]. Esta librería, actúa como una API abstracta entre la aplicación y las APIs de varias redes sociales[12]. De esta manera, es factible acceder al perfil de usuario mediante una estructura rica, sencilla y estandarizada.

[10]http://www.lua.org/
[11]http://hybridauth.sourceforge.net/
[12]La lista completa de redes sociales y demás proveedores de información soportadas por esta herramienta, puede ser consultado en http://hybridauth.sourceforge.net/userguide.html

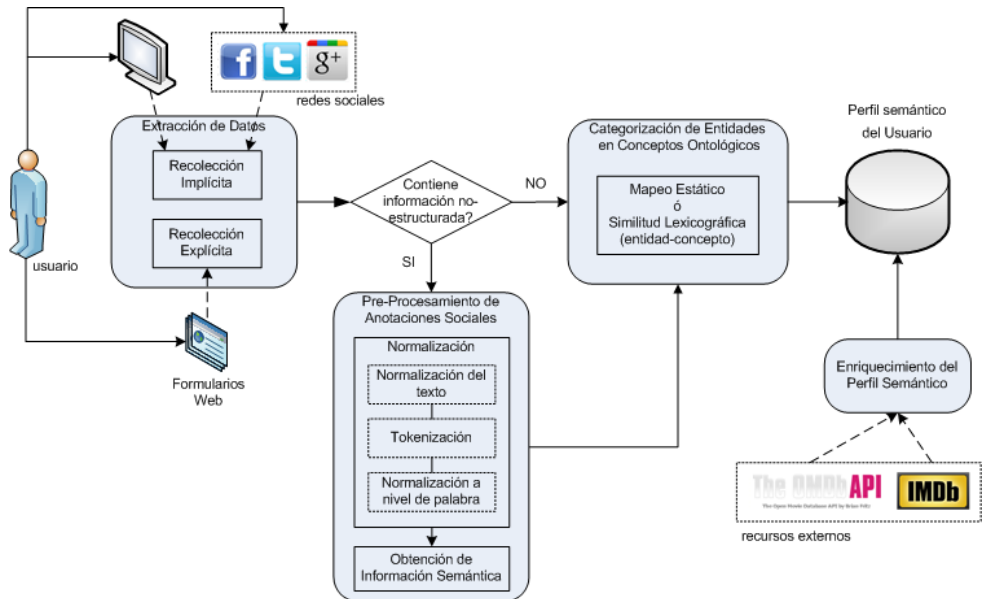


Figura 4. Arquitectura del Sistema

La Figura 5, muestra un ejemplo del acceso al perfil básico de un usuario de Facebook. Se puede observar que la información recuperada contiene datos como el nombre, ocupación, género, edad, correo y localización.

```
object(stdClass)[69]
  public '@context' =>
    object(stdClass)[68]
      public 'name' => string 'http://schema.org/name' (length=22)
      public 'occupation' => string 'http://schema.org/occupation' (length=28)
      public 'gender' => string 'http://schema.org/gender' (length=24)
      public 'age' => string 'http://schema.org/age' (length=21)
      public 'email' => string 'http://schema.org/email' (length=23)
      public 'location' => string 'http://schema.org/location' (length=26)
  public 'age' => int 24
  public 'email' => string 'abuelcore@hotmail.com' (length=21)
  public 'gender' => string 'male' (length=4)
  public 'location' => string 'Loja, Ecuador' (length=13)
  public 'name' => string 'Sebastian Roman' (length=15)
  public 'occupation' => string 'Computer Science' (length=16)
```

Figura 5. Acceso al perfil básico de Facebook usando HybridAuth

4.2 Pre-Procesamiento de Anotaciones Sociales

Algunas redes sociales como Twitter proveen importante información embebida en descripciones textuales que merecen ser consideradas. Por ejemplo, la figura 6 muestra dos ejemplos de mensajes en Twitter que podrían ofrecer información sobre las preferencias televisivas de un usuario. En este caso la etiqueta @ChampionsLeague representa la página oficial de la UEFA Champions League.

Observando el último capítulo de TheBingBangTheory

Próximo partido: CUARTOS FINAL -IDA- @ChampionsLeague:
Atlético Madrid-Barcelona
Fecha: miércoles 9 de abril
Hora: 13:45
Estadio: Vicente Calderón
Transmisión: **Fox Sports**
Fuerza Atlético

Figura 6. Ejemplos de mensajes en Twitter

Para facilitar el análisis de estos datos, la información textual tiene que ser separada en entidades y asignada a un vocabulario compartido. El proceso de filtrado usado en el sistema ejecuta un proceso secuencial donde la salida de un paso es usado como entrada al próximo. En las siguientes secciones se explica este procedimiento con más detalle:

4.2.1 Normalización

Esta fase agrupa tres componentes que limpian y separan el texto de entrada en palabras. La fase de *normalización del texto* ejecuta operaciones que permiten la eliminación de texto extraño, puntuación (por ejemplo, paréntesis, que se utiliza para marcar sinónimos o contexto de uso), o los puntos en las abreviaturas (ejemplo, I.S.B.N). La fase de *tokenización* divide un texto en palabras. Finalmente, el proceso de *normalización a nivel de palabra* reconoce y anota aquellas palabras pertenecientes a categorías especiales (horarios, fechas, números, etc), expande contracciones, reconoce y normaliza los errores tipográficos e identifica palabras compuestas.
Para la ejecución de estas actividades se han usado diferentes herramientas de procesamiento de lenguaje natural. La herramienta Tokenizer[13] para ejecutar la división de texto en palabras y las herramientas GATE[14] y Freeling[15] para la recuperación y administración de entidades.

4.2.2 Obtención de Información Semántica.

Para poblar la ontología que modela el perfil con conceptos asociados a las entidades descubiertas en las anotaciones de las redes sociales, se necesita una base de conocimiento semántica general y multi-dominio.

[13]http://nlp.stanford.edu/software/tokenizer.shtml
[14]https://gate.ac.uk/
[15]http://nlp.lsi.upc.edu/freeling/

En este trabajo, se propone extraer esta información desde Wikipedia. Los artículos en Wikipedia describen diferentes tipos de entidades: personas, lugares, empresas, etc, proporcionando descripciones, referencias e incluso fotos de las entidades descritas.
Puesto que las entidades descubiertas pueden tener diferentes significados para diferentes contextos, es necesario ejecutar un proceso de desambiguación. Por ejemplo, la entidad "ruta" puede estar refiriéndose a un mensaje en Twitter que haga referencia a las propiedades curativas de esta planta medicinal[16], o una anotación de me gusta en Facebook sobre la ruta o trayecto del spondylus en el Ecuador[17]. Una de las secciones en los artículos en Wikipedia está dedicada a ofrecer posibles opciones de desambiguación del término en cuestión. Siguiendo con el ejemplo del término "ruta", a continuación se muestra algunos ejemplos de la página de desambiguación ofrecida por Wikipedia para esta palabra:

- En algunos países hispanoaméricanos, nombre con el que se conoce a las *carreteras*, un camino de dos manos, generalmente asfaltado, para el tránsito vehicular interurbano.
- En comunicaciones, a una *ruta de enlaces*, un conjunto de puntos de comunicación que conectan dos puntos extremos.
- En botánica, al género *Ruta*, el tipo de la familia de las rutáceas.
- …

Procesando esta información y las categorías que ofrecen los artículos en Wikipedia, es posible inferir el significado correcto del término buscado. Una vez que este proceso es ejecutado para cada entidad descubierta, el sistema ejecuta el proceso de categorización de entidades a conceptos ontológicos, como se describe en la siguiente sección.

4.3 Categorización de Entidades en Conceptos Ontológicos

Para aquellas redes sociales que devuelven datos estructurados como respuesta (ej. Facebook y Google+), el sistema ejecuta el proceso de asignación de una entidad recuperada a un concepto en la ontología mediante un mapeo previamente establecido. La Tabla 1, muestra un ejemplo de las correspondencias identificadas entre el modelo de objetos devuelto por HybridAuth y los conceptos en la ontología. La información presentada en la figura corresponde a las entidades recuperadas desde el perfil básico de Facebook.

[16]"La ruda (Ruta) es un género de subarbustos siempreverdes fuertemente aromatizados de 2 a 6 metros de altura, de la familia de las Rutaceae".
[17]"La Ruta del Spondylus es una vía a lo largo de la costa de Ecuador que combina muchos de los elementos que comprenden la cultura del país".

Tabla 1. Correspondencias entre etiquetas y conceptos de la ontología

Términos de la Ontología	Etiquetas	Descripción
name	name	Nombre del usuario
profession	occupation	Ocupación del usuario
gender	gender	Género del usuario
age	age	Edad del usuario
mail	email	Correo del usuario
place_residence	location	Lugar de residenciadel usuario

En el caso de las redes sociales que ofrecen información textual de las anotaciones sociales de un usuario (ej. Twitter), se está trabajando en la implementación de un procedimiento que descubra coincidencias morfológicas entre el nombre y categorías de las entidades de entrada y los nombres de las clases en la ontología. Aquellas clases en la ontología con mayor cantidad de nombres similares al nombre y categoría de la entrada serán elegidos como los conceptos en donde se almacenarán las instancias correspondientes.
En ambos casos, para poblar los datos en la ontología se usan sentencias SPARQL. La Figura 7 muestra un ejemplo de los comandos SPARQL usados para instanciar los datos personales del perfil de un usuario con datos extraídos desde la red social Facebook.

```
PREFIX po: <http://purl.org/ontology/po/>
INSERT IN GRAPH <http://ucuenca.edu.ec/perfil/> {
  <http://ucuenca.edu.ec/perfil/mau>
     po:name "Mauricio Espinoza";
     po:profession "Ingeniero de Sistemas";
     po:gender "masculino";
     po:age 39;
     po:place_residence "Cuenca"
}
```

Figura 7. Ejemplo de comandos SPARQL para poblar la ontología del perfil de usuario.

4.4 Enriquecimiento del Perfil Semántico

El proceso de enriquecimiento de la ontología del perfil de usuario hace uso de diferentes servicios REST que permiten completar información sobre los programas vistos por el usuario [17]. Para explicar mejor el método de enriquecimiento propuesto, se plantea el siguiente ejemplo. En el proceso de extracción de datos (ver sección 4.1), el sistema ha detectado que el usuario Juan ha visto el programa Zoey en el canal de televisión Ecuavisa. La información recuperada desde la guía de programación electrónica muestra que dicha serie será emitida el martes 4 de enero de 2014 a las 15h30. Junto con esta información se muestra una breve descripción del capítulo.
Aunque esta información puede ser útil para recomendar nuevamente esta programación al usuario, sería ideal poder contar con información adicional sobre la serie, de manera

que sea posible inferir otras recomendaciones. Usando los servicios OMDb API[18] y TheTVDB.com[19] es posible recuperar datos adicionales sobre la serie Zoey, por ejemplo su categoría, quién es el director, quiénes son los actores, etc. La Figura 8, muestra un extracto del resultado sobre la consulta de la serie Zoey. Esta nueva información es actualizada en la base de conocimiento del sistema para que el sistema recomendado la utilice al momento de ofrecer sugerencias de programación al usuario.

```
<root response="True">
<movie title="Zoey 101"
  year="2005–"
  rated="N/A"
  released="09 Jan 2005"
  runtime="30 min"
  genre="Comedy, Drama, Family"
  director="N/A"
  writer="Dan Schneider"
  actors="Jamie Lynn Spears, Paul Butcher, Christopher M
  assey, Erin Sanders"
  ....
  type="series"/>
</root>
```

Figura 8. Resultado del servicio OMdb a la consulta sobre la serie Zoey

5. TRABAJOS RELACIONADOS

En esta sección se describen algunos trabajos relacionados con esta propuesta. En el enfoque presentado en [18], se describe un sistema de recomendación personalizada de televisión. El sistema confía en las preferencias del usuario provistas explícitamente, evitando así la recuperación automática de información sobre el usuario. Aunque el sistema hace un análisis de las preferencias basado fundamentalmente en las clasificaciones explícitas de los contenidos vistos por el usuario, el sistema emplea además la información sobre el contenido no visto, lo que indica un método primitivo para detectar desinterés.

En el proyecto Nazou[20] se propone una ontología para modelar un usuario. En la ontología se definen los conceptos que representan las características del usuario y se identifican las relaciones entre éstas características individuales. El modelo (después de su población) es usado por herramientas de presentación para ofrecer contenido y navegación personalizada. El modelo es empleado además en herramientas de organización de contenido (por ejemplo, clasificación de elementos en función de las preferencias del usuario).

En [19], el autor propone la expansión automática de metadatos, un método utilizado para enriquecer los metadatos de los programas de televisión sobre la guía de programación electrónica (EPG) y un diccionario de conceptos asociado. La información transaccional, como la grabación, búsqueda y votación de programas es procesada automáticamente por un motor basado en reglas, con el fin de actualizar las interfaces de usuario. Un filtrado colaborativo es utilizado para rastrear las preferencias del usuario inducida por pares.

En [20] se describe una guía de programas de televisión personal. Esta guía almacena información estereotipada acerca de las preferencias de visualización de televisión de diferentes categorías de usuarios, por ejemplo, una ama de casa. Una ontología de dominio general es empleada para modelar el espacio de conceptos sobre las preferencias de televisión, incluyendo categorías jerárquicas de programas. El conocimiento ontológico estructurado permite que la guía pueda asignar eficazmente diferentes caracterizaciones de programas de TV a un solo vocabulario de categorías de género.

El proyecto europeo NoTube[21] finalizó en enero de 2012, con el objetivo de llevar Internet y televisión juntos a través de contenidos y modelos de datos compartidos. Los integrantes del proyecto desarrollaron diferentes servicios, APIs y aplicaciones para una televisión personalizada, basada en la recomendación de contenidos de televisión en función de los intereses de los usuarios. El perfil de los intereses de un usuario se genera automáticamente a partir de las actividades ejecutadas en la Web por los usuarios que utilizan la API Beancounter desarrollada en NoTube. El contenido de ese perfil se expone en línea y se controla a través de una interfaz de usuario dedicada para ello. Allí el usuario puede seleccionar qué recursos va a utilizar y se permite agregar, eliminar o modificar las ponderaciones de los temas seleccionados en el perfil. El Beancounter monitorea los cambios en las actividades sociales de la Web y actualiza el perfil del usuario.

En el trabajo presentado en [21] se aborda el problema de la sobrecarga de información mediante la definición de ontologías de alto nivel, así como ontologías de dominio. Los autores describen además un escenario para la asistencia médica externa, donde la información dependiente del contexto es usada para el tratamiento del paciente. En [22], los autores proponen una jerarquía para modelar los intereses de un usuario. En su propuesta, los autores sugieren que las anotaciones efectuadas sobre páginas Web pueden ser usadas para identificar los intereses generales y específicos de los usuarios.

[18]http://www.omdbapi.com/
[19]http://thetvdb.com/
[20]http://nazou.fiit.stuba.sk/
[21]http://www.notube.tv

6. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se presenta una propuesta genérica que permite capturar, manipular, y serializar el perfil de un usuario de TV digital. El proceso de construción del perfil considera el uso de diferentes recursos ontológicos relacionados al dominio permitiendo la creación de una red de ontologías. Esta red de ontologías genera una base de conocimiento que permite inferir las preferencias televisivas de un usuario. La combinación de la información de las redes sociales con conocimiento disponible en la Web semántica es en nuestra opinión un enfoque poderoso y prometedor para proporcionar flexibilidad y multidominio en los sistemas de recomendación.

Además, se han presentado técnicas que permiten recuperar información desde las redes sociales y anotar semánticamente esta información mediante ontologías de dominio. Estas mismas ideas pueden ser usadas para dar significado a datos de diferentes dominios y que poseen características similares como: ambiguedad, formatos libres de esquemas o dificultad de procesamiento automático.

Como trabajo futuro, se incluye la evaluación del aporte individual de cada una de las redes sociales involucradas en esta investigación a la información de los hábitos de consumo televisivo de un usuario. Se espera obtener información útil que permita aplicar procedimientos similares en otras aplicaciones sensibles al contexto.

AGRADECIMIENTOS

La investigación descrita en este trabajo 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. Se agradece la ayuda de Juan José Saenz y Sebastián Román en la implementación del módulo de conexión a redes sociales.

REFERENCIAS

[1] V. Saquicela, M. Espinoza, J. Mejía, and B. Villazón-Terrazas, "Reduciendo la sobrecarga de información en usuarios de televisión digital," in Proceedings of the *Workshop on Semantic Web and Linked Data*, 2014.

[2] D. Tsatsou, M. Loli, V. Mezaris, R. Klein, M. Kober, T. Kliegr, J. Kuchar, M. Mancas, J. Leroy, and L. Nixon, 2012, specification of user profiling and contextualisation. LinkedTV Project Deliverable 4.1.

[3] A. Ebersbach, M. Glaser, and R. Heigl, 2010, social web. UVK Verlagsgesellschaft mbH.

[4] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, 2001.

[5] E. Portmann, "The social semantic web," in *The FORA Framework*, ser. Fuzzy Management Methods. Springer Berlin Heidelberg, 2013, pp. 13–36.

[6] J. G. Breslin, A. Passant, and S. Decker, *The Social Semantic Web*. Springer Berlin Heidelberg, 2010.

[7] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198–208, Apr. 2006.

[8] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in Proceedings of the 17th International Conference on World Wide Web, ser.WWW '08, 2008.

[9] N. WeiBenberg, A. Voisard, and R. Gartmann, "Using ontologies in personalized mobile applications," in Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems, ser. GIS '04, 2004.

[10] M. Grüninger and M. S. Fox, "Methodology for the design and evaluation of ontologies," in *International* Joint Conference on Artificial Inteligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[11] M. Espinoza and V. Saquicela, "Modelando los hábitos de consumo televisivo usando tecnología semántica," in Proceedings of the IX Congreso de Ciencia y Tecnología ESPE 2014, 2014.

[12] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Entity matching in online social networks." in *SocialCom*. IEEE, 2013, pp. 339–344.

[13] L. Getoor and A. Machanavajjhala, "Entity resolution: Tutorial," University of Mariland, http://www.cs.umd.edu/ getoor/ Tutorials/ER_VLDB2012.pdf.

[14] D. Eguez, P. Guerra, N. Nebot, A. Ortega, and M. Santoro, "Perfiles de consumo de redes sociales en jóvenes: Una perspectiva cuantitativa desde el marketing," Universidad Casa Grande, Facultad de Comunicación, 2010.

[15] F. Vélez and M. Espinoza, "Definiendo los actores del sistema de recomendación para usuarios de tv digital," Reporte Técnico del DCC, https://drive.google.com/file/d/0BzHHoovgc5fBc WptbEpjYlRQWjg/edit?usp=sharing.

[16] K. Palacio, H. Alban, M. Espinoza, V. Saquicela, J. Avila, and X. Riofrío, "Análisis de la influencia de las propiedades semánticas en los sistemas de recomendación," Revista Politécnica, Escuela Politécnica *Nacional*, vol. 34, 2014.

[17] V. Saquicela, M. Espinoza, K. Palacio, and H. Albán, "Enriching electronic program guides using semantic technologies and external resources," in *Proceedings* of the XL Latin American Computing Conference, ser. CLEI'14, 2014.

[18] A. Barragáns, J. Pazos, A. Fernández, J. García, and M. López, "What's on TV Tonight? An Efficient and Effective Personalized Recommender System of TV Programs," IEEE Transactions on Consumer Electronics, vol. 55, no. 1, pp. 286–294, feb 2009.

[19] T. Tsunoda and M. Hoshino, "Automatic metadata expansion and indirect collaborative filtering for tv program recommendation system," *Multimedia Tools Appl.*, vol. 36, no. 1-2, pp. 37–54, Jan. 2008.

[20] L. Ardissono, C. Gena, P. Torasso, F. Bellifemine, A. Difino, and B. Negro, "User modeling and recommendation techniques for personalized electronic program guides," in Personalized Digital Television Targeting Programs to Individual Viewers. Kluwer Academic Publishers, 2004, pp. 3–26.

[21] F. Bobillo, M. Delgado, and J. Gómez-Romero, "Representation of context-dependant knowledge in ontologies: A model and an application." *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1899–1908, 2008.

[22] H. R. Kim and P. K. Chan, "Learning implicit user interest hierarchy for context in personalization," in Proceedings of the 8th International Conference on Intelligent *User Interfaces*, ser. IUI '03. New York, NY, USA: ACM, 2003, pp. 101–108.