

Collecting Data about Television Contents in the OntoTV System

A model for collecting television content data from various existing sources, in order to improve the available information about digital television programs

José Luis Redondo-García, Alvaro E. Prieto, Adolfo Lozano-Tello
Universidad de Extremadura. Quercus Software Engineering Group
Escuela Politécnica, Campus Universitario s/n, 10071
Cáceres, Spain
jluisred@unex.es, aeprieto@unex.es, alozano@unex.es

Nowadays, digital television and the in particular the terrestrial one, offers a great variety of digital television contents and channels. However, the available information about the broadcasted programs is scarce and not always immediately accessible. It is necessary to develop new strategies that retrieve all the possible television content information, in order to improve the viewers' experience.

The main objective of this research is the definition of a new input module for the OntoTV content management system, that (1) collects information from various existing platforms, (2) enriches it by accessing additional resources, (3) integrates into a single document all the relevant data, and (4) represents this information in the knowledge base by using ontologies. Then, it is possible to perform advanced operations over the stored data, like more precise queries for the user.

Keywords-digital television; ontology; collection; XMLTV;

I. INTRODUCCIÓN

A pesar de que hoy en día existe una gran variedad de plataformas digitales desde las que es posible ver televisión, en la mayoría de ocasiones la información sobre los contenidos que se ofrece es insuficiente. El usuario, para solucionar este inconveniente, se ve obligado a consultar en fuentes alternativas aquellos detalles que no se recogen en las guías originales.

Es común el caso del televidente que, tras revisar la oferta televisiva de un canal de TDT (Televisión Digital Terrestre), encuentra interesante un determinado contenido, por ejemplo "The Walking Dead". Pero a menudo sólo hay disponible sobre él una breve descripción. Si quiere conocer más datos tendrá que recurrir, por ejemplo, a búsquedas en Internet. Para evitar esta escasez en la información que proporcionan los proveedores, sería útil un sistema que recopile autónomamente la información e integre de manera transparente todos los datos sobre programas disponibles.

Por otro lado, ciertas funcionalidades como las búsquedas personalizadas o las recomendaciones, resultan cada día más familiares y cercanas para el usuario. Estas operaciones aparecen ya en algunos sitios Web, como el conocido YouTube. Pero sólo tiene sentido trasladarlas a escenarios como el de la TDT si la información disponible en ellos es rica y detallada, y esto todavía no ocurre.

La implementación real de estos mecanismos no supone actualmente un problema. Existen sistemas de televisión digital con amplias posibilidades en programación de aplicaciones e interactividad, como MHP [1] o Google TV¹. Este último en particular tiene una gran proyección de futuro, incluyendo soporte para casi todo tipo de plataformas de televisión digital, ejecución de aplicaciones de terceros, o compatibilidad gran variedad de dispositivos y terminales de usuario. Estos sistemas pueden incorporar un gestor de información sobre contenidos televisivos con funcionalidades avanzadas que les otorgue valor añadido de cara a sus usuarios.

De esta forma surge un nuevo enfoque de gestión de información sobre contenidos televisivos, en el que los emisores de contenidos, los desarrolladores de sistemas de televisión digital y por supuesto los propios televidentes, resultan beneficiados.

II. ONTOLOGÍAS PARA LA GESTIÓN DE CONTENIDOS TELEVISIVOS

Una forma de representar la información relativa a contenidos televisivos y preferencias de televidentes es mediante el uso de ontologías [2]. El sistema descrito en [3] presenta una arquitectura que contempla el uso de las mismas para dar soporte a algoritmos de recomendación. Las ontologías propuestas (usuario, contexto, contenido, y categorías) constituyen una buena aproximación para

¹ <http://www.google.com/tv/>

representar datos relativos al dominio televisivo y realizar inferencias sobre ellos. También el proyecto AVATAR [4] utiliza ontologías en formato OWL, centrándose especialmente en el almacenamiento de las preferencias de los usuarios. Recurre a técnicas de razonamiento semántico para realizar operaciones de recomendación más precisas. Pero en ambos casos no se hace referencia a la forma en que se consiguen los datos sobre contenidos televisivos que se almacenan en estas ontologías.

Existen otros trabajos que sí contemplan la integración de distintas fuentes de datos televisivos. NoTube [5] es un gestor de contenidos televisivos capaz de acceder a guías electrónicas de varias plataformas. Rescata preferencias y comportamientos de los usuarios analizando su actividad en portales como Twitter o Facebook. Por otro lado el sistema Sensee [6] propone la integración de diversas fuentes de datos televisivos, como la programación de la BBC², guías en XMLTV³, o descripciones de películas del portal IMDB⁴.

III. EL SISTEMA ONTOTV

A pesar de que existen proyectos que utilizan ontologías para la representación e integración de información sobre contenidos televisivos procedentes de distintas fuentes, no se ha encontrado ninguno que describa con detalle un mecanismo de recolección de datos que opere de forma consensuada para obtener información correcta y relevante.

El sistema OntoTV (ONTOlogy-based management system for digital TeleVision) es un sistema gestor de datos sobre contenidos televisivos que ofrece una nueva forma de recolectar y acceder a la información sobre los programas que han sido, están siendo, o serán emitidos en las distintas plataformas digitales. Debido a que incorpora mecanismos adecuados de adquisición de datos, el sistema puede ofrecer al usuario descripciones detalladas de contenidos y le permite realizar operaciones avanzadas de búsqueda y recomendación con alto grado de personalización.

A. Antecedentes

El sistema OntoTV fue presentado anteriormente en [7], donde se estableció el uso de un modelo basado en ontologías para representar la información relativa al dominio televisivo sobre contenidos y televidentes. Igualmente se describieron operaciones de búsqueda y recomendación y la mecánica de interacción del usuario con el sistema.

Además se mostró un mecanismo para la lectura de datos en el flujo de emisión de la TDT (Televisión Digital Terrestre) con el que es posible acceder a la información sobre contenidos que los proveedores incluyen en él. Pero esta información es muy escasa, por eso en el documento actual se amplía y mejora este proceso de captura de datos.

OntoTV contiene un nuevo módulo de entrada de datos más general, que engloba tanto el caso específico de la lectura de datos de la TDT como otros procesos similares de adquisición de información para el resto de plataformas de televisión

digital. También se incluyen componentes que acceden a fuentes alternativas para complementar datos y algoritmos de selección de información relevante.

B. Características Generales

A continuación se muestran las características más destacadas del sistema OntoTV:

1. Incorpora un modelo ontológico para el almacenamiento y representación de la información. El objetivo es conseguir una correcta estructuración de los datos y mejorar el funcionamiento de los módulos de la recomendación y búsqueda de contenidos.
2. Presenta una arquitectura de tipo modular y centralizada. Existe un módulo principal que realiza tareas de almacenamiento de datos y ejecución de operaciones de búsqueda y recomendación que es utilizado por el resto de componentes de entrada, de salida, y de transformación de los datos.
3. Implementa una arquitectura de tipo cliente/servidor. Hoy en día la mayor parte de los servicios interactivos que se ofrecen en Internet siguen este mismo patrón, porque se libera a los terminales de realizar tareas costosas, y se minimiza la duplicación de datos entre los distintos terminales de usuario.
4. Busca cumplir las premisas de escalabilidad y flexibilidad ante futuros cambios. Pretende sentar las bases para la creación de un sistema gestor de información sobre contenidos de televisión digital de carácter abierto y superior en funcionalidades a los implementados en los decodificadores convencionales.
5. Está pensado para ser ejecutado en distintas plataformas, ofreciendo un servicio global de gestión de contenidos independiente del sistema de televisión en el que se ejecute. Podrán acceder a él clientes que posean decodificadores basados en MHP, dispositivos de Google TV, o cualquier Media Center.

C. Módulo de Recolección de Datos

La arquitectura del sistema OntoTV que se muestra en la Figura 1 contiene tres módulos distintos:

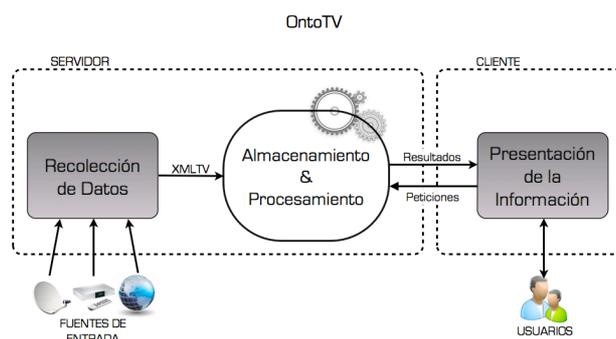


Figura 1. Esquema de los módulos principales de OntoTV

- Un módulo de “Recolección de Datos”. Este módulo, que se describirá con mayor detalle en las siguientes subsecciones, se encarga de recopilar toda la

2 <http://backstage.bbc.co.uk/>

3 <http://wiki.xmltv.org/index.php/XMLTVFormat>

4 <http://www.imdb.com/>

información posible sobre contenidos televisivos e incorpora mecanismos de acceso a las diversas fuentes existentes.

- Otro módulo de “Almacenamiento y Procesamiento”, que toma todos estos datos, los representa utilizando ontologías, y permite ejecutar operaciones de búsqueda y recomendación sobre ellos.
- Finalmente el usuario interactúa con el sistema enviando peticiones para la ejecución de operaciones y recibiendo los resultados de las mismas, a través del módulo de “Presentación de la Información”. Se ha empleado la experiencia previa adquirida en [8] sobre implementación de aplicaciones para televisión digital.

Tanto el módulo de “Recolección de Datos” como el de “Almacenamiento y Procesamiento” están implementados en el lado del servidor. Sólo el de “Presentación de la Información” se ejecutará en el terminal del usuario.

IV. EL MECANISMO DE RECOLECCIÓN DE DATOS

En este apartado se describe con mayor detalle el mecanismo que permite que la información procedente del exterior sea incorporada al sistema.

A. Formato de los Datos de Entrada

El sistema de televisión digital que se propone necesita de un formato de datos que permita recoger de manera completa la información de las fuentes de datos externas, para posteriormente incorporarla al sistema.

Se han considerado tres alternativas. La primera de ellas ha sido el estándar TV-Anytime [9], por su alta capacidad para describir contenidos y preferencias de usuarios, y por estar claramente orientada al ámbito televisivo, como lo demuestra el hecho de haber sido desarrollado gracias al esfuerzo conjunto de los principales fabricantes del sector. Otra opción que se ha tenido en cuenta ha sido el formato MPEG7 [10], un estándar basado en XML y desarrollado para describir los atributos, la semántica y la estructura de cualquier tipo de contenido multimedia. Su punto a favor reside en su versatilidad y su proyección de futuro. En contra tiene que resulta demasiado general y extenso e incluye atributos de muy bajo nivel no necesarios en el ámbito televisivo. Por último está el ya mencionado formato XMLTV, que al igual que TV-Anytime, está orientado específicamente al ámbito televisivo. Su mayor desventaja es que las descripciones que ofrece son poco detalladas; apenas se permite almacenar preferencias de usuarios y se tiende a duplicar información. Pero precisamente esta simplicidad provoca que su uso esté bastante extendido en diversas fuentes de información sobre contenidos.

A pesar de que en el futuro el sistema debe contemplar los tres formatos, se ha decidido que el mecanismo de recolección de datos descrito en este documento se centre en el estándar XMLTV, por ser el más utilizado en la actualidad y porque los elementos descriptivos recogidos en el formato, aunque lejos de ser los ideales, son suficientes para las descripciones sobre contenidos televisivos que se han manejado en las pruebas.

B. Componentes dentro del Módulo de Recolección de Datos

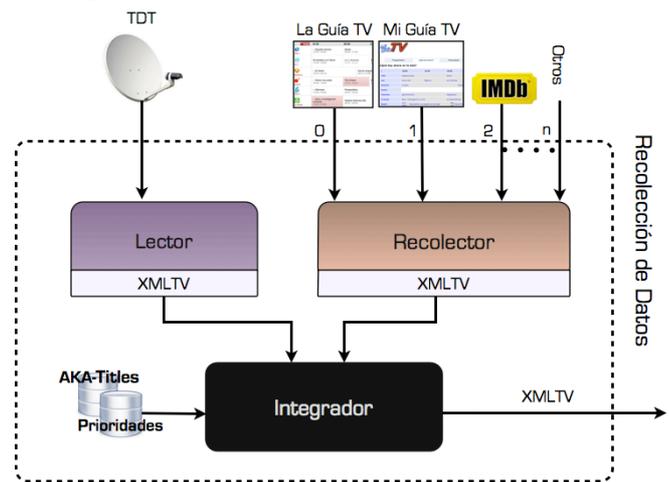


Figura 2. Esquema del módulo de recolección de datos

En la Figura 2 se pueden observar los distintos componentes que forman el módulo y las relaciones que se establecen entre ellos:

- En primer lugar, están los componentes de tipo “Lector” dedicados a la extracción de datos de las plataformas de televisión digital soportadas por el sistema. Suponemos que el sistema cuenta con hardware para la recepción de señal DVB-T. Por ello se implementa el mecanismo que ya fue descrito en [7], consistente en utilizar el estándar DVB-SI y DVB-EIT para acceder a la información contenida dentro del flujo de emisión de la TDT. Esta información se transforma a formato XMLTV.
- OntoTV, además de leer guías de programación directamente de las plataformas disponibles (TDT, televisión por cable...) puede acceder a fuentes alternativas, principalmente en Internet, para evitar la incompletitud en la información. El componente “Recolector” engloba a un conjunto de agentes que se conectan a servidores externos, normalmente utilizando el protocolo TCP/IP, para rescatar otras guías de programación e información complementaria.
- El último componente se denomina “Integrador”, y es el encargado de incluir en un solo archivo de datos toda la información en XMLTV procedente de los dos componentes anteriores. Debido a que es posible que el sistema consiga rescatar descripciones de un mismo programa desde varias de las fuentes disponibles, pueden producirse solapamientos e incongruencias entre ellas que deben ser resueltos. En el apartado C.2 se describe cómo se solucionan estas ambigüedades.

C. Fases del Proceso de Recolección

El mecanismo de recolección de datos sobre contenidos televisivos está dividido en dos fases: recolección y fusión. Como el formato XMLTV ha sido el elegido para la entrada de datos, también se utiliza para el intercambio de resultados parciales entre estas etapas.

1) Recolección

En esta fase se realiza la lectura directa de datos desde las fuentes contempladas en el sistema. El objetivo es interpretar el formato que utiliza una determinada entrada y transformar la información leída al tipo de datos soportado por el sistema, en este caso XMLTV. Para ilustrar el funcionamiento de esta fase se va a recurrir a una prueba que utiliza información real. En su desarrollo se ha configurado el acceso a las siguientes fuentes:

- Información incluida en la señal de la TDT, mediante el procedimiento descrito en [7].
- Acceso a “La Guía TV”⁵. Se trata de un portal Web con información relativa a horarios de emisión de contenidos para los principales canales de televisión. También incluye descripciones más o menos detalladas de algunos de ellos. Para realizar la conversión del formato HTML de la página a XMLTV, se utilizará el paquete de código abierto xmltv-0.5.59⁶.
- Acceso a “Mi Guía TV”⁷. Se trata de un portal Web de características parecidas al del caso anterior, por lo que se procederá de utilizando también xmltv-0.5.59.

```

<!-- FUENTE: SEÑAL DE TDT -->
<programme channel="8" start="20101214130000" stop="20101214140000">
  <title>Los Simpsons</title>
  <sub-title>Maridos y cuchilladas y El día que murió la viuda</sub-title>
  <desc>Los momentos más divertidos de la serie, anécdotas, juegos, sigue la serie en Internet y pásatelo en gr... Piensa en amarillo. bros descubren que tiene una ma... de nacimiento en su cuerpo, le convierten en el Ele...</desc>
</programme>

<!-- FUENTE: LA GUÍA TV -->
<programme start="20101214140000 +0100" channel="Antena-203.1" lang="es">
  <title lang="es">Los Simpson</title>
  <category lang="es">serie</category>
</programme>

<!-- FUENTE: MI GUÍA TV -->
<programme start="20101214140000 +0100" stop="20101214143000" lang="es">
  <title lang="es">Los Simpson</title>
  <desc lang="es">Los Simpson son una familia políticamente de clase media, que habita en la ficticia localid... nacieron en 1987 como una sección de 'El show de... donde aparecían cinco minutos en cada emisión. Lo...</desc>
  <date>20050908</date>
  <category lang="es">Animación</category>
  <category lang="es">Programas</category>
  <category lang="es">Series</category>
  <length units="minutes">30</length>
</programme>

```

Figura 3. Fragmentos XMLTV recolectados de distintas fuentes de datos

En la Figura 3 se observan tres fragmentos de ficheros XMLTV procedentes de cada una de las fuentes, y relativos a un capítulo de la serie “Los Simpsons” emitido en España por el canal “Antena 3” el 14 de Diciembre de 2010. Se aprecian diferencias en el nivel de detalle ofrecido por cada uno: por

5 <http://www.laguia.tv/>

6 <http://sourceforge.net/projects/xmltv/>

7 <http://www.miguia.tv/>

ejemplo el de “La Guía TV” resulta realmente escaso, mientras que el de “Mi Guía TV” contiene abundante información acerca de las categorías asociadas al contenido. Con todos estos archivos de datos ya disponibles se aplica la siguiente fase.

2) Fusión

Ya se dispone de la información sobre contenidos televisivos que se buscaba, pero ésta se encuentra repartida en cada uno de los ficheros generados. Por ello en esta fase se realiza una selección de los datos más relevantes y se unifican en un solo archivo. De esta tarea se encarga el componente “Integrador” descrito anteriormente y mostrado en la Figura 2. Una vez que finaliza su ejecución proporciona un archivo único de datos con toda la información en XMLTV que antes estaba disgregada. Estos son los pasos que lleva a cabo:

a) *Agrupar descripciones correspondientes a un mismo contenido.* Aunque puede parecer una tarea trivial, en ocasiones no es inmediato determinar si descripciones procedentes de distintas fuentes hacen referencia a un mismo contenido. Puede ocurrir que su nombre no coincida exactamente, que la hora de inicio se haya retrasado, etc. Por ello a la hora de tomar una decisión al respecto se aplicará un algoritmo que tiene en cuenta distintos criterios:

- Similitud espacio-temporal del contenido (S_{E-T}). Si un programa pertenece al mismo canal, y comienza y finaliza a una hora similar, entonces la posibilidad de que se trate del mismo programa es alta.
- Similitud en los títulos (S_T). Se aplican en este campo funciones de comparación relativa de cadenas, que informan del grado de parecido entre los operandos. Para el caso de películas, los Aka-Titles proporcionados por IMDB permiten identificar una misma película a pesar de que su título se presente en alguno de los idiomas alternativos.
- Similitud global (S_{Global}). Dadas dos descripciones, se buscan palabras que aparezcan en ambas a la vez, sin importar el lugar exacto en que lo hagan. Si el número de ítems coincidentes es alto, la probabilidad de que sea el mismo contenido también lo será.

Para cada similitud S_x se obtiene un valor numérico. El algoritmo finaliza realizando la suma ponderada de estos resultados parciales, donde los pesos indican el grado de importancia otorgado a cada uno de ellos. Si el resultado supera un determinado umbral, se considera que las dos descripciones estudiadas pertenecen a un mismo contenido.

$$C_A \approx C_B \Leftrightarrow P_1 \cdot S_{E-T} + P_2 \cdot S_{Titulo} + P_3 \cdot S_{Global} \geq Umbral \quad (1)$$

Donde C_A y C_B son contenidos televisivos, S_x los resultados parciales de cada similitud y P_i su peso asociado. Además:

$$S_x \in [0, 1], \quad P_1 + P_2 + \dots + P_n = 1, \quad 0 < Umbral \leq 1$$

Tras las pruebas realizadas, se han fijado como pesos más adecuados los siguientes: $P_1=0.6$, $P_2=0.25$, $P_3=0.15$.

b) *Fusionar las descripciones pertenecientes a un mismo grupo.* Una vez que se tienen localizadas todas las descripciones correspondientes a un mismo contenido hay que ver qué información puede aportar cada una de ellas. Si una

descripción tiene relleno un campo que el resto traía vacío, se toma inmediatamente. Si por el contrario se producen solapamientos de información, los campos se concatenan si es posible. Si no lo es, se eliminan los que procedan de fuentes con menor importancia. En este trabajo se ha optado por dar la prioridad más alta a la información leída directamente de la TDT por ser la que ofrecen los propios proveedores. Al final de este paso se obtiene una descripción única de cada contenido que incorpora toda la información recuperada de cada una de las fuentes de datos consultadas.

c) *Elaborar la guía única.* Para terminar, todas las descripciones generadas en el paso anterior se exportan a un mismo documento en formato XMLTV.

V. RESULTADOS

Cuando finaliza el proceso de recolección de datos toda información se almacena ya en un mismo documento en formato XMLTV sin redundancias, correctamente estructurado, y listo para ser incluido en la base de conocimiento.

```
<programme start="20101214140000 +0100" channel="Antena3.Onto"
  <title lang="es">Los Simpsons</title>
  <sub-title>Maridos y cuchilladas y El día que murió la v
  <desc lang="es">Los momentos más divertidos de la serie,
    juegos, sigue la serie en Internet y pásatelo en g
    Piensa en amarillo. bros descubren que tiene una m
    de nacimiento en su cuerpo, le convierten en el EL
    Los Simpson son una familia políticamente incorrec
    de clase media, que habita en la ficticia localida
    nacieron en 1987 como una sección de 'El show de T
    donde aparecían cinco minutos en cada emisión. La
  </desc>
  <date>20050908</date>
  <length units="minutes">30</length>
  <category lang="es">Animación</category>
  <category lang="es">Programas</category>
  <category lang="es">Series</category>
</programme>
```

Figura 4. Fragmento XMLTV tras aplicar el proceso de recolección

En la Figura 4 se observa cómo ha mejorado la información existente sobre el capítulo de “Los Simpsons”, antes dispersa entre los distintos ficheros de las fuentes de entrada y ahora integrada bajo un mismo elemento de tipo “programa”:

- Los campos que sólo aparecían en uno de los ficheros de origen, se han añadido inmediatamente, como el caso de “sub-title”.
- Campos coincidentes con posibilidad de fusionarse en uno sólo, como por ejemplo “desc”, así lo han hecho.
- Si no tenía sentido fusionar, se ha dado prioridad a los campos procedentes de la TDT, como en el caso de “title”, que aparecía en cada uno de los tres ficheros iniciales con pequeñas variaciones.

Se observa que gracias al mecanismo de recolección utilizado ha aumentado la calidad de las descripciones y por tanto las posibilidades que se ofrecen al usuario.

A. Incorporación de Resultados a la Base de Conocimiento

Para que la información recién obtenida sea incluida en la base de conocimiento, se lleva a cabo el proceso de lectura y de transformación del fichero XMLTV a formato OWL, el utilizado en las ontologías del sistema. En OntoTV el encargado de realizar esta tarea es el módulo de “Almacenamiento y Procesamiento”. Para ello utiliza la API Jena⁸.

B. Ontologías

A pesar de que el objetivo de este documento no es entrar en detalles acerca de las ontologías empleadas en OntoTV, éstas serán descritas brevemente para comprender dónde van almacenándose exactamente los datos capturados.

La primera de las ontologías es la de datos televisivos, denominada Onto-TVData. En ella se almacena toda la información que el sistema ha adquirido sobre contenidos. Está basada en la utilizada por el proyecto AVATAR, disponible para su consulta en <http://avatar.det.uvigo.es/software.html>. Su estructura presenta múltiples subclases que establecen la jerarquía de categorías (deporte, películas, series...). Además, gran parte de sus propiedades encuentran correspondencia directa con campos específicos del formato XMLTV, por lo que se facilita el proceso de incorporación de los ficheros de datos a la base de conocimiento.

Existe una segunda ontología denominada Onto-UserProfile, donde se almacena información relativa a las preferencias de usuarios sobre contenidos televisivos. Gracias a ello, se pueden realizar consultas y recomendaciones altamente personalizadas. Los campos del formato XMLTV que se incluyen en ella son aquellos relacionados con la valoración del televidente sobre el contenido y sus opiniones personales.

C. Interfaz Web de Consulta de Información

Para comprobar que los resultados generados por el proceso de recolección de datos son satisfactorios, se ha utilizado una interfaz Web que permite el acceso a los datos contenidos en la base de conocimiento.

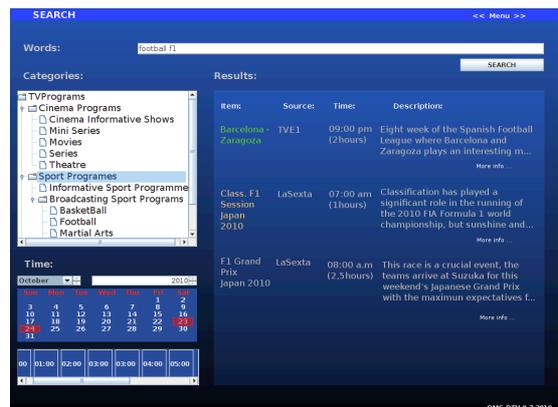


Figura 5. Interfaz Web para el acceso a información sobre contenidos

⁸ <http://jena.sourceforge.net>

Únicamente pueden realizarse operaciones de consulta, introduciendo palabras clave o navegando por la jerarquía de clases de la ontología. Pero resulta suficiente para mostrar la ventaja que para el usuario supone el disponer de una información mucho más completa que la que se recolectaría utilizando mecanismos convencionales.

VI. LINEAS FUTURAS

Los resultados obtenidos por el módulo de “Recolección de Datos” de OntoTV son muy satisfactorios. Pero el proyecto continúa en desarrollo, especialmente en los siguientes aspectos:

- En el módulo de entrada, los ficheros XMLTV obtenidos durante el proceso de recolección de datos pueden ser enriquecidos utilizando recursos que proporcionen de información adicional sobre campos vacíos o incompletos. Por ejemplo, se recurrirá a páginas Web que los proveedores de contenidos mantienen en Internet, y a portales que recojan valoraciones de los usuarios sobre los contenidos.
- Por otro lado, se debe aumentar la compatibilidad con formatos avanzados como TV-Anytime y MPEG-7, analizados en este documento. Aún no existen suficientes fuentes que ofrezcan datos en este formato, pero están llamados a ser los estándares de referencia en años posteriores.
- En el módulo de procesamiento deben añadirse nuevas operaciones avanzadas y refinar las ya existentes: búsqueda de contenidos, recomendaciones, elaboración de guías de programación a medida, etc. Para mejorar el grado de personalización se implementará un aprendizaje dinámico de las preferencias de los usuarios, a medida que estos interactúan con el sistema.
- El modelo de aplicación propuesto, basado en una arquitectura cliente-servidor, permite que sea implementado sobre cualquier sistema de televisión digital. Desarrollar versiones para múltiples plataformas, como por ejemplo las ya mencionadas Google TV y MHP, permitirá convertir OntoTV en un marco común de acceso a información sobre contenidos televisivos.

VII. CONCLUSIONES

Actualmente los usuarios de televisión digital no tienen acceso a información suficiente sobre los contenidos televisivos que se están emitiendo o van a ser emitidos. Para solucionar este inconveniente se ha desarrollado un módulo de recolección de datos que está integrado dentro del sistema OntoTV. En él se implementa un mecanismo para la obtención de información sobre contenidos televisivos que contempla:

- El acceso a varias fuentes de datos sobre contenidos televisivos, para extraer la información disponible en ellas.

- La integración de toda la información recolectada, seleccionando aquellos datos más relevantes y eliminando redundancias.

El resto de módulos del sistema toman estos datos recolectados para llevar a cabo sus funcionalidades. Adicionalmente, y para comprobar la validez del modelo, se ha utilizado una aplicación Web de acceso a la base de conocimiento que permite realizar consultas y observar el nivel de detalle ofrecido en las descripciones de contenidos.

Actualmente es necesario seguir mejorando ciertas partes del proceso de recolección, como la fase de agrupación de descripciones correspondientes a un mismo contenido. Los algoritmos de similitud global y en los títulos están en fase de desarrollo para incorporar diccionarios de sinónimos y procesamiento de lenguaje natural. Pero el objetivo principal se ha conseguido, y el sistema OntoTV proporciona a los usuarios información mucho más abundante y completa que la ofrecida por los tradicionales sistemas televisivos. Esta información es además fundamental para que las operaciones avanzadas, como consultas personalizadas y recomendaciones, sean realmente efectivas.

AGRADECIMIENTOS

Este trabajo está patrocinado por la beca de investigación PD10006 de la Junta de Extremadura, por el proyecto TIN2008-02985 del Ministerio de Ciencia e Innovación, y por el Fondo Europeo de Desarrollo Regional (FEDER).

REFERENCIAS

- [1] J. Piesing, “The DVB multimedia home platform (MHP) and related specifications,” *Proc. IEEE*, vol. 94, no. 1, pp. 237–247, Jan 2006.
- [2] A. Gomez-Perez, O. Corcho, and M. Fernandez-Lopez, “Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web,” Springer-Verlag, New York, 2004.
- [3] Y. Naudet, A. Aghasaryan, Y. Toms, and C. Senot, “An Ontology-Based Profiling and Recommending System for Mobile TV,” *SMAP '08*, Prague, pp. 94–99, December 2008.
- [4] Y. Blanco-Fernández, J.J. Pazos Arias, A. Gil-Solla, M. Ramos Cabrer, M. López Nores, J. García Duque et al, “Exploiting synergies between semantic reasoning and personalization strategies in intelligent recommender systems: A case study,” *Journal of Systems and Software*, vol. 81, pp. 2371–2385, 2008.
- [5] B. Schopman, D. Brickley, L. Aroyo, C. van Aart, V. Buser, R. Siebes, et al. “NoTube: making the Web part of personalised TV,” *Proceedings of the WebSci10: Extending the Frontiers of Society Online*. Apr 2010.
- [6] P. Bellekens, L. Aroyo, G. Houben, A. Kaptein, K. van der Sluijs, “Semantics-Based Framework for Personalized Access to TV Content: The iFancy Use Case,” Springer, ISWC/ASWC 07, pp. 887–894, 2007.
- [7] J.L. Redondo-García, P. Valiente-Rocha, A. Lozano-Tello, “Ontology-based system for content management in Digital Television,” *CISTI 2010*, Santiago de Compostela, pp. 277–283, June 2010.
- [8] J.L. Redondo García, J.L. González-Sánchez, A. Gazo-Cervero, J. Corral-García, “The New Digital Television and the Interactivity in its Multimedia Applications,” *SIGMAP 2009*, pp.167-172.
- [9] A. Butkus, M. Petersen, “Semantic modelling using TV- Anytime Genre metadata”. *euroITV*, LNCS Lecture Notes in Computer Science, vol. 4471, Berlin, pp. 226-234, 2007.
- [10] B. S. Manjunath, P. Salembier, and T. Sikora, “Introduction to MPEG-7 Multimedia Content Description Interface”, John Wiley & Sons Ltd., June 2002.