SISTEMA PARA LA RECUPERACIÓN DE NOTICIAS DIGITALES

Ing. Aramis Romero Carballea[1]. aramis@uci.cu
Ing. Susana Yaque Rivera[1]. susanay@uci.cu
Ing. Felix Ivan Romero Rodríguez[1]. firomero@uci.cu
Ing. Yanary Hernández Sosa[1]. yanary@uci.cu

[1] Departamento Señales Digitales. Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km. 2 ½. Torrens, municipio de La Lisa. La Habana, Cuba.

RESUMEN

Las agencias de noticias tienen como misión principal la divulgación de la información, por lo que la recuperación de las noticias y todos sus componentes asociados es de vital importancia. En Cuba existe la Agencia de Información Nacional (AIN) encargada de divulgar la información a través de diversos medios. En esta agencia no es posible obtener la noticia completa, siendo esto un elemento importante para el redactor encargado de la elaboración de la noticia. Precisamente el presente trabajo está enmarcado en el objetivo de desarrollar un sistema de recuperación y extracción completa de las noticias publicadas en Internet, permitiendo su posterior almacenamiento en un sistema gestor de base de datos. Para el desarrollo de la misma se seleccionó un conjunto de tecnologías y herramientas de acuerdo con las políticas de software libre del país.

Palabras Clave: búsqueda, extracción, noticias, recuperación.

INTRODUCCIÓN

En la actualidad se pueden encontrar disimiles agencias de noticias cuyo propósito fundamental es mantener una extensa red de corresponsales en todo el mundo y divulgar sus servicios, ejemplo Deutsche Presse-Agentur (DPA), Associated Press (AP), Agence France-Presse (AFP), Reuters (REU). En Cuba existe la Agencia de Información Nacional (AIN) la cual tiene como objetivo principal difundir el acontecer noticioso nacional e internacional. Para lograr la difusión de las noticias se apoya en diversas fuentes, dígase sitios Web, periódicos nacionales y agencias de prensas. Además posee los siguientes medios para la propagación de los contenidos: radio, un canal de televisión y un sitio web.

En la AIN no disponen de algún mecanismo que les permita lograr una completa obtención de la noticia y sus elementos asociados para su posterior procesamiento y difusión por los medios. Por esta razón, se trazó para la presente investigación el siguiente objetivo general: desarrollar una aplicación que permita automatizar la recuperación, extracción y almacenamiento de las noticias y sus componentes asociados para la AIN para lograr garantizar los procesos de gestión de la noticia llevado a cabo por los editores de la agencia.

MATERIALES Y MÉTODOS

Antes de comenzar con la recuperación, extracción y almacenamiento de las noticias y debido a la necesidad que tienen los editores de la agencia de estar

actualizados del acontecer nacional e internacional, se hizo necesario estudiar los formatos de redifusión de contenidos web pues estos son una forma de organizar la información que se encuentra en la web. La redifusión hace posible el filtrado de las publicaciones de páginas que son de interés para cada usuario (aquellas en las que se suscribe) y, mediante un software específico, le pueden llegar las nuevas noticias publicadas de su interés, evitando tener que visitar un número, en ocasiones, demasiado extenso de páginas web para comprobar si se han producido actualizaciones. Es decir produce un gran ahorro de tiempo, ya que es posible acceder rápidamente a todos los contenidos nuevos publicados en varios sitios, sin tener que visitarlos uno por uno (1).

Las dos principales familias de formatos de redifusión web son Atom y RSS (2). El Formato de Redifusión Atom es un fichero en formato XML usado para Redifusión web y RSS son las siglas de Really Simple Syndication, un formato XML para sindicar o compartir contenido en la web.

Luego de un exhaustivo estudio sobre ambos formatos se seleccionó RSS para el desarrollo del sistema que se propone en la presente investigación debido a que está más desarrollado y publicado que la Atom lo cual es un elemento determinante en cuanto al tema de redifusión de contenidos en la web. Con el uso de RSS el usuario puede recibir y clasificar rápidamente información de diferentes sitios, garantizando así una mayor cantidad de información en menos tiempo, lo cual es de vital importancia para la AIN en su afán de obtener resultados factibles en el proceso de divulgación de las noticias a través de los medios de difusión de los cuales dispone. Una vez definido el formato de redifusión a emplearse seleccionó la herramienta para la recuperación de la información.

Para lograr la recuperación de la información en internet existen varias herramientas, ejemplo los metabuscadores y los motores de búsquedas o buscadores. Estas herramientas garantizan una búsqueda de manera rápida y dinámica. Hasta el momento no existe un metabuscador específico que sea capaz de interactuar con los servicios web, es decir utilizan los servicios web de los motores de búsqueda que se indexan en su contenido. Por lo que sería redundante y poco eficaz, utilizar un metabuscador en un sistema que cuente en su implementación con la asociación de algún motor de búsqueda. Por tal motivo se descarta esta herramienta en la presente investigación y se procede a estudiar los motores de búsqueda.

Buscadores o motores de búsqueda.

Los buscadores o motores de búsqueda son aplicaciones robustas que manejan grandes bases de datos de referencias a páginas web recopiladas por medio de un proceso automático, sin intervención humana. Uno o varios agentes de búsqueda recorren la web, a partir de una relación de direcciones inicial y recopilan nuevas direcciones generando una serie de etiquetas que permiten su indexación y almacenamiento en la base de datos. Un motor no cuenta con subcategorías, sino con avanzados algoritmos de búsqueda que analizan las páginas que tienen en su base de datos y proporcionan el resultado más adecuado a una búsqueda. También almacenan direcciones que les son remitidas por los usuarios.

Actualmente existen muchos motores de búsqueda pero los principales son: *Yahoo, Bing* y *Google*. Estos son fáciles de usar, rápidos, fiables y además los desarrolladores no tienen que preocuparse en añadir datos estructurados para cada buscador sino que prevalece un lenguaje común para todos permitiendo optimizar el indexado de las páginas y, por tanto, mejorar el posicionamiento en las páginas de resultados (3)**.**

Acorde al objetivo actual que se persigue para la aplicación, el uso de *Yahoo* es muy pobre, presenta escasa documentación y una muy pobre actualización de sus servicios, dejando depreciados la mayoría de ellos. Por otro lado se descarta el uso de *Bing* debido a medidas políticas tomadas por *Microsoft* sobre este buscador, dejando inactiva la prestación de sus servicios. Por lo que, debido a lo investigado a profundidad, se decidió analizar y caracterizar los servicios del potente motor de búsqueda de Internet, *Google*.

Google es un fuerte y potente motor de búsqueda, pero sin embargo se guiará la integración de dicho motor con el presente sistema hacia la principal característica que identifica el mismo, la Búsqueda de Noticias.

RESULTADOS Y DISCUSIÓN

Se alcanzó una solución integral capaz de gestionar todos los procesos de recuperación de la información noticiosa de la AIN. Dicho sistema da solución al problema de redacción actual en la agencia, automatizando así un conjunto de funcionalidades que hasta el momento se vienen realizando manualmente.

Recuperación de la información

A continuación se explican las dos vías encontradas para garantizar la recuperación de la información a través del sistema que se propone. Una es haciendo uso de un lector RSS y la otra variante es a través de la integración de Google como motor de búsqueda.

1ra variante- Casi todos los sitios web en Internet cuentan en su diseño con la integración de un canal RSS, para lograr así mantener actualizado al usuario en el mundo noticioso de una manera eficiente y rápida. A través de la dirección de un sitio web, se realiza la recuperación de las noticias que están publicadas en dicha dirección para adquirir un listado con los titulares de las noticias.

2da variante-Integrar un motor de búsqueda (Google con su servicio web de Google Search Ajax API) por medio del cual el usuario puede buscar cualquier tema de su interés y el sistema por su parte busca mediante una URL todas las noticias que se relacionen con la petición del usuario. Como resultado de este proceso se obtiene una lista con los titulares de las noticias encontradas.

Por cualquiera de las variantes anteriores el acceso a la información para su posterior recuperación es a través de una dirección URL y se obtiene un listado con los titulares de las noticias. Sin embargo para lograr la extracción de la información esto no es suficiente.

Extracción de la información

El formato de redifusión RSS está compuesto por el formato XML el cual brinda una estructura organizada que favorece a la hora de acceder a la información contenida en el mismo. En este caso solo se desea obtener del XML el título y la fecha de publicación de la noticia. Para el caso del motor de búsqueda mediante la URL se obtienen una serie de resultados los cuales van a contener una fecha de publicación y un título. Teniendo extraídos el título y la fecha de publicación se pasa a extraer los demás datos de la noticia. Por cualquiera de las variantes (lector RSS y motor de búsqueda), se opera de manera similar teniendo siempre la URL.

Almacenamiento de la información en el sistema.

Una vez extraída toda la información referente a la noticia por medio de las dos vías mencionadas anteriormente se procede a almacenar la noticia en la estructura definida para esta función, en este caso, una base de datos operacional. La misma contará con 5 tablas para el almacenamiento de los datos, una primera tabla para almacenar todos los datos referentes a las noticias, una segunda tabla para guardar todo lo referente a las fuentes web agregadas por el usuario al sistema y otras tres tablas las cuales dependen de la tabla noticia para almacenar las medias. En este caso, a la hora de almacenar las medias se guardará la dirección local de la media una vez haya sido descargada, esta dirección indicará a que noticia pertenece la media almacenada.

Dificultades en la Evaluación de los sistemas de Recuperación de Información

1. El texto no tiene estructura clara y no es fácil de analizar (4).
2. No existe “la respuesta correcta”. Cada documento puede ser más o menos relevante, y esto varía según el usuario y la situación. (4)
3. La velocidad y espacio importa, pero más la calidad de la búsqueda. (4)
4. La Recuperación de Información busca una aproximación a la respuesta sobre lo que el usuario busca. (4)

En la Recuperación de Información se buscan los documentos relevantes, por lo que la búsqueda exhaustiva es impracticable. Las técnicas automáticas de análisis de contenido y de clasificación de los documentos así como los métodos de evaluación de la efectividad de la recuperación constituyen los tres pilares sobre los que se fundamenta el desarrollo de los sistemas de recuperación de información. La dificultad estriba no solo en la extracción de la información sino también en la determinación de su relevancia. El propósito de las estrategias de recuperación automática consiste en obtener *todos* los documentos relevantes al mismo tiempo que obtener la *menor cantidad posible* de documentos irrelevantes (5).

Efectividad de la Recuperación - Precisión y Exhaustividad.

$$Exhaustividad = \frac{Número\ \ de\ \ documentos\ \ relevantes\ \ recuperados}{Número\ \ Total\ \ de\ \ documentos\ \ relevantes}$$

$$Precision = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ \ de\ documentos\ recuperados}$$

Evaluación del sistema.

En el Sistema Recuperador de Noticias Digitales se recuperó una muestra de 100 documentos HTML para procesar, de estos, 10 resultaron documentos irrelevantes, o sea, documentos que su URL no se pudo procesar y por tanto fue imposible obtener sus datos. De los 90 restantes, 80 fueron recuperados, dando aproximadamente un 88% de exhaustividad. Por otra parte, se obtuvo otro 80% de precisión. Ambos factores de evaluación cuentan con un elevado índice de porcentaje, mientras mayor sea la precisión y exhaustividad, mayor será la factibilidad y escalabilidad del sistema.

CONCLUSIONES

La evaluación de un sistema recuperador de la información, en este caso, el propio Sistema Recuperador de Noticias Digitales permitió determinar su factibilidad y precisión para cumplir sus objetivos llegando a cubrir las necesidades de información existentes en la AIN.

BIBLIOGRAFÍA

1. Torre, Aníbal de la. Web 2.0 en Educación. 4.1 La redifusión -"sindicación"- de contenidos web. [En línea] 2006. http://platea.pntic.mec.es/vgonzale/web20_0809/conten/temas/Tema_4.1.htm#arriba.
2. Wikipedia. Redifusión web. [En línea] http://es.wikipedia.org/wiki/Redifusi%C3%B3n_web.
3. Velasco, JJ. Bitelia. Google, Bing y Yahoo! se unen para optimizar la indexación de webs. [En línea] 2011. http://bitelia.com/2011/06/google-bing-yahoo-unidos-optimizar-indexacion-webs..
4. Rijsbergen, C. J. van. Information retrieval. London: Butterworth : s.n., 1979.
5. Comeche, JA Martínez. Los Modelos Clásicos de Recuperación de Información. [En línea] 2006. eprints.ucm.es/5979/1/Modelos_RI_preprint.pdf.

www.sociedadelainformacion.com

Edita:

Director: José Ángel Ruiz Felipe
Jefe de publicaciones: Antero Soria Luján
D.L.: AB 293-2001
ISSN: 1578-326x