Sistema de indexación automática de contenidos multimedia

Urtzi Urdapilleta Roy, David Conejero Olesti, Xavier Anguera Miro,
David Cacenabes Vázquez, Fco. Javier Caminero Gil

División de Tecnología del Habla, Telefónica I+D
{urtzi, dco, xanguera, davidcv, fjcg}@tid.es

***Abstract* — La gran cantidad de material multimedia que se genera hoy en día hace difícil al usuario poder encontrar de manera sencilla aquello que busca. El indexado automático de audio/vídeo es un área que está suscitando gran interés, ya que permite indagar en el contenido de los documentos multimedia y extraer información relevante. La mayoría de sistemas existentes permiten analizar un número limitado de características, pero no están pensados para interactuar con otros sistemas complementarios para enriquecer la salida del indexado. En esta publicación presentamos una arquitectura de indexado que permite la implementación de módulos especializados que interactúen entre sí para obtener un resultado que es accesible vía web.**

I. Introducción

La sociedad de la información en la que vivimos actualmente nos proporciona, cada vez más, una ingente cantidad de contenidos multimedia. Toda esta información ha de estar estructurada y catalogada para poder ser de utilidad, ya que en caso contrario, su posterior localización y consulta sería imposible. Tradicionalmente, esta catalogación se venía realizando manualmente por personas (como en el caso de las emisiones de radio y televisión, publicaciones de prensa, etc.), pero es patente la necesidad creciente de poder automatizar, al menos en parte, este proceso.

Debido a la magnitud del problema que plantea la indexación automática, es necesario seguir una aproximación parcial y aplicar técnicas específicas adaptadas a cada escenario concreto planteado. En la bibliografía existen muchos ejemplos de sistemas de indexado acústico y/o de vídeo [1][2][3]. En la mayoría de estos sistemas el indexado se reduce a la transcripción automática de la voz. Otros sistemas que obtienen otras características del audio como información sobre los locutores, el idioma hablado, etc… suelen presentarse como sistemas independientes sin una clara correlación entre ellos. En la presente publicación presentamos un sistema de indexado automático de audio que consta de diferentes módulos independientes usados para analizar el audio desde distintos puntos de vista. El sistema de indexado define un entorno común para todos los módulos que comparten la extracción y el preprocesado de la señal de entrada y los ficheros de entrada y salida. Esto permite desarrollar sistemas muy flexibles que pueden modificarse según el tipo de datos a analizar (tipo de programación, TV/radio, etc.) y que acepten configuraciones de los módulos de indexado en cascada o en paralelo.

La estructura de la presente publicación es la siguiente: en el apartado II se describe la arquitectura general del sistema de indexado propuesto. En los apartados III, IV y V se describen las tres partes principales del sistema, que son la adquisición de los datos, el conjunto de módulos de indexado y la presentación de los resultados por web. Finalmente se encuentran las conclusiones y bibliografía.

II. Arquitectura del sistema de indexado

Con la finalidad de obtener un indexado incremental, se ha creado un sistema de indexado para el ámbito de emisiones de radio y televisión, con la arquitectura que presentamos a continuación.

El sistema está dividido en tres subsistemas, tal como se puede ver en la figura 1. El primero de ellos realiza la adquisición automática de los contenidos y su preparación inicial para poder ser analizados por el subsistema de indexado. El segundo realiza el proceso de indexación propiamente dicho y se plantea como una serie de algoritmos especializados en las diferentes problemáticas, que según el caso trabajarán en serie o en paralelo para la obtención de los resultados. Se ha decidido utilizar varios formatos de ficheros XML como sistema de representación de los resultados e intercambio de información, ya que este lenguaje facilita la portabilidad de los resultados así como su modificación incremental por los diferentes algoritmos. Las ejecuciones de los algoritmos están planificadas de forma automática para que se lleven a cabo sobre los nuevos contenidos adquiridos en el orden adecuado. Finalmente, el último subsistema permite consultar y visualizar los resultados obtenidos. La realización de las consultas y presentación de los resultados se ha planteado mediante la creación de una aplicación web dinámica, capaz de mostrar los contenidos multimedia. Se utiliza el lenguaje XPath [4] para llevar a cabo las consultas necesarias sobre los ficheros de los resultados, facilitando en gran medida el proceso de extracción de los datos de los mismos.

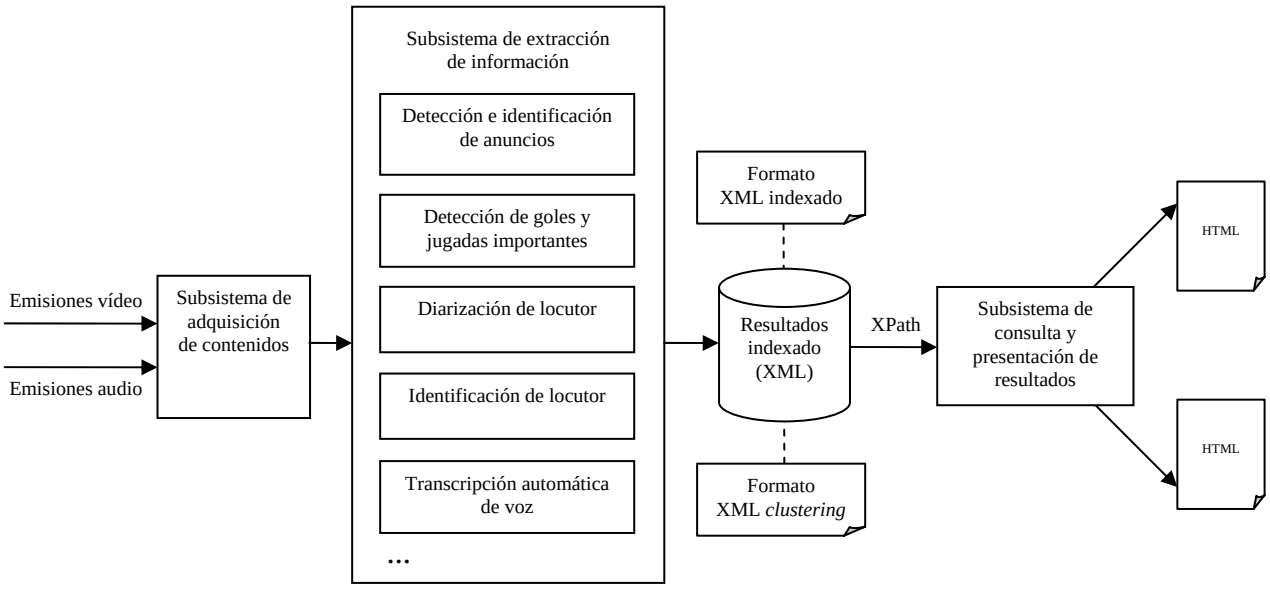

Fig. 1. Visión general del sistema de indexación automática de contenidos multimedia.

III. ADQUISICIÓN DE CONTENIDOS

Las capturas audiovisuales están basadas en la grabación de contenidos emitidos en la Televisión Digital Terrestre (DVB-T). El sistema funciona sobre un sistema operativo Linux y está íntegramente basado en herramientas de código abierto. El proceso de adquisición de contenidos consiste en cuatro etapas, tal como se puede ver en la figura 2.

En DVB-T los canales se envían agrupados en *múltiplex* frecuenciales de 4 ó 5 canales diferentes [5]. Así, cuando se ajusta el sintonizador de televisión a una frecuencia concreta, se está captando un flujo de datos (*Transport Stream*) que contiene los canales de audio, vídeo, subtítulos e información adicional de cada uno de los canales que transporta. La selección de canal entre estos 4 ó 5 se hace a posteriori. La primera parte del proceso consiste en capturar este flujo, haciendo un volcado a disco. Una vez finalizada la captura, se tiene grabado el conjunto de canales que acompañan al que se quiere grabar.

La segunda parte consiste en la demultiplexación del contenido del flujo de datos, dejando únicamente el canal de interés. Esta tarea se realiza a través de una aplicación software diseñada para tal efecto. Ésta coge como entrada el flujo grabado en la primera parte. La salida que proporciona este bloque del sistema consiste en un flujo de vídeo y otro de audio, ambos de baja compresión; en caso de que existieran subtítulos, incluye también un flujo de imágenes que conforman los subtítulos según el estándar de DVB-T y un fichero de texto con los subtítulos extraídos del teletexto. Los que se encuentran en formato imagen son procesados por un sistema de OCR (*Optical character recognition*) específicamente diseñada para subtítulos, que además del texto, extrae información de sincronización temporal. Nótese que se tienen dos fuentes diferentes de subtítulos: los de teletexto y los del canal de datos de DVB-T. Esto se debe a que las cadenas no incorporan la misma información en un sistema que en otro. Los subtítulos se capturan para su posterior uso en los algoritmos de indexado.

La tercera etapa consiste en la codificación de los contenidos para que ocupen menor espacio en disco, y además puedan ser servidos por red a través de *streaming* (algo necesario para el subsistema de presentación de resultados). El *codec* elegido para el vídeo es h.264, o MPEG-4 parte 10 [6]. Se trata de un *codec* de elevada tasa de compresión, manteniendo una buena calidad de imagen. Para codificar el audio se ha escogido el *codec* AAC de compresión con pérdidas. Toda la información se encapsula en el contenedor MPEG-4 Parte 14, conocido como MP4, que permite distribuir flujos por Internet.

La cuarta etapa consiste en la extracción de los parámetros MFCC del flujo de audio sin comprimir, para facilitar su uso en las posteriores etapas de indexado. Al concluir el proceso de captura, se dispone de un fichero que contiene el canal que se quería grabar, codificado para su correcta distribución como flujo de datos a través de la red. Se dispone también de dos fuentes de subtítulos en formato texto, para su posterior uso en diferentes algoritmos de indexado. También se dispone de un fichero con la parametrización MFCC del flujo de audio y una versión inicial del fichero XML de indexado con referencias a todos los ficheros anteriores, que será el punto de partida para el funcionamiento de los diferentes algoritmos.

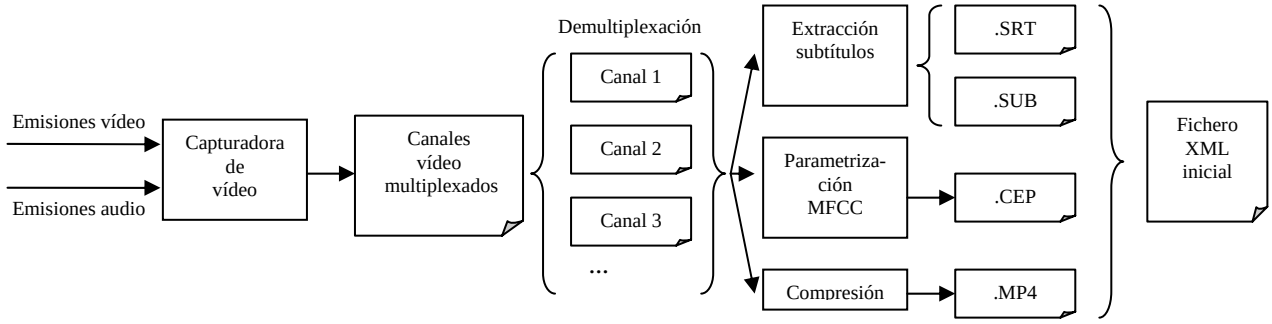

Fig. 2. Visión general del subsistema de adquisición de contenidos.

IV. Subsistema de extracción de información

La filosofía con la que se ha planteado la creación del subsistema de extracción de información ha sido la de generar un sistema flexible y versátil al que se le puedan ir añadiendo funcionalidades de forma incremental a medida que se vayan detectando nuevas necesidades. Por todo ello, se ha diseñado un sistema como un grupo de módulos independientes especializados en ejecutar tareas de indexado concretas. La función de los módulos es extraer una información específica del fichero multimedia y guardarla en formato XML.

Se han definido dos tipos de ficheros XML. Uno de ellos, el asociado a los ficheros multimedia (generado por el subsistema de adquisición de contenidos) sirve para describir el contenido de ese fichero, y en él se guarda información como los subtítulos, la trascripción, los anuncios, etc… cada uno de ellos acompañado por su tiempo de inicio y fin. El segundo tipo de ficheros XML sirve para almacenar información sobre las agrupaciones (*clustering*) que se pueden dar en distintos ficheros (por ejemplo repeticiones de anuncios). Mientras los ficheros asociados al documento multimedia sirven para describirlo, el segundo tipo de documentos XML sirve para indexar la base de datos multimedia en un tema concreto. Estos mismos ficheros XML son usados para pasar información entre los distintos módulos.

Actualmente se está trabajando con las capturas audiovisuales emitidas a través de TDT. Los módulos que se han puesto en marcha o que están en vías de desarrollo son:

1) Extracción de subtítulos. Los textos del teletexto y OCR se guardan en el XML asociado al contenido multimedia.
2) Reconocimiento de voz. Permite transcribir lo que se está diciendo.
3) Detección de anuncios. Se recogen los distintos anuncios que hay en la grabación.
4) Clasificación de anuncios. Se detecta cuantas veces se repite un anuncio. Esta información se almacena en un fichero de *clustering* de anuncios.
5) Detección de momentos estelares en partidos de fútbol locutados. Determina en qué momentos se da alguna situación excitante a lo largo de un partido de fútbol cuando está retransmitido por un locutor.
6) Diarización de locutor. Segmenta la señal en distintos trozos que se corresponden a distintos locutores.
7) Identificación de locutor. Permite identificar el locutor que está hablando en cada momento.

A partir de la información recogida con estos módulos se pueden construir un gran número de aplicaciones. Por ejemplo, de los momentos estelares de partidos de fútbol se pueden generar resúmenes automáticos de los partidos. De la detección de anuncios repetidos se pueden contabilizar cuantas veces aparece un anuncio, a que horas y en que canales. Los subtítulos y la transcripción automática se pueden usar para hacer búsquedas de contenidos o "*topic detection*". A partir de la identificación del locutor se puede buscar las apariciones de este locutor en los distintos medios, etc.

V. Web de consulta de los resultados de la indexación

Para la visualización de los resultados del sistema de indexación se ha optado por la creación de una aplicación web con contenido dinámico, ya que esta tecnología permite dotar a las páginas de la riqueza visual de las aplicaciones de escritorio, sin perder las ventajas de despliegue de la tecnología web. Debido a que el sistema ha de ser escalable y permitir la rápida incorporación de las nuevas tecnologías de indexado, se ha desarrollado utilizando el lenguaje Ruby (junto con su *framework* de desarrollo web, Rails) por permitir éste un desarrollo ágil y rápido. Las facilidades que incorpora Rails para la utilización de la tecnología Ajax, permite dotar del dinamismo a la aplicación. El acceso a los ficheros XML con los resultados se realiza mediante consultas XPath, que permiten recorrer el árbol de nodos de forma sencilla y flexible, evitando tener que realizar un parseado diferente para cada nuevo tipo de consulta.

La web está dividida en diferentes secciones, organizadas acorde a las diferentes tecnologías de indexado. Dentro de cada sección, existen diferentes subsecciones o vistas, que permiten acceder a la información de indexado desde distintas perspectivas. Como casos ilustrativos, presentamos las vistas asociadas a las tecnologías de detección de anuncios y momentos estelares en los partidos de fútbol, como se puede ver en la figura 3.

En el caso de la sección de detección de anuncios, existen tres vistas. La primera de ella muestra la lista con los anuncios identificados, ordenada por el número de apariciones (esto nos permite evaluar la importancia de una campaña publicitaria). Una vez seleccionado un anuncio en concreto, podemos ir a la segunda vista, en la que se detallarán las apariciones del mismo en los distintos canales y franjas horarias (esto nos permite evaluar si la emisión de una campaña publicitaria se ajusta realmente con el contrato acordado con la cadena). Finalmente, la tercera vista nos permite explorar todas las grabaciones de la base de datos por fechas y canal, y visualizar los anuncios detectados automáticamente (esto permite validar el correcto funcionamiento del sistema de detección).

En el caso de la sección de detección de los momentos importantes en los partidos de fútbol, existen dos vistas. En la primera de ellas se muestra la lista de partidos de la jornada que el usuario escoja. El usuario puede elegir un partido de la lista, y de este modo accederá a la segunda vista. Esta segunda vista permite generar automáticamente un resumen del partido escogido, con sus momentos más importantes. El usuario puede elegir la duración del resumen, lo que permite ajustar el umbral que clasifica un momento del partido como importante o no.

Fig. 3. Vistas detalladas de las apariciones de un anuncio y del resumen de un partido de fútbol.

VI. Conclusiones

El sistema presentado en esta publicación permite aunar múltiples tecnologías de indexado bajo un mismo *framework*. Esto resulta muy adecuado para el desarrollo y pruebas de nuevos algoritmos de indexado, ya que provee de un marco común sencillo y suficientemente flexible para la incorporación de tecnologías muy heterogéneas, con un bajo impacto sobre los desarrollos ya realizados. Así mismo, la aproximación incremental y colaborativa entre las diferentes partes del sistema, permite escalar y abordar de forma adecuada las tecnologías de indexado más complejas.

Además de las tecnologías presentadas, actualmente se está trabajando en la incorporación de las tecnologías de diarización e identificación de locutor, centradas en el entorno de los telenoticias. Estas tecnologías permitirán indexar y localizar las intervenciones de los personajes de actualidad de forma automática. Combinando también la transcripción automática de voz, será posible registrar las declaraciones de cada personaje.

Referencias

[1] JL Gauvain, L Lamel, G Adda, “Transcribing Broadcast News for Audio and Video Indexing”, Transactions on communications of the ACM, volume 43, Issue 2, February 2000.

[2] S. Renals, D. Abberley, D. Kirby, and T. Robinson. “Indexing and retrieval of broadcast news”. Speech Communication, 32(1-2):5–20. 2000.

[3] S.J. Young, M.G. Brown, J.T. Foote, G.J.F. Jones and K. Sparck Jones, “Automatic Indexing for Multimedia Retrieval and Browsing”, Proc. Intl. Conf. Acoustics, Speech and Signal Processing ICASSP 97, 1:199-202. Munich, April 1997.

[4] J.Clark, S. DeRose, “XML Path language Version 1.0”, W3C Recommendation, 16 November 1999. http://www.w3.org/TR/xpath

[5] “H.264: Advanced video coding for generic audiovisuals services”, ITU-T Recommendation, November 1997.

[6] “Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television”, ETSI Recommendation, November 2004.