Generación de versiones especializadas de la Wikipedia

Mario García Moreira [(1)], Luis Mora Torres [(2)], Cristina Abad [(3)]
Ingeniero en Computación Especialización Sistemas Tecnológicos [(1)], Ingeniero en Computación Especialización Sistemas de Información [(2)], M.Sc., Profesor [(3)];
Facultad de Ingeniería de Electricidad y Computación (FIEC)
Escuela Superior Politécnica del Litoral (ESPOL)
Campus Gustavo Galindo, Km 30.5 Vía Perimetral
Apartado 09-01-5863
marioalbert85@gmail.com, luisma1985@gmail.com, cabad@fiec.espol.edu.ec

Resumen

En este trabajo se presenta una alternativa para generar una enciclopedia especializada con temas previamente definidos basada en la Wikipedia, usando la librería Hadoop y los servicios Web de Amazon para el procesamiento distribuido de la enciclopedia original, de tal manera que los resultados generados se puedan guardar en un dispositivo portable para ser consultados en cualquier computadora sin la necesidad de conectarse a Internet.

Palabras Claves: *Wikipedia, Hadoop, Mapreduce, Enciclopedia Portable, Educación.*

Abstract

In this work, is presented an alternative to generate an specialized encyclopedia with topics previously defined based on Wikipedia, using Hadoop library and the Amazon Web Services for the distributed processing of the original encyclopedia, so the generated results could be saved in a portable device to be consulted in any computer without a connection to Internet.

1. Introducción

La Wikipedia es una enciclopedia libre que se ha convertido en uno de los recursos más consultados para quienes tienen la posibilidad de conectarse a Internet. Sin embargo su acceso debe ser en-línea, dejando a un lado a quienes por motivos de orden geográfico o económico no pueden acceder a ella.

Como alternativa, la Wikipedia permite descargarse todo el contenido de la misma, pero la gran cantidad de artículos que tiene en su sistema hace que sea—para fines prácticos—imposible descargarla para un usuario común.

El presente trabajo se ofrece una alternativa para generar una versión personalizada de la Wikipedia, de tal manera que los resultados generados se puedan guardar en un dispositivo portable para ser consultados en cualquier computadora sin la necesidad de conectarse a Internet.

2. Descripción del problema

Internet se ha convertido en uno de los principales recursos didácticos tanto para estudiantes como para profesores de escuelas. Sin embargo, en nuestro país existen muchas familias, escuelas y centros comunitarios que tienen computadores personales con características básicas, pero sin conexión a esta red.

Si bien en la actualidad, es posible conectarse a Internet a través de una serie de medios, incluyendo DSL, cable modem, y servicios celulares, en la práctica, el costo mensual de dichos servicios los hace difíciles de conseguir para muchos sectores. En el 2009, el ingreso de usuarios en Internet en el Ecuador oscila entre el 9% y 13% con una mayor concentración de proveedores de Internet en las ciudades de Quito y Guayaquil. [1]

3. Análisis y Diseño

El procesamiento de los datos de entrada está compuesto principalmente por dos algoritmos: el de Selección de Artículos y el de Unión y Limpieza de artículos.

Selección de Artículos.- Este algoritmo se encarga de examinar los artículos de forma individual e incluir aquellos que cumplan con los criterios de búsqueda, de acuerdo a la lista de temas de interés, y crear un archivo que contiene los títulos de los artículos que fueron seleccionados de acuerdo con los siguientes criterios:

- Tema está contenido en el título del artículo.
- Las categorías que contiene el Artículo están asociadas por lo menos una de las categorías a las que está relacionado el Tema.
- Si el Artículo seleccionado es una página de redirección entonces el Artículo al cual es redirigido es agregado a la lista de Artículos seleccionados.
- En el caso de que el Artículo no haya sido seleccionado debido a que ninguno de los Temas tiene relación ni con su título ni con sus categorías y este sea una página de redirección a otro artículo, entonces se verifica que el artículo al cual es redirigido tenga relación con alguno de los temas buscados. En el caso de ser así, entonces tanto el Artículo original como el que se redirigió son agregados a la lista de Artículos seleccionados.

Unión y limpieza de artículos.- Este algoritmo se encarga de juntar en un solo archivo los datos XML de cada página seleccionada y limpiar todos los elementos redundantes dentro de los artículos, entre los cuales se encuentran:

- Links hacia artículos que no fueron seleccionados dentro del primer algoritmo.
- Links a referencias de pies de páginas.
- Links de idiomas.
- Links externos.

3.1 Primera fase.

El primer proceso MapReduce para generación de la versión especializada de la enciclopedia tiene como entradas el dataset de la Wikipedia en línea y la lista de Temas de interés y sus categorías asociadas. Este trabajo se encuentra repartido en dos funciones una Map y otra Reduce, la función Map contiene el algoritmo de Selección de Artículos mencionado anteriormente y la función Reduce se encarga de unir los títulos de los artículos seleccionados en un solo archivo el cual es la salida de esta fase.

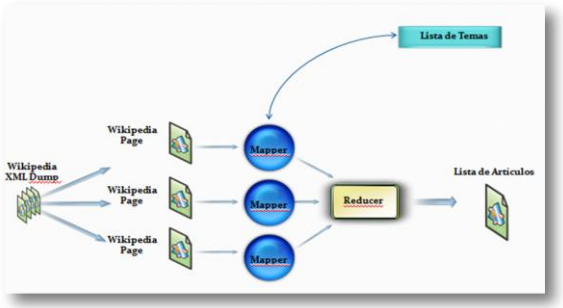

Figura 1. Primer proceso mapreduce para la selección de artículos.

3.2 Segunda Fase.

Esta fase tiene como entrada la lista de artículos seleccionados en la primera fase y el dataset de la Wikipedia. Está compuesta por dos funciones: una Map que contiene el algoritmo de limpieza mencionado anteriormente y una Reduce que se encarga de unir todos los artículos limpios de datos redundantes en un archivo XML con el mismo formato del dataset original.

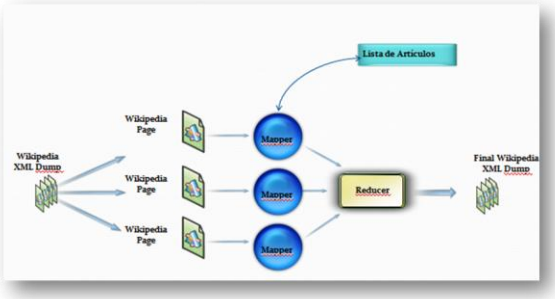

Figura 2. Segundo proceso mapreduce para la unión y limpieza de artículos.

4. Implementación

Se creó la librería WikiGen para la implementación de los algoritmos previamente descritos. Está integrada por clases de Java que contienen la funcionalidad de la misma. Así también se hizo uso de las librerías de Hadoop, Cloud9 y además del uso de expresiones regulares para las búsquedas dentro de los artículos. Con este propósito las librerías java.util.regex.Pattern y java.util.regex.Matcher de Java han sido utilizadas.

También se hizo uso de una librería externa llamada Cloud9. Esta es una librería en lenguaje Java que incluye clases para la lectura de datos de diversos datasets para ser procesados de manera distribuida usando el Framework Hadoop. Contiene un paquete para la lectura de artículos de Wikipedia desde su dataset y construir un modelo de objetos para que la manipulación de las diferentes secciones de los artículos dentro de las funciones Map en un trabajo MapReduce. Han sido usados específicamente los paquetes edu.umd.cloud9.collection.wikipedia y edu.umd.cloud9.util.

Una vez generado el archivo XML con los artículos de la Wikipedia utilizando los algoritmos de selección y limpieza, este es cargado dentro de un ambiente que en forma previa ha sido preparado para que el archivo sea leído y presentarlo al usuario en una interfaz sencilla y comprensible.

Este ambiente está conformado por el servidor portable llamado Server2Go [2] que contiene las librerías de PHP, una base de datos MySql Lite y el MediaWiki [3] que permiten procesar y leer el archivo XML generado con anterioridad, transformándolo en un modelo relacional para ser presentado en un entorno gráfico. Todo esto es la infraestructura base que le permite al usuario final observar y buscar dentro de los datos generados de forma rápida y en una interfaz amigable.

Figura 3. Ejemplo de enciclopedia personalizada para un año de educación básica.

5. Pruebas y Resultados.

Los escenarios seleccionados para las pruebas fueron el segundo y tercer año de educación básica para los cuales se preparan las listas de los Temas y Categorías asociadas a cada año de estudio. Los temas fueron seleccionados en base a la Malla Curricular del Ministerio de Educación que se encuentra en la página Web de “Educar Ecuador” [4].

```
Numero natural;Matematica,Geometria,Aritmetica
Unidades y decenas;Matematica,Geometria,Aritmetica
Numero ordinal;Matematica,Geometria,Aritmetica
Dolar;Monedas,Ecuador
Teoria de numeros;Matematica,Geometria,Aritmetica
Semirrecta;Matematica,Geometria,Aritmetica
Cuerpo humano;Anatomia humana
Numero cardinal;Matematica,Geometria,Aritmetica
Suma;Matematica,Geometria,Aritmetica
Resta;Matematica,Geometria,Aritmetica
Conjunto;Matematica,Geometria,Aritmetica
...
...
```

Figura 4. Extracto de un archivo de temas de interés y categorías relacionadas.

Las entradas para las pruebas fueron:

- El dataset de la Wikipedia en español.
- Los archivos de temas y categorías asociadas. Uno con 30 temas y otro con 50.

El Dataset de la Wikipedia fue subido al S3 de Amazon y para procesarla con nuestra librería se levantaron clústers en EC2. Desde el nodo maestro se descargó directamente del Amazon S3 el dataset y se subieron al nodo maestro el jar de la librería y los archivos de los temas para posteriormente cargar todo en el HDFS.

Se hicieron pruebas con 3 y 7 nodos para dos diferentes archivos de categorías. El tiempo total de ejecución es la suma de los tiempos individuales de cada uno de los algoritmos previamente presentados.

Tabla 1. Resultado con 3 nodos esclavos y 30 Temas

# Tareas map	# Tareas reduce	Tamaño XML generado	Tiempo	Proceso
26	1	1.2 MB	18m 36s	
			13m 4s	Select articles
			5m 32 s	Clean articles

Tabla 2. Resultado con 3 nodos esclavos y 50 Temas

# Tareas map	# Tareas reduce	Tamaño XML generado	Tiempo	Proceso
26	1	4.3 MB	22m 55s	
			16m 18s	Select articles
			6m 37s	Clean articles

En las pruebas con tres nodos se pudo apreciar que hubo un incremento en el tiempo de ejecución de 18 a 22 minutos aproximadamente entre las pruebas con 30 y 50 artículos, lo que se justifica ya que si se incrementa el número de temas a ser buscado se incrementa el tiempo para Unir el resultado y hacer la limpieza de los enlaces rotos.

Tabla 3. Resultado con 7 nodos esclavos y 30 Temas

# Tareas map	# Tareas reduce	Tamaño XML generado	Tiempo	Proceso
26	1	1.2 MB	10m 2s	
			7m 26s	Select articles
			2m 36 s	Clean articles

Tabla 4. Resultado con 7 nodos esclavos y 50 Temas

# Tareas map	# Tareas reduce	Tamaño XML generado	Tiempo	Proceso
26	1	4.3 MB	13m 2s	
			9m 57s	Select articles
			3m 12s	Clean articles

5. Conclusiones y Recomendaciones

5.1. Conclusiones

En este trabajo se desarrolló una herramienta que permite la Generación de Versiones Especializadas de la Wikipedia, las cuales pueden ser almacenadas en un CD o pen drive, para su posterior consulta fuera de línea.

La herramienta presentada, en combinación con un adecuado mecanismo de actualización y distribución, permitiría reducir los problemas de accesibilidad a información a la Wikipedia para quienes no tienen acceso a Internet.

Wikigen permite generar a los usuarios una versión personalizada de la Wikipedia con la información que necesitan para realizar sus tareas educativas, trabajos investigativos, datos informativos, etc.

5.2. Recomendaciones

En nuestro país la demanda de televisores y DVDs es mayor que la de computadoras, por ello una futura mejora que puede ser aplicada al Wikigen es la creación de una versión que permita almacenar su contenido dentro de un DVD y poder ser observado desde un televisor, con ello aseguramos un mayor acceso a la información llegando en esta forma a aquellos usuarios que no cuentan con un computador en sus hogares pero sí poseen un Televisor y un DVD. La posible interfaz que se propone es una desarrollada en un software con tecnología multimedia que permita generar contenidos que interactúen y reaccionen con el uso de un control de Televisión y DVD.

Para poder implementar las ideas descritas en el presente trabajo, habría que diseñar también un esquema de actualizaciones periódicas (por ejemplo

anuales) de las distribuciones personalizadas de la Wikipedia.

6. Agradecimientos

A Dios, a nuestros padres, a nuestros profesores y a nuestros más dilectos amigos por su incondicional apoyo y amplia colaboración en el curso de nuestra vida universitaria

12. Referencias

[1] Carrión, Hugo. *Internet, calidad y costos en Ecuador.* [En línea] 26 de Agosto de 2009. [Citado el: 2006 de Agosto de 2009.] http://www.imaginar.org/docs/internet_2009.pdf

[2] Server2Go - *Self configurable WAMPP Stack.* [En línea] [Citado el: 28 de Agosto de 2009.] http://www.server2go-web.de.

[3] MediaWiki/es. [En línea] *MediaWiki, The Free Wiki Engine*, 21 de Noviembre de 2007. [Citado el: 28 de Agosto de 2009.] http://www.mediawiki.org/wiki/MediaWiki/es..

[4] Reforma Curricular para la Educación Básica. [En línea] http://www.educarecuador.ec/_upload/Reformacurribasica.pdf.