ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"GENERACIÓN DE VERSIONES ESPECIALIZADAS DE LA WIKIPEDIA"

INFORME DE MATERIA DE GRADUACIÓN

Previo a la obtención del Título de:

INGENIERO EN COMPUTACIÓN ESPECIALIZACIÓN

SISTEMAS TECNOLÓGICOS

INGENIERO EN COMPUTACIÓN ESPECIALIZACIÓN

SISTEMAS DE INFORMACION

Presentado por:

Mario Alberto García Moreira

Luis Manuel Mora Torres

Guayaquil – Ecuador

AÑO 2009

AGRADECIMIENTO

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

DEDICATORIA

A todos los que aportan con su conocimiento y tiempo para llevar a este país adelante pese a todas las adversidades.

TRIBUNAL DE GRADO

_____

MsC. Cristina Abad R.

PROFESORA DE LA MATERIA DE GRADUACIÓN

_____

MSc. Xavier Ochoa

PROFESOR DELEGADO POR EL DECANO

DECLARACION EXPRESA

"La responsabilidad del contenido de este Proyecto de Graduación, nos corresponde exclusivamente; y el patrimonio intelectual de la misma, a la Escuela Superior Politécnica del Litoral"

(Reglamento de exámenes y títulos profesionales de la ESPOL)

Mario Alberto García Moreira

Luis Manuel Mora Torres

RESUMEN

En este trabajo se presenta una alternativa para generar una enciclopedia especializada con temas previamente definidos basada en la Wikipedia, usando la herramienta 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.

ÍNDICE GENERAL

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 presenta 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.

1. PLANTEAMIENTO DEL PROBLEMA

1.1 Antecedentes

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).

1.2 Wikipedia

Wikipedia es una enciclopedia en línea colaborativa y libre que contiene una gran cantidad de información al alcance de todos quienes tienen acceso a Internet.

Entre los servicios que pone a disposición esta enciclopedia en línea, se encuentra la posibilidad de ser descargada en su totalidad de manera gratuita. Esta funcionalidad permite consultarla fuera de línea. Lastimosamente, resulta difícil distribuir la Wikipedia en su totalidad, ya que por la gran cantidad de datos que contiene no entra en dispositivos externos de almacenamiento secundarios como CDs, DVDs y memorias flash. En Agosto del 2009 el tamaño de la Wikipedia comprimida en inglés es de aproximadamente 5.2GB y en español es 806MB (2). Adicionalmente, la tasa de crecimiento de la misma hace que su tamaño aumente considerablemente con el paso del tiempo.

Las Figuras 1 y 2 ilustran el crecimiento de la Wikipedia en español.

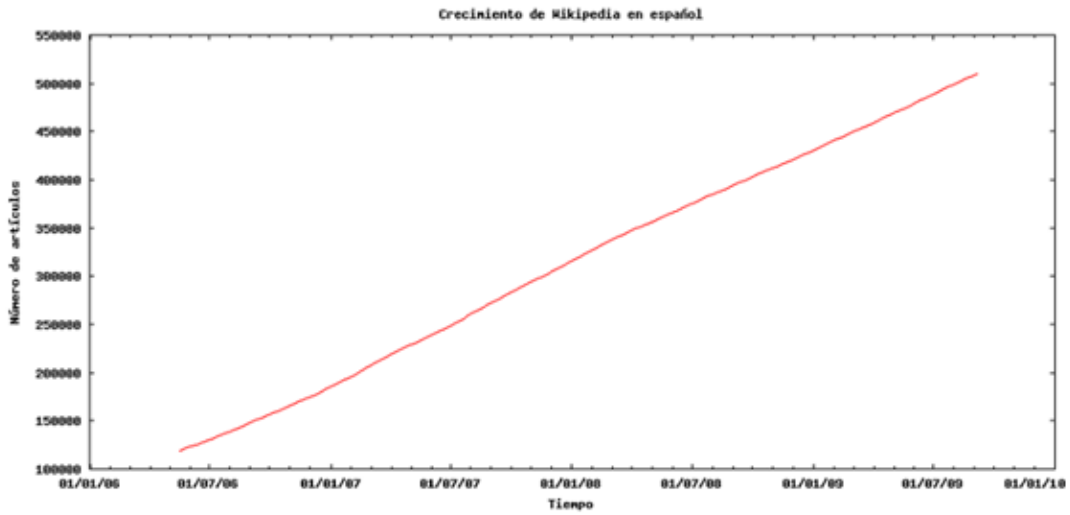

Figura 3- Crecimiento de la Wikipedia en español desde enero del 2006 hasta enero del 2010 (3)

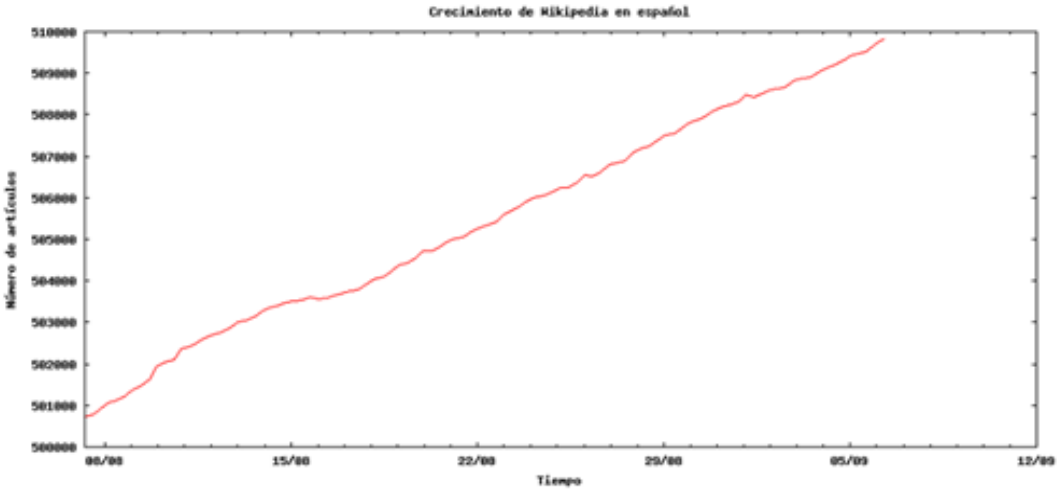

Figura 4- Crecimiento mensual desde Agosto del 2008 hasta Diciembre del 2009 (3)

Existen varios proyectos alrededor del mundo cuyo objetivo ha sido generar una versión especializada de la Wikipedia incluyendo contenido relevante a un país o idioma. Un ejemplo es la Wikipedia en alemán que empezó en el año 2001 con su versión línea y posteriormente fue distribuida en CD en noviembre del 2004 con 132000 artículos y 1200 imágenes. Para noviembre del 2005 fue distribuida con aproximadamente 300000 artículos y 100000 imágenes (4).

El presente trabajo complementa y extiende el trabajo realizado por distribuciones personalizadas como la descrita arriba, ya que constituye una herramienta que permite generar este tipo de distribuciones de manera eficiente y a bajo costo.

1.2.1 Fiabilidad de datos.

Dado que el presente proyecto sugiere el uso de información de la Wikipedia con fines educativos, resulta necesario e interesante dedicar una pequeña sección a la discusión y controversia de la fiabilidad que encontramos en los datos de la misma.

La fiabilidad de los datos a los cuales tenemos libre acceso en la Wikipedia ha sido un tema de debate muy extenso. Su objetivo de “distribuir gratuitamente la totalidad del saber mundial a cada persona del planeta, bajo una licencia libre que permite modificarlo, adaptarlo, reutilizar lo y redistribuirlo libremente”, encabeza una lista de beneficios entre los que mencionamos: que el servicio es gratuito, rápido, se puede corregir de forma inmediata, tiene capacidad ampliamente desarrollada y es una útil herramienta educativa. Sin embargo también podemos mencionar algunas desventajas que fueron resumidas en julio de 2006 por Sam Vaknin en un artículo que publicó en American Chronicle bajo el título 'Los seis pecados de la Wikipedia' (5; 6), entre los que postula:

1. No se conoce a los autores de los artículos.
2. Es anárquica.
3. La fuerza es el principal derecho editorial, pues la autoridad va en función de la cuantía de participación sin importar la calidad.
4. Está contra el verdadero conocimiento, pues los expertos reconocidos son rechazados y atacados en ella. ¿Quiénes son los eruditos?

5. No es una enciclopedia aunque se presente como tal.
6. Es una fuente de difamación y violaciones del copyright .

La misma enciclopedia cita que el contenido de cada artículo debe ser verificado y solo admite como fuentes confiables a los artículos cuyo contenido ha sido publicado luego de un análisis y evaluación por parte de expertos especializados en el tema que está siendo objeto de estudio.

Sin embargo, otras investigaciones han demostrado que el proceso de depuración de la información de la Wikipedia es, en la práctica, bastante bueno, al punto de que una investigación hecha por la revista Nature sugirió que los artículos científicos en la Wikipedia se aproximaban en gran parte por su precisión a los de la enciclopedia Británica (7).

1.2.2 Derechos de autor

El contenido de la Wikipedia es licenciado bajo la Creative Commons Attribution/Share-Alike License 3.0 (Unported) así como bajo la GNU Free Documentation License. Estas licencias permiten el uso comercial de los contenidos reutilizados, siempre y cuando los usos que se les den sean de acuerdo a los términos de uso de la licencia respectiva (8).

1.3. Objetivos del proyecto

El presente proyecto de estudio tiene como objetivos fundamentales:

- Construir un marco de trabajo para seleccionar artículos de la Wikipedia que se basen en temas previamente especificados.
- Generar una versión personalizada que pueda ser accedida sin conexión a Internet para usuarios sin mayor conocimiento técnico.
- Ofrecer a los usuarios una interfaz sencilla y amigable para la lectura de la información generada.

1.4 Justificación del proyecto

La Wikipedia es un gran recurso que se encuentra a disposición de la Humanidad, pero debido a que actualmente el uso de Internet en nuestro país continua siendo limitado, se detecta una profunda brecha digital; es decir una gran diferencia entre aquellos que pueden conectarse y la inmensa mayoría que no tiene acceso a este beneficio. Por ello se debe buscar una alternativa para que sea posible el acceso a la información contenida dentro de la red de una manera desconectada y que la distribución bajo este concepto sea rápida y de fácil acceso.

1.5 Alcance

El proyecto está enfocado al área de educación primaria aunque su diseño es flexible (y además adaptable a otras áreas). Así mismo, su interfaz es sencilla y simple de utilizar lo que permite a cualquier usuario, independiente de su grado de experiencia manipular y buscar la información que necesite.

Dentro de los alcances del mismo está el poder ser utilizado en modo desconectado es decir desde cualquier computador que cuente o no con una conexión a Internet; es decir, que el programa sea completamente portable y puede ser almacenado en cualquier tipo de dispositivo.

2. MARCO TEÓRICO

2.1 Cloud Computing y los Servicios Web de Amazon (AWS)

El término Cloud Computing, o computación en las nubes, se refiere a las aplicaciones entregadas como servicios sobre la Internet y el hardware y software en los datacenters que proveen estos servicios. Los servicios en sí mismos son conocidos como *SaaS (Software as a Service)*, mientras que a las plataformas y hardware subcontratado de esta manera se lo conoce como *IaaS (Infrastructure as a Service).*

El hardware y software del datacenter es lo que llamaremos *Cloud (Nube)*. Cuando una *Cloud* está disponible para alquilar al público, la llamamos *Public Cloud (Nube Pública).* El servicio que es ofrecido por esta, es conocido como *Utility Computing*, por ejemplo las Amazon Web Services (AWS), Google App Engine y Microsoft Azure (9).

Características de estos nuevos servicios son:

- La ilusión de recursos infinitos disponibles bajo demanda.
- La eliminación de un compromiso por adelantado de los usuarios de La Nube.

- La habilidad de pagar por el uso de recursos en base a las necesidades de cada uno.

Entre las empresas que ofrecen servicios de alquiler de recursos a otras se encuentra Amazon la cual oferta sus AWS (Amazon Web Services). AWS es un conjunto de servicios que dan la facilidad de construir aplicaciones Web robustas apoyándose sobre la infraestructura de los mismos (10). Entre los servicios que ofrece se encuentran:

- *Amazon Simple Storage Service (S3):* es un servicio que ofrece un almacenamiento seguro para cualquier tipo de dato ya sea personal o que puede ser usado en una arquitectura distribuida.
- *Amazon Elastic Compute Cloud (EC2):* Es un servicio que permite la ejecución de múltiples servidores bajo demanda, dependiendo de las necesidades del usuario, evitando en esta forma la necesidad de comprar computadores para realizar una tarea que requiera una gran capacidad de procesamiento y con ello el uso de un extenso numero de computadores.

2.2 Paradigma MapReduce y Hadoop

MapReduce es un modelo de programación que permite el procesamiento masivo de datos a gran escala de manera paralela, básicamente el usuario debe plantear el diseño de su solución como dos funciones: Map y Reduce.

La operación Map se aplica a los datos de entrada y los agrupa en una lista ordenada tipo clave/valor, luego procesa este resultado dentro de la función Reduce que agrupa los valores por medio de las claves (11).

Hadoop es un framework que permite ejecutar aplicaciones sobe una gran cantidad de computadores. Para su implementación utiliza el paradigma MapReduce y un sistema de archivos distribuido propio (12).

2.3 Edición de páginas en Wikipedia

La edición de páginas puede ser hecha por cualquier usuario siempre y cuando no esté restringido el artículo. Para poder editarlo se hace uso de un lenguaje especial con una sintaxis sencilla. El texto que forma un artículo es una mezcla de texto normal y etiquetas de este lenguaje que en conjunto se la llama Wikitexto.

2.3.1 Enlaces internos

Un enlace interno es el que apunta a un artículo dentro de la misma Wikipedia. Se lo forma agregando corchete doble al inicio y final de la palabra o frase que se desea convertir en link. Por ejemplo si queremos hacer un link al artículo que habla sobre matemáticas sólo debemos poner [[matemáticas]] en el Wikitexto. Aparecerá la palabra matemáticas como un

enlace al artículo Matemáticas. Pero si la palabra es diferente al nombre del artículo hay que agregar la palabra y el artículo entre corchetes separados por una barra lateral. Por ejemplo, para enlazar la palabra *ecuatoriano* al artículo *Ecuador* sería [[Ecuador|ecuatoriano]]. También son considerados enlaces internos enlaces a otras wikis de la Fundación Wikimedia y enlaces entre idiomas.

Si se quiere enlazar con imágenes o con categorías pero no se desea que se muestre la imagen ni se categorice el artículo, se debe hacer lo siguiente:

- Para enlazar con categorías es necesario poner dos puntos : de esta forma, [[:Categoría:Mi_categoría]]. Por ejemplo Categoría:Escritores se obtiene con [[:Categoría:Escritores]].
- Para enlazar a imágenes del mismo modo, Imagen:Ejemplo.jpg se obtiene con [[:Imagen:Ejemplo.jpg]].

En el caso de enlaces en la misma página, en lugar de poner el nombre del artículo se pone el nombre de la sección precedido de # así: **[[#sección|sección]]** (si se pone **|sección** el enlace aparece con un feo #). También se puede utilizar #top, el cual sube a la parte superior de la página.

2.3.2 Enlaces externos

El Wiki detecta automáticamente los enlaces externos que empiecen por http://. Por ejemplo, un enlace a otra Web: http://www.espol.edu.ec. El mismo enlace se puede hacer poniendo dentro de los corchetes tanto la dirección Web como el título del enlace, separadas por un espacio. Este ejemplo: Página principal de la ESPOL, se ha conseguido de esta manera: [http://www.espol.edu.ec Página principal de la ESPOL] (13).

3. ANÁLISIS Y DISEÑO

3.1 Análisis de datos de entrada.

La librería se alimenta fundamentalmente de dos entradas: la primera es el dataset de la Wikipedia en español y la segunda es la lista de temas de interés con sus categorías asociadas.

- *Dataset*: Tiene un tamaño de 3.2 GB, se encuentra en idioma español y fue descargado con fecha de actualización del 17 de julio del 2009. El dataset se encuentra en un formato XML y contiene los artículos de la Wikipedia sin el historial de revisión, imágenes y sonidos. Cada artículo de la enciclopedia está representado entre las etiquetas <page>, sin embargo para nuestro análisis hacemos uso del wikitexto que es el que se encuentra dentro de los tags <text> de cada página.

```
<page>
  <title>Producto escalar</title>
  <id>15722</id>
  <revision>
    <id>27215338</id>
    <timestamp>2009-06-13T19:23:59Z</timestamp>
    <contributor>
      <username>Raulshc</username>
      <id>607921</id>
    </contributor>
    <comment>retoques</comment>
    <text xml:space="preserve">En matemáticas el producto escalar  ......</text>
  </revision>
</page>
```

Figura 3 - Representación de una página de la Wikipedia en formato XML

- *e temas de interés:* Esta lista contiene los temas que serán incluidos en la versión especializada de la Wikipedia. Para realizar una búsqueda más exacta se debe ingresar por lo menos una categoría a la cual esta enlazado el tema a buscar.

  Por ejemplo si se busca Números Naturales entonces este tema puede ser asociado a la categoría Números o Matemáticas y deberá ser ingresado con cualquiera de los siguientes formatos:

```
Números Naturales;Números
```

```
Números Naturales;Números,Matemáticas
```

```
Números Naturales;Números,Matemáticas,Aritmetica
```

Como se puede ver el formato de la lista de temas de interés es:

```
[TEMA];[CATEGORIA 1],[CATEGORIA 2]..,[CATEGORIA N]
```

3.2 Algoritmo utilizado

La librería está compuesta principalmente de dos algoritmos: el de *Selección de artículos* y el de *Unión y limpieza* de los mismos.

- *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 una 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 en 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 titulo 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 página.
  - Links de idiomas.
  - Links externos.

3.2.1 Primera fase

En esta sección se revisará la primera fase de la generación de la versión especializada de la Wikipedia, la misma tiene como entradas el dataset de la Wikipedia 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.

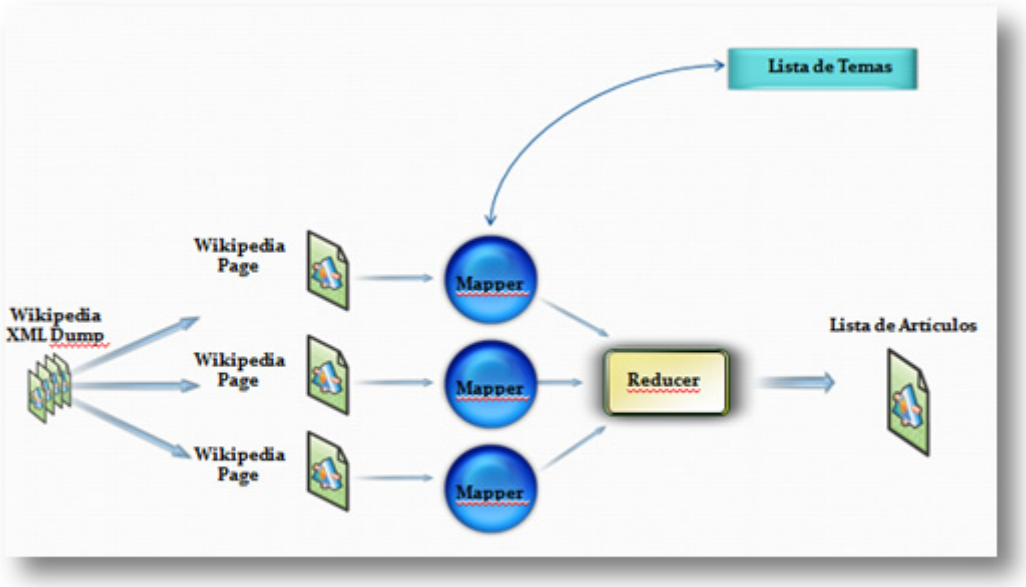

Figura 6 - Primer proceso MapReduce para la selección de artículos de la enciclopedia.

3.2.2 Segunda fase

En esta sección se revisará la segunda fase de la generación de la versión especializada de la Wikipedia, la misma 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.

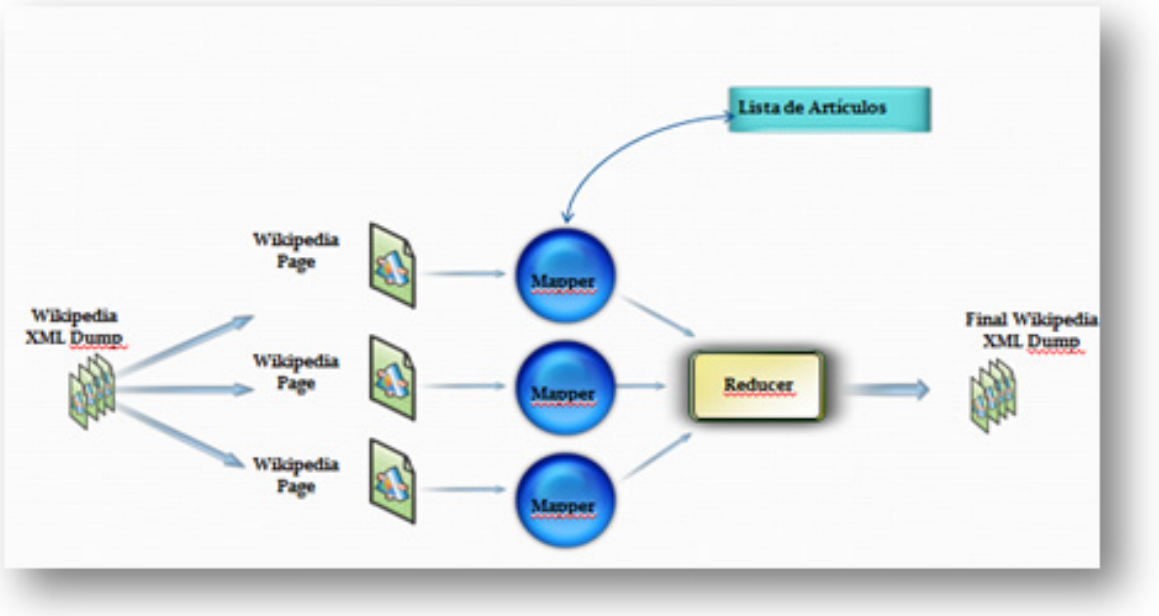

Figura 7 - Segundo proceso MapReduce para la Unión y Limpieza de artículos seleccionados.

4. IMPLEMENTACIÓN Y PRUEBAS

4.1 Detalle de la implementación

En esta sección se describe la implementación de la librería Wikigen que ha sido implementada basada en el diseño propuesto en el capitulo anterior y la librería Cloud[9] que ha sido utilizada para facilitar el trabajo con el dataset de la Wikipedia.

4.1.1 Wikigen

Esta ha sido desarrollada haciendo uso de un conjunto de herramientas y plataformas que nos proveyeron agilidad y productividad. Está integrada por clases de Java que contienen la funcionalidad de la misma. Así también se hizo uso de las librerías de Hadoop, Cloud[9] 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.

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 que contiene las librerías de PHP, una base de datos MySql Lite y el MediaWiki 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 8 - Ejemplo de Wikipedia Personalizada para primer año de Educación Básica

4.1.2 Cloud[9]

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. Fue desarrollada para su uso en los cursos de Cloud Computing en la Universidad de Maryland. 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.

En nuestro proyecto fueron usados específicamente los paquetes *edu.umd.cloud9.collection.wikipedia* y *edu.umd.cloud9.util*.

4.1.3 Expresiones Regulares

Las expresiones regulares son una forma de describir un conjunto de cadenas basándose en características comunes compartidas por cada una de las cadenas del conjunto. Pueden ser utilizadas para buscar, editar o manipular texto y datos.

Dentro de Java el paquete que permite el uso de expresiones regulares es java.util.regex el cual contiene un conjunto de clases para la creación de Expresiones Regulares.

4.2 Software utilizado para las pruebas

En las pruebas con los resultados generados por la librería se utilizó el siguiente software:

- *MediaWiki 1.15.1.* Es un software CMS para wikis con licencia GNU desarrollado en sus orígenes por Magnus Manske y patrocinado por la Wikipedia. Se caracteriza por su fácil instalación y posee una amplia variedad de parámetros configurables escritos en PHP (14). Dentro del proyecto se encuentra ya embebido en MediaWiki el cual es utilizado para leer el archivo generado por la librería de búsqueda, el mismo ha sido adaptado a las necesidades del proyecto modificando sus propiedades y su interfaz con el objetivo de que los usuarios no necesiten realizar ninguna configuración de este, y su uso sea transparente para ellos
- *Server2Go Mini-Package V. 1.7.3*. Servidor Web que corre sin ninguna instalación y en dispositivos protegidos contra escritura. Una aplicación Web instalada en este servidor puede ser usada directamente desde un cdrom, memoria usb o cualquier carpeta en el disco duro sin configurar sus componentes. Esta versión trae pre-configurado Apache 2.0.63, PHP 5.2.10 y MySql 5.0.41 (15).
- *Firefox Portable 3.5.2.* Versión portable de este popular navegador Web. El incluir un navegador que se ejecute cada vez que se inicie el servidor permite que siempre que se abra la enciclopedia no haya

problemas con la visualización de estilos CSS como suele suceder al abrir una misma página Web en distintos navegadores.

4.3 Pruebas y Resultados

Los escenarios seleccionados para las pruebas fueron el segundo y tercer año de educación básica para los cuales se prepararon 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 su página Web (16).

```
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 9- Extracto de un archivo de temas de interés y categorías relacionadas.

Para las pruebas se tuvieron como entradas:

- El dataset de la Wikipedia en español.

- El archivo de Temas de interés y Categorías asociadas. Se subieron dos archivos, uno con 30 temas y otro con 50 temas correspondientes a segundo y tercer año de educación básica respectivamente.

El dataset de la Wikipedia fue subido al S3 de Amazon y para procesarla con nuestra librería se levantaron clústeres 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 esclavos para cada archivo de temas subido. El tiempo total de ejecución es la suma de los tiempos individuales de cada proceso MapReduce correspondiente a cada algoritmo presentado, el de selección de artículos y el de Unión y Limpieza.

4.3.1 Resultados con 3 nodos esclavos.

# Temas	# Tareas map	# Tareas reduce	Tamaño del XML Resultado	Tiempo	Proceso MapReduce
30	26	1	1.2 MB	Total: 18m 36s	
				1) 13m 4s	Select articles
				2) 5m 32s	Clean articles
50	26	1	4.3 MB	Total: 22m 55s	
				1) 16m 18s	Select articles

				2) 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 buscados se incrementa el tiempo para Unir el resultado y hacer la limpieza de los enlaces rotos.

4.3.2 Resultados con 7 nodos esclavos.

# Temas	# Tareas map	# Tareas reduce	Tamaño del XML Resultado	Tiempo	Proceso MapReduce
30	26	1	1.2 MB	Total: 10m 2s	
				1) 7m 26s	Select articles
				2) 2m 36s	Clean articles
50	26	1	4.3 MB	Total: 13m 9s	
				1) 9m 57s	Select articles
				2) 3m 12s	Clean articles

Como se puede observar en las pruebas con 7 nodos se produjo una reducción del tiempo de generación del XML de resultado tanto para 30 temas como para 50 temas, sin embargo como era de esperarse, el tamaño del XML generado no varía.

Una vez obtenido el archivo XML con los temas seleccionados y limpios de links y datos innecesarios, se procedió a cargarlo dentro del servidor potable haciendo uso del MediWiki para leer el archivo y transformarlo en un modelo relacional con el objetivo de brindarle al usuario un entorno gráfico enriquecido y sencillo para la lectura, búsqueda y organización de la información.

Para evitar problemas de compatibilidad de navegadores se ha agregado al contenido un navegador portable en este caso *Firefox Portable 3.5.2.*

Todo este contenido fue integrado dentro de un dispositivo de almacenamiento (en nuestro caso un CD), para ser leído desde cualquier computador. El peso de todo el contenido es de aproximadamente 300MB lo que permite llevarlo en casi cualquier dispositivo de almacenamiento que se usa en la actualidad, brindándole al usuario la facilidad de portarlo consigo y acceder a sus beneficios en el momento que lo desee y que lo requiera.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

1. 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.
2. 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.
3. 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.

4. Entre las principales y más importantes ventajas que ofrece una Wikipedia Personalizada a los usuarios se menciona:

   Facilitarles una interfaz Web sencilla y fácil de utilizar.

   Brindar la capacidad de portabilidad desde cualquier dispositivo de almacenamiento que hoy en día se manejan en forma práctica.

   Contar con herramientas internas y un navegador portable que le evitan a los usuarios problemas de compatibilidad, permitiéndole así abstraerse del funcionamiento interno de la librería.

5. El tiempo de procesamiento fue razonablemente disminuido gracias al uso de un sistema distribuido alquilado (versus haber hecho pruebas en el clúster local del laboratorio de Sistemas Distribuidos de la ESPOL, que solamente consta de tres nodos).
6. En el análisis realizado hemos generado un archivo que contiene los artículos seleccionados, que, en conjunto con el servidor portable y las herramientas necesarias para la ejecución y lectura de nuestra enciclopedia no supera los 300MB.

Recomendaciones

1. En nuestro país la demanda de televisores y DVDs es mayor que la de computadoras[1], 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.
2. Sería recomendable también, validar las distribuciones generadas, realizando pruebas de uso y encuestas a al menos 30 estudiantes de primaria y uno o más profesores de ese nivel, de alguna escuela fiscal de la ciudad.
3. 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.
4. En caso de desear implementar las ideas descritas en el presente trabajo en el Ecuador, se debe diseñar los procesos que permitan la

[1] Según un estudio realizado por el Instituto Nacional de Estadísticas y Censos (INEC), en el Ecuador la venta de productos electrodomésticos para el hogar al año 2007 fue de 600 millones de dólares, siendo las ventas de televisores y DVDs las más representativas del total (18), lo que nos muestra el mayor grado de consumismo de estos enseres frente al de los computadores que tienen un menor grado de acceso.

distribución de las versiones personalizadas de la Wikipedia a sus usuarios objetivos.

5. La interfaz debería permitir que el usuario fácilmente pueda conocer la fecha en que la distribución fue generada, de tal manera que no haya confusiones sobre la actualidad o frescura de su contenido.
6. Sería de utilidad facilitar el añadir contenido a las versiones personalizadas y poder guardar dichos cambios (respaldar) para poder incluirlos en futuras generaciones de la distribución. Agradecemos al Dr. Jorge Calderón, Director del CICYT, por su sugerencia al respecto.

Bibliografía

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. **Fundación Wikimedia.** Wikimedia Downloads. [En línea] 26 de Agosto de 2009. [Citado el: 26 de Agosto de 2009.] http://download.wikipedia.org.

3. **Gráficos del crecimiento de la Wikipedia.** Wikipedia. [En línea] 21 de 9 de 2009. http://es.wikipedia.org/wiki/Usuario:Chewie.

4. **Colaboradores de Wikipedia.** Wikipedia en alemán. [En línea] Wikipedia, La enciclopedia libre., 22 de Julio de 2009. [Citado el: 27 de Agosto de 2009.] http://es.wikipedia.org/wiki/Wikipedia_en_alemán.

5. **EL PAIS.COM.** ¿Debemos fiarnos de la Wikipedia? [En línea] 10 de Junio de 2009. [Citado el: 27 de Agosto de 2009.] http://www.elpais.com/articulo/sociedad/Debemos/fiarnos/Wikipedia/elpeputec/20090610elpepisoc_1/Tes.

6. **Vaknin, Sam.** The Six Sins of the Wikipedia. [En línea] American Chronicle, 2 de Julio de 2006. [Citado el: 27 de Agosto de 2009.]

7. **Colaboradores de Wikipedia.** Reliability of Wikipedia. [En línea] Wikipedia, la enciclopedia libre, 27 de Agosto de 2009. [Citado el: 27 de Agosto de 2009.] http://en.wikipedia.org/wiki/Reliability_of_Wikipedia.

8. —. Derechos de autor - Wikipedia, la enciclopedia libre. [En línea] [Citado el: 05 de Septiembre de 2009.] http://es.wikipedia.org/wiki/Wikipedia:Derechos_de_autor.

9. **M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia.** Above the Clouds: A Berkeley View of Cloud Computing. *University of California at Berkeley.* [En línea] Febrero de 2009. http://berkeleyclouds.blogspot.com/2009/02/above-clouds-released.html.

10. What is AWS? [En línea] [Citado el: 26 de Agosto de 2009.] http://aws.amazon.com/what-is-aws/.

11. **Ghemawat, J. Dean and S.** *MapReduce: Simplified Data Processing on Large Clusters.* Berkeley, CA, USA : s.n., 2004.

12. Hadoop Wiki. [En línea] [Citado el: 26 de Agosto de 2009.] http://wiki.apache.org/hadoop/.

13. **Colaboradores de Wikipedia.** Ayuda:Cómo se edita una página. [En línea] Wikipedia, la enciclopedia libre. [Citado el: 27 de Agosto de 2009.]

14. **MediaWiki contributors.** 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.

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

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

17. **Josune Cordoba Torrecilla, Pedro Cuesta Morales.** *Adaptando un sistema de Wikis para su uso educativo.*

18. **INEC.** Comercio Interno. [En línea] [Citado el: 19 de Septiembre de 2009.] http://www.inec.gov.ec/web/guest/descargas/basedatos/inv_eco/com_int.