Videacademia: sistema de almacenamiento y consulta de videos educativos en HD mediante internet

Videacademia: storage and retrieval system of educational videos in HD via Internet

Abelardo Rodríguez-León,[1*] Irving Espinoza-Calvo,[2] Carlos J. Genis-Triana,[3] Hector A. Andrade-Gomez[4]

[1] Departamento de Sistemas y Computación, LCI, Instituto Tecnológico de Veracruz.
Miguel Ángel de Quevedo 2779, col. Formando Hogar, CP 91860. Veracruz, México
* Correo-e: 1arleon@itver.edu.mx

PALABRAS CLAVE:

sistemas web de altas prestaciones,
repositorio de videos HD,
videoteca educativa

RESUMEN

Existen en internet varios sistemas para el almacenamiento y reproducción de videos; sin embargo, casi ninguno tiene una forma eficiente de consulta de contenidos. Debido a esto, es difícil para el usuario encontrar la información que sea de su interés, pues primero debe revisar material que resulta irrelevante. Para solucionar esto se creó un sistema web que, además de administrar un depósito de videos de contenido académico, cuenta con una innovación que no está presente en sistemas similares: consiste en una base de datos que guarda los metadatos de los videos, la cual ayuda a realizar búsquedas más eficientes de los contenidos, en función de varios rubros, como pueden ser institución, tema, nivel educativo, autor, etcétera. Dicho sistema se llama Videacademia y tiene como objetivo servir a la comunidad académica para la explotación de información en forma de videos grabados en HD. Actualmente se encuentra en funcionamiento en fase alfa, con un contenido precargado de más de 300 videos, incluida toda la videoteca del Instituto Tecnológico de Veracruz, la cual ha sido digitalizada en alta definición para incorporarla a la plataforma.

KEYWORDS:

high performance web system,
HD video repository,
educational video library.

ABSTRACT

There are several systems for video data management and retrieval. However, almost all of them lack of an efficient way of searching content. This makes it difficult to find information because it is necessary to look at a huge amount of irrelevant information in order to find what it is needed. In order to solve this problem, we present a novel approach for video data management and retrieval. It consists on a web application that, in addition of managing an academic video repository, it also uses a data base with metadata which allows for much more efficient search using several criteria such as institution, topic, author, academic degree for which the video was made for, etc. The name of this system is Videacademia and it was made with the intention of being used by the academic community as a tool to get information in the form of high definition videos. The system is now in the alpha phase with more than 300 videos, including the video repository of the Instituto Tecnológico de Veracruz, which has been digitalized in its entirety and incorporated to the system.

1 INTRODUCCIÓN

Hace apenas unas décadas la educación no era tan accesible como ahora; el volumen de información disponible resulta inconmensurable. Las tecnologías de la información y comunicación proporcionan una gran oportunidad para mejorar los procesos educativos a través de esquemas modernos como es el caso del *e-learning*. De acuerdo con Sutton [1], "la eficacia del video en línea sólo es superada por la palabra de boca en boca en su capacidad para influir en las personas", y esto sólo es posible gracias a la utilización de videotecas digitales.

Una videoteca digital es un sitio donde se almacena y mantiene información digital para su consulta por las personas que así lo deseen. YouTube es una de las mayores videotecas que hay en Internet, es de muy fácil manejo, ya que el material está clasificado por categorías.

Los videos de esta plataforma son una extraordinaria fuente de información pero, a pesar de su auge, es desaprovechada para el proceso de enseñanza-aprendizaje. En general, las personas están acostumbradas a buscar información textual, al estilo de un libro de instrucciones, cuando tratan de explicar o entender un procedimiento concreto; sin embargo, el video se compenetra mejor con este tipo de tareas porque permite visualizar las diferentes acciones necesarias para llevar a cabo las tareas.

YouTube ha crecido con gran rapidez, llega a más de 100 millones de clips vistos al día, con sólo un pequeño equipo responsable de escalar el portal. ¿Cómo gestionar ese tráfico para todos los usuarios? ¿Cómo ha evolucionado desde que Google adquirió el sitio? ¿Cuál es el secreto para soportar tan alta cantidad de peticiones?

En un artículo publicado por Hoff [2] se describe la arquitectura del sitio de distribución de videos más importante del mundo. El artículo se basa en Cuong [3], quien explica cómo se consiguió que la arquitectura del sitio se adapte a un rápido crecimiento entre los meses de marzo y julio de 2006, cuando pasó de 30 a 100 millones de videos vistos al día. La clave de su éxito se basa en permitir un gran escalamiento del sitio, tanto a nivel de *software* como de *hardware*.

Según Henderson [4], el enfoque de escalabilidad utilizado por YouTube es el modelo que siguen varios servicios web importantes como Facebook, Flickr, entre otros; ya que tienen en común un alto grado de complejidad computacional debido al número de usuarios y a la gran cantidad de contenido que poseen, pero también deben permitir la evolución del sitio para poder hacer frente a esta situación.

Este enfoque de escalabilidad es el usado en el desarrollo del proyecto Videacademia, el cual pretende ser un portal innovador que permita almacenar, clasificar y visualizar en alta definición contenidos culturales, didácticos y pedagógicos de varias categorías de forma totalmente gratuita. Para un mejor desarrollo se ha dividido el proyecto en módulos, varios de los cuales se han desarrollado de manera progresiva en los años recientes.

2 DESARROLLO

Este proyecto tiene como base el conocimiento generado por el proyecto: "Sistema multiagente para distribución por demanda de video de alta definición sobre internet 2" (clave DGEST 2189.09P). Éste proporcionó los conocimientos necesarios para el tratamiento (captura, almacenamiento y distribución) de video de alta definición en internet 2 [5]; se usaron codificadores de video muy eficientes, como MPEG-4, sobre un *cluster* de computadoras para mejorar el almacenamiento de videos. Dicho proyecto se desarrolló de agosto de 2009 a julio de 2010.

Con estos conocimientos se planeó, presentó y desarrolló "Videacademia: repositorio y visualización de videos digitales educacionales en HD mediante internet e internet 2" (clave DGEST 4399.11-P). Dicho proyecto plantea la arquitectura general del sistema, define los módulos necesarios para su funcionamiento y desarrolla versiones beta (funcionales solamente, sin adecuada validación) de varios de ellos.

El método de ingeniería de *software* empleado fue la división modular y la integración progresiva por prototipos funcionales. Para ello se dividió el núcleo del proyecto en cuatro módulos, que se pueden observar en la figura 1 y que se describen mas adelante. Cada unos de estos módulos fue asignado y desarrollado por alumnos en proyectos de residencias profesionales.

2.1 Módulo de captura de video

En la figura 1 se nota el módulo de captura de video en los cuatro cuadros de la parte izquierda. Hay diversas alternativas para que el usuario pueda subir videos al sistema. Este elemento consistió en un proceso ad-

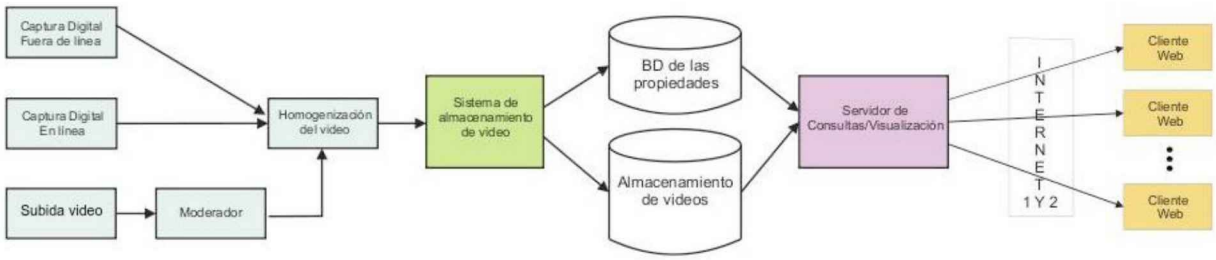

Figura 1. Arquitectura general del sistema Videacademia

ministrativo que indica cómo se deben codificar los videos a alimentar en el sistema dependiendo las diversas fuentes de que procedan (permanece en fase beta). Las fuentes de videos pueden ser por captura digital fuera de línea (actualmente se ocupa para anexar videos en VHS); captura digital en línea (es una opción que permitirá capturar video en tiempo real y subir al sistema), y por subida a través de la red (solo para usuarios registrados).

2.2 Módulo de homogeneización de video

Proporciona los métodos de compresión de video (codificación) para que su transmisión sea más efectiva en función de la demanda. Como este trabajo tiene un alto costo computacional, se apoya en el uso del *cluster* nopal para llevar a cabo su labor y poder codificar el video (de ser necesario de manera continua), para después almacenarlo en el repositorio.

Para realizar la codificación se usa el *software* libre ffmpeg con *codec* X264 para obtener todos los videos suministrados al sistema en alta definición. El ffmpeg se preparó (fase beta) para trabajar en un algoritmo paralelo multinivel que corre en los nodos del *cluster*.

2.3 Módulo sistema de almacenamiento

Este módulo consiste en dos componentes importantes, el primero de ellos es el "almacenamiento", para el que fue necesario la configuración de un RAID que sirve como depósito masivo de los videos.

Se eligió un sistema RAID 0 porque las necesidades actuales del proyecto van orientadas a guardar la mayor cantidad posible de información, y no tanto a contar con un sistema de almacenamiento con redundancia y monitoreo. RAID 0 maximiza el uso de los discos duros, ya que suma las capacidades de todos los discos independientes para formar una sola unidad virtual con una capacidad de almacenamiento conjunta muy superior a la de cualquier otro RAID, o a utilizar todos los discos por separado.

Después de implementar el RAID, se le evaluó simulando un ambiente de producción. Para hacer las pruebas se usó la aplicación Palimpsest. Los resultados obtenidos indican, de forma apabullante, que es mucho mejor utilizar RAID en lugar de discos independientes, ya que demostró ser entre 1.5 y 6 veces más rápidos que los discos en varios parámetros.

El segundo componente del sistema de almacenamiento es la "base de datos (BD) de las propiedades". En este componente se desarrolló e implementó un esquema completo de base de datos para la clasificación y administración de información, tanto de usuarios como de las características de los videos. Se puede considerar que este componente es el núcleo del sistema.

Se requirió el desarrollo de un modelo de BD correctamente definida y de rápida respuesta para poder hacer frente a un proyecto con las características planteadas por Videacademia. El diseño incluyó la construcción del "diccionario de datos", el "modelo entidad-relación", el "modelo relacional" y el "modelo físico de la base de datos".

Para la implementación se eligió el sistema gestor de base de datos MySQL, ya que presentó una mayor cantidad de características positivas para el proyecto, tanto para cumplir los objetivos a corto plazo, como para su posible crecimiento a futuro. En cuanto al almacenamiento de las tablas, se eligió InnoDB como motor principal, en lugar de MyISAM, debido a las ventajas funcionales que proporciona el primero. El motor MyISAM deja al sistema operativo la tarea de administrar la caché de lecturas y escrituras de los

Tabla 1. Resumen de resultados de la evaluación de la base de datos

Tipo de operación	Registros	Operaciones	Tiempo (seg.)	Cores	CPU promedio (%)	Motor
Inserción	251,494	10,000	75.34	3	6.90	InnoDB/MyISAM
Búsqueda simple (una palabra)	261,495	1,000	58.27	8	100.00	InooDB
Búsqueda simple (una frase)	261,495	1,000	27.48	8	100.00	MyISAM
Modificación campo status	261,495	10,000	6.29	8	4.78	InooDB
Borrado físico	261,495	10,000	0.649	8	51.80	InnoDB/MyISAM

registros, mientras que InnoDB realiza él mismo la tarea combinando cachés de registro y de índice. Una característica positiva de InnoDB es que almacena físicamente los registros con base en la clave primaria, mientras que MyISAM los guarda en el orden en que fueron añadidos. Finalmente, InnoDB se recupera de un problema volviendo a ejecutar sus *logs*, mientras que MyISAM necesita revisar todos los índices y tablas que hayan sido actualizados y reconstruirlos si los cambios no han sido escritos en disco. Por lo tanto, en cuanto a recuperación de problemas, InnoDB requiere siempre el mismo tiempo aproximadamente, mientras el utilizado por MyISAM aumenta con el tamaño de la base de datos en la recuperación. Se usó el motor de almacenamiento MyISAM como motor secundario para emplear una tabla espejo, solamente usada para búsqueda. Esta tabla lleva una copia de los videos semiindependiente de las tablas de la base datos principal.

En la fase de implementación de la BD se hicieron pruebas del esquema elegido. El gestor MySQL sólo ofrece tres métodos. Se encontró que el primero era muy restrictivo; el segundo limita las búsquedas al uso de una sola palabra clave y no permite el uso de frases enteras, y el tercer método permite el uso de frases. Para resolver el problema, se empleó una combinación de las últimas dos alternativas, las cuales se activarán dependiendo si se utiliza una palabra o frase de búsqueda.

Para la evaluación de la BD se emplearon las herramientas mysqlslap y monitor del sistema, las cuales daban a conocer tiempos de respuesta del servidor, así como el consumo total de CPU que realiza el servidor bajo situaciones estresantes. Como se puede ver en la tabla 1, algunas operaciones de BD, sobre todo de escritura y búsqueda de información, requirieron muchos recursos, incluso el cien por ciento del consumo de procesador, por lo que dieron tiempos de respuesta de alrededor de 1 minuto para las operaciones que se realizaron simulando 100 usuarios simultáneos y 10,000 o 1,000 operaciones de cada tipo.

2.4 Módulo servidor de consultas/visualización

Se desarrolló una interfaz gráfica *web* por completo, para que los diversos usuarios pudieran acceder a ella de forma remota, tanto para subir videos como para realizar consultas y visualización, dependiendo del rol o privilegios elegidos.

Para el sistema se crearon varios perfiles de usuarios con diferentes atributos cada uno: *Anónimo*, *Registrado* y *Administrador*; cada uno de ellos con diferentes capacidades de manejo del sistema. El trabajo desarrollado en este módulo fue crear las interfaces web necesarias para que cada tipo de usuario pudiera acceder con sus correspondientes limitaciones a los datos del sistema. Para ello se utilizó PHP con mySql y Javascript con hojas de estilo; se instaló y configuró un servidor web normal ssh remoto para desarrollo y pruebas remotas, y también se utilizó MySql Workbench para facilitar el manejo de la BD.

3 RESULTADOS

El avance de este proyecto se puede considerar, en términos generales, en prototipo fase alfa. Se han concretado varios periodos de publicación en modo prueba en internet. A la fecha se han cargado más de 300 videos con sus respectivas clasificaciones.

Actualmente se encuentra funcionando y disponible de manera pública en la dirección http://sgca.itver.edu.mx/videacademia Se está realizando una carga inicial de videos desde noviembre de 2012. La página inicial se puede observar en figura 2.

Figura 2. Interfaz web inicial del sistema Videacademia

4 CONCLUSIONES Y TRABAJOS FUTUROS

Este proyecto demuestra la posibilidad real de crear y poner en funcionamiento proyectos que en otros contextos son éxitos comerciales, como es el caso de un repositorio de videos similar a YouTube. Por supuesto, nuestro objetivo no es competir con los gigantes de los medios, ya que nuestra infraestructura y público meta difieren mucho. Nuestra finalidad es crear un sistema que sirva para almacenar de forma organizada contenidos educativos en video para todos los niveles, principalmente en habla hispana.

La característica que distingue este proyecto, y que puede ser tanto un factor a favor como en contra, es la posibilidad de agregar los metadatos de los videos; es decir, información relevante que permita buscar y clasificar los contenidos por: autor, tema, nivel, escuela, materia, etc. Puede ser favorable porque permite hacer una búsqueda más limpia y organizada, dejando fuera muchos videos que no son del interés del usuario. Puede ir en contra ya que requiere que los usuarios capturen esos datos, sin los cuales, las búsquedas de contenidos serían infructuosas o poco útiles, como suele suceder en YouTube, por ejemplo, cuando al buscar por una o más palabras, a veces se presentan muchos videos que no tienen relación con lo que se busca.

Hay muchas ideas de mejoras en torno a este sistema; por ejemplo, implementar un modelo de negocios que nos permita hacerlo autofinanciable; evaluar el uso de la interfaz web actual o implementarle características técnicas novedosas (como transmisión en vivo de videos o suministro de video acorde al ancho de banda). Si bien técnicamente son posibles estas mejoras, es preciso tener cautela porque requieren tiempo y recursos.

REFERENCIAS

1. Sutton, Adam T. Content Marketing: Videos attract 300% more traffic and nurture leads. En *Marketing* Sherpa, disponible en: http://www.marketingsherpa.com/sample.cfm?ident=32075. Publicado el 14 de Diciembre de 2011.
2. Hoff, Tod. YouTube Architecture. En *High scalability*, disponible en: http://highscalability.com/youtube-architecture, publicado el 12 de Marzo de 2008.
3. Cuong, Do. "Seattle Conference on Scalability: YouTube Scalability". Disponible en http://video.google.com/videoplay?docid=-6304964351441328559. Publicado el 23 de Junio de 2007.
4. Henderson, Cal. *Building Scalable Web Sites*. California: O´Reilly, 2006.
5. Mares-Rosas, L. A., Rodríguez-León, A., Rivera-López, R. Modelado de agentes de *software* para un sistema de videoconferencia en tiempo real bajo Internet 2. *Congreso Internacional de Informática y Computación – ANIEI,* 2010, 455-460.

Acerca de los autores

Abelardo Rodríguez León es Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Veracruz (1990), Maestro en Ciencias Computacionales por la Universidad Veracruzana (1996) y Doctor en Ciencias Computacionales por la Universidad Politécnica de Valencia, España (2007). Actualmente es profesor investigador en el Departamento de Computación y Sistemas del Instituto Tecnológico de Veracruz. Sus áreas de interés incluyen: la computación de alto rendimiento, paralelismo y *grid*, además de estudios de modelos gráficos en 3D.

Irving Espinoza-Calvo es Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Veracruz (2013). Actualmente se desempeña como desarrollador independiente. Sus áreas de interés incluyen: diseño y desarrollo de aplicaciones web, diseño de bases de datos y programación en Java.

Héctor Andrade Gómez es Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Veracruz (1985), Maestro en Ciencias Computacionales por el Instituto Tecnológico y de Estudios Superiores de Monterrey, campus Morelos (1992) y Doctor en Ciencias Computacionales por la Universidad de Florida, Estados Unidos (2001). Actualmente es profesor investigador en el Departamento de Sistemas y Computación del Instituto Tecnológico de Veracruz. Sus áreas de interés incluyen: lenguajes de programación, cómputo móvil y desarrollo web.

Carlos Julián Genis-Triana es Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Veracruz (2002) y Maestro en Ciencias Computacionales por la misma institución (2006). Actualmente es profesor del Departamento de Sistemas y Computación y Jefe del Centro de Computo del Instituto Tecnológico de Veracruz. Sus áreas de interés incluyen: lenguajes de programación, cómputo distribuido, seguridad informática y desarrollo web.