ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"Análisis de la WEB de la ESPOL y afines, utilizando Hadoop

como una plataforma de procesamiento masivo de datos"

INFORME DE PROYECTO DE GRADUACIÓN:

Previo a la obtención del Título de

INGENIERO EN COMPUTACION ESPECIALIZACION

SISTEMAS MULTIMEDIA

INGENIERO EN COMPUTACION ESPECIALIZACION

SISTEMAS TECNOLOGICOS

Presentado por:

Cinthia Piedad Martínez Montero

Carlos Fernando Barcos Sinche

GUAYAQUIL – ECUADOR

Año: 2009

// A G R A D E C I M I E N T O

A todas las personas que de uno u otro modo colaboraron en la realización de este trabajo y especialmente a la Ing. Cristina Abad Directora de Tesis, por su invaluable colaboración.

Cinthia Martínez

Carlos Barcos

D E D I C A T O R I A

A mi madre por ser ejemplo de virtud y dedicación, A mi padre ejemplo de trabajo incansable, a mis hermanos la alegría de mi vida.

Cinthia Martínez

A DIOS por la vida y las oportunidades que me ha brindado. A mi padre y madre por todo el apoyo y comprensión que me han brindado a lo largo de mi vida académica. A mis hermanos por sus valiosos consejos.

Carlos Barcos

TRIBUNAL DE SUSTENTACIÓN

Ing. Jorge Aragundi Rodríguez
SUBDECANO DE LA FIEC
PRESIDENTE

Msc. Cristina Abad Robalino
DIRECTOR DEL PROYECTO
DE GRADUACIÓN

Msc. Carmen Vaca Ruiz
MIEMBRO PRINCIPAL

Ph. D. Enrique Peláez
MIEMBRO PRINCIPAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este proyecto de graduación nos corresponden exclusivamente, y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL)

____________________________ ____________________________

Cinthia Martínez Montero   Carlos Barcos Sinche

RESUMEN

El presente documento muestra los resultados del análisis de la red de la ESPOL, utilizando Hadoop como plataforma de procesamiento masivo de datos. Gracias al estudio que se ha realizado, se ha podido demostrar que la estructura de la Web de la ESPOL no tiene propiedades de pequeño mundo (no es una red libre de escala), forma que usualmente toman muchas de las redes reales, y que tiene gran incidencia en la "navegabilidad y accesibilidad de la información en grandes redes de documentos" [18]. Esto dificultaría la exploración de la Web de la ESPOL, y tendría una incidencia negativa en la percepción de la utilidad (a los usuarios) de nuestra Web.

Para este estudio, utilizamos los índices obtenidos de la indexación de los enlaces entrantes como salientes de las páginas Web del dominio espol.edu.ec. Estos datos fueron procesados para así obtener la cantidad de enlaces entrantes y salientes para cada uno de ellos. Además, los mismos datos nos permitieron conseguir la distribución estadística de enlaces (entrantes y salientes) de las paginas del dominio de la ESPOL, y así poder comprobar que la misma no tiene las propiedades de una distribución de *ley*

de potencias (power law), un criterio fundamental que debe cumplir una red para poder ser clasificada como *libre de escala* (scale free).

Finalmente, para validar este análisis se ha considerado estudios previos a las redes de otras universidades, que sí muestran una estructura pequeño mundo.

ÍNDICE GENERAL

ABREVIATURAS

API: Interfaz de programación de aplicaciones.

Crawl: Araña de la web; indexador.

Csv: Formato de archivo separado por comas.

HDFS: Sistema de Archivos distribuido de Hadoop.

ESPOL: Escuela Superior Politécnica del Litoral.

FIEC: Facultad de Ingeniería Eléctrica y Computación.

HTML: *Lenguaje de Marcas de Hipertexto,* es el lenguaje de marcado predominante para la construcción de páginas web.

Hum/soc: Sub-sitios pertenecientes a la categoría Humanidades y Ciencias Sociales.

ICM: Instituto de Ciencias Matemáticas.

Log: Logaritmo.

MIT: Instituto Tecnológico de Massachusetts.

Nat/tech: Sub-sitios de la categoría ciencias Naturales y Tecnología.

Path Nets: Sub-grafo que contiene los caminos más cortos.

SCC: Strongest Connected Component.

SSH: Intérprete de órdenes seguro.

SSHD: Open Secure Shell, conjunto de aplicaciones que permiten realizar comunicaciones cifradas a través de una red, usando el protocolo SSH.

Topic drift: Link transversals

UK: Reino Unido (Escocia, Inglaterra, Irlanda del Norte y Gales).

URL: Localizador Uniforme de Recursos.

Web: *Red Global Mundial (*World Wide Web*).*

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

En los últimos meses hemos visto un gran interés por parte de la comunidad de la ESPOL, en incrementar su reconocimiento académico nacional e internacional. Como punto de referencia podemos tomar el Ranking Mundial de las Universidades proporcionado por el Laboratorio de Cibermetría del Consejo Superior de Investigaciones Científica de España[1] donde la ESPOL consta en el puesto 62 de las universidades latinoamericanas.

Debido a esto la ESPOL está implementando una política para mejorar la accesibilidad de la información hacia el contenido publicado en su sitio Web. El sitio Web de la ESPOL posee enlaces a los diferentes institutos, centros y unidades dentro de la misma. Sin embargo, es fácil notar que su navegabilidad no es tan sencilla debido a que no existe un amigable flujo de navegación. Además, cuando se empezó a desarrollar el presente proyecto no existían enlaces que nos permitan conocer las publicaciones científicas realizadas por estudiantes o personal docente de la ESPOL.

[1] Disponible en línea en http://www.webometrics.info/top100_continent_es.asp?cont=latin_america

Más específicamente, se planteó la hipótesis de que la Web de la ESPOL no tiene la forma de una red “pequeño mundo”, forma que usualmente toman muchas de las redes reales y que tiene gran incidencia en la “navegabilidad y accesibilidad de la información en grandes redes de documentos” [18]. Esto dificultaría la exploración de la Web de la ESPOL, y tendría una incidencia negativa en la percepción de la utilidad (a los usuarios) de nuestra Web.

CAPÍTULO 1

1. PLANTEAMIENTO DEL PROBLEMA

1.1 Motivación

En estudios anteriores se ha demostrado que muchas de las redes reales (ya sean en el campo de la biología, sociología, o informática) en lugar de tomar una forma de red aleatoria, suelen tomar una forma denominada “pequeño mundo” [20]. Este tipo de redes se caracteriza por poseer varios nodos concentradores. Los nodos concentradores poseen muchos más enlaces hacia otros nodos que los nodos normales en la red. Por ejemplo, una de las razones por las que el sitio de Wikipedia se ha hecho famoso es porque, al poseer esta característica, es muy fácil encontrar información útil publicada en esta enciclopedia [27]. En la Web, la forma pequeño mundo facilita la navegación, ya que desde un sitio poco visitado, hay enlaces a sitios muy visitados (concentradores), lo que facilita que encontremos la información. Una Web, como la de la ESPOL, que no tiene forma de red libre de escala, no permite ubicar fácilmente la información y ve reducida su utilidad (hacia sus usuarios).

Como hemos mencionado, uno de los índices utilizados para la medición del Ranking Mundial de Universidades de Webometrics [1] son las publicaciones científicas y su impacto en el medio que se ve reflejado en los enlaces de sitios externos hacia el sitio de la ESPOL. También, se considera la cantidad de páginas dentro de la institución que los buscadores Web pueden encontrar. Debido a esto podemos inferir que si mejoramos la estructura de navegación de los sitios de la ESPOL podríamos captar una mayor cantidad de Internautas, y también, aumentar el número de páginas de la institución que pueden ser halladas utilizando un navegador. Por ejemplo, a principios del 2008, si buscábamos la página del Grupo de Visualización Científica y Sistemas Distribuidos de la ESPOL en Google, no era posible hallarla[2]. Esto tenía una repercusión negativa para la institución, ya que en el sitio www.visid.espol.edu.ec, podemos encontrar información sobre proyectos de investigación, publicaciones científicas, etc.

---

[2] Actualmente es posible encontrar el sitio del VISID, debido a las actividades que inició en el 2008 la ESPOL con el fin de aumentar su visibilidad.

1.2 Antecedentes

Como se ha mencionado la ESPOL se encuentra actualmente en el puesto 62 del ranking latinoamericano de universidades proporcionado por el Laboratorio de Cibermetría del Consejo Superior de Investigaciones Científicas de España. Este ranking se obtiene a través de un proceso automatizado, a partir de un indicador combinado que tiene en cuenta tanto el volumen de los contenidos Web como la visibilidad y el impacto de estas publicaciones Web de acuerdo al número de enlaces externos entrantes.

Según las investigaciones realizadas acerca de las estructuras de navegación en la Internet, y nuestro análisis al problema, la estructura que posee el sitio Web de la ESPOL no parecía poseer una forma "pequeño mundo". Esto representaría una deficiencia en la exploración de contenidos en la Web, y por lo tanto en la utilidad de la misma.

1.3 Objetivos

Mediante el uso de una plataforma de procesamiento masivo y distribuido para la exploración de datos Hadoop, realizar un análisis de la estructura que siguen los diferentes sitios Web de la ESPOL para determinar, de forma certera, cuál es el principio de dicha estructura, con miras a mejorar su navegabilidad y el valor a los usuarios de la misma.

1.4 Justificación

El objetivo de esta investigación es analizar sistemáticamente la estructura de la Web de la ESPOL para determinar con certeza su forma, identificar causales de esta anomalía (ya que como se mencionó anteriormente, al auto-formarse la Web usualmente toma una forma pequeño mundo), y proponer posibles soluciones para mejorar la exploración de contenido en la Web de la ESPOL. Esto le permitiría a la institución dar a conocer con mayor énfasis las publicaciones científicas realizadas por la ESPOL, además de brindar información acerca de todo lo referente a las carreras, centros de investigaciones, profesores, materias, etc. De hecho, los propios miembros del Laboratorio de Cibermetría del Consejo Superior de Investigaciones Científicas de España, que realiza el ranking de universidades en base a Webometrics, indican que “si el rendimiento Web de una institución se encuentra por debajo de los esperado de acuerdo a su excelencia académica, los dirigentes universitarios deberían reconsiderar su política Web, promoviendo el incremento substancial del volumen y la calidad de sus publicaciones electrónicas”.

Para realizar este estudio, fue necesario procesar todas las páginas Web de la ESPOL, y analizar los enlaces entre ellas. Para esto, utilizamos Hadoop. Hadoop [3] es un framework que combina un sistema de archivos distribuidos con un algoritmo de computación distribuida denominada Map-Reduce. Esta combinación le da a Hadoop la capacidad de indexar y manipular grandes cantidades de datos en menor tiempo que otros sistemas tales como RDBMS, Grid computing, Volunteer Computing [22]. Su potencia sobrepasa esta función o utilidad, ya que por su característica de tratar grandes volúmenes de información, puede ser utilizado para la exploración de datos, búsqueda de patrones, análisis de hipervínculos, entre otros. Hadoop fue elegido como herramienta para el presente trabajo por la facilidad de instalación y configuración que nos provee, el hardware necesario para su instalación no tiene un costo elevado. Además, las tareas de búsqueda, análisis e indexación pueden ser fácilmente divididades en subtareas, lo cual sigue el modelo del algoritmo Map-Reduce.

Cabe recalcar que, si este estudio quisiera hacerse de manera manual tendría costos elevados de tiempo y poca confiabilidad, y estaría propenso a errores producto de tal mecanismo. En el caso de optar por el desarrollo de una aplicación que realice el procesamiento tendríamos problemas similares ya que el costo en tiempo sigue siendo alto debido a la falta de pruebas necesarias para verificar el correcto funcionamiento de la aplicación, y la misma sería propensa a errores de programación (bugs). Mientras que la plataforma Hadoop nos garantiza muchos beneficios ya que ha sido diseñada con este propósito. Hadoop es actualmente utilizada por empresas que procesan datos masivos como Yahoo!, Facebook, Amazon, Twitter, entre otras [22].

1.5 Alcances y limitaciones

Este proyecto tiene como alcance:

- Buscar y analizar los enlaces tanto entrantes como salientes hacia el sitio www.espol.edu.ec.
- Estudiar el tipo de estructura que tiene el sitio de la ESPOL www.espol.edu.ec.
- Plantear posibles soluciones para mejorar la estructura de la ESPOL.
- Implementar un clúster de procesamiento masivo con las plataformas Linux-Hadoop-Nutch.

Las limitaciones que tiene el presente proyecto:

- La cantidad de datos a bajar es grande por tal motivo el tiempo de obtención de datos, procesamiento de la información y análisis es un factor altamente limitante.

- Dependemos de la disponibilidad ininterrumpida de los servidores que alojan los sitios que forman parte de la ESPOL.
- Dependemos de la energía eléctrica que forman parte de la alimentación de los equipos que son parte de las herramientas del presente proyecto, así como también de los servidores sobre los cuales están alojados los sitios que forman parte de la red de la ESPOL, ya que sin la disponibilidad dichos servidores nos es imposible obtener la información que requerimos para el presente proyecto.

CAPÍTULO 2

2. FUNDAMENTOS TEÓRICOS

2.1 Conceptos básicos

2.1.1 Redes pequeño mundo

En los años 60 ***Stanley Milgram*** hizo lo que él denominó ***Experimentos del Mundo Pequeño*** que dio origen a los llamados ***seis grados de separación*** [20]. Esta idea consiste en que dos ciudadanos cualesquiera en EE.UU. están separados por una cadena de nos más de seis conocidos de distancia.

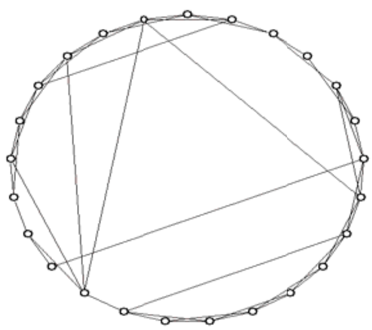


FIGURA 2.1. RED PEQUEÑO MUNDO [25]

Una red ***pequeño Mundo*** es un tipo de grafo con un alto grado de agrupamiento (clústering) con una longitud del camino promedio entre dos nodos bastante pequeña [8] [35]. Un ejemplo de esto lo podemos ver en una red social, en la cual los nodos son las personas y los enlaces son la relación que ellos mantienen con otros miembros de la red.

A) Aleatoria **B) Independiente de escala**

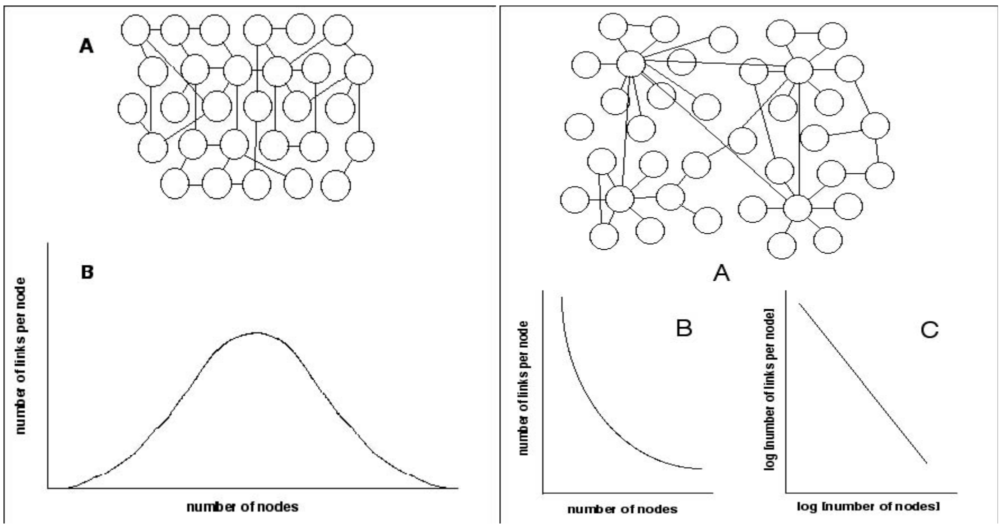


FIGURA 2.2. REPRESENTACIÓN ESTANDAR DE REDES ALEATORIAS E INDEPENDIENTES DE ESCALA [25]

En una red ***pequeño mundo*** podemos reconocer principalmente dos propiedades:

- Dos nodos cualesquiera dentro de la red se comunican entre sí por medio de un camino de nodos intermedio relativamente pequeño. El tamaño de este camino crece de forma logarítmica con el número de nodos de la red. [35]

- Posee valores altos de coeficiente de agrupamiento (clústering coefficient). Este valor indica que aunque dos nodos cualesquiera en la red no están conectados de forma directa, existe una gran probabilidad de que se conecten a través de otros nodos en la red. [35]

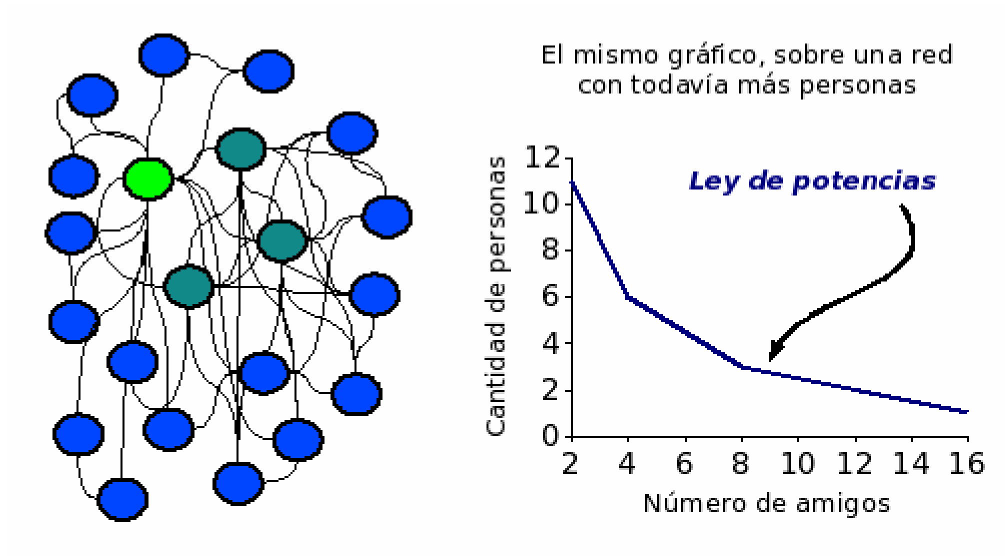


FIGURA 2.3. CARACTERISTICA DE LA RED LIBRE ESCALA[25]

Redes de libre escala

Una *red de libre escala* es una red cuya distribución de enlaces sigue a la de la *Ley de las Potencias*. Esto significa que la fracción P(k) de nodos en una red que tienen K conexiones hacia otros nodos, para grandes valores de K, es

$$P(k) \approx k^{-\gamma}$$

Donde γ es una constante cuyo valor generalmente se encuentra entre 2 y 3 ($2< \gamma <3$), aunque ocasionalmente aparecen excepciones.

Una red de libre escala se define como una red que contiene algunos nodos que se encuentran altamente conectados. Es decir que poseen un gran número de enlaces hacia otros nodos, aunque el grado de conexión de casi todos los nodos es bajo.[36]

El interés en las redes de libre escala creció en 1999 cuando Albert-László Barabási y sus colegas de la Universidad de Notre Dame crearon un mapa de la topología de una porción de la red en la que encontraron que algunos nodos que ellos llamaban concentradores (hubs) tenían una cantidad considerable de enlaces que otros nodos en la red. Además notaron que el número de enlaces que conectan a un nodo tenía la distribución de *ley de potencias* [9].

Ejemplos de redes de libre escala:

La red de amistades entre personas. También se puede extender hasta las redes de llamadas telefónicas, envíos de postales y correo electrónico [28].

- Las redes de distribuciones eléctricas, en las cuales existen estaciones enormes que abastecen a zonas enormes y al mismo tiempo a transformadores pequeños.
- Las redes de comercio internacional, ya que los países desarrollados que son la minoría, concentran la mayor cantidad de intercambio de bienes.
- Las redes de interacción de proteínas en el metabolismo celular, en donde unas cuantas proteínas aparecen en la mayoría de reacciones mientras que la mayoría aparecen solo en situaciones específicas.

Ley de potencias (power law)

La *ley de potencias* es un tipo de relación matemática entre dos cantidades. Si una cantidad es la frecuencia y la otra el tamaño del evento en sí, entonces la relación es una distribución de *ley de potencias* si el tamaño del evento incrementa de forma en que la frecuencia del evento decrementa lentamente. Por ejemplo, un terremoto con el doble de largo en duración es 4 veces menos frecuente en suceder.[37]

Una relación en forma de *ley de Potencias* entre dos escalares cuantitativos X y Y es aquella que puede expresarse de la siguiente manera:

$$y = ax^k$$

Donde a es la constante de proporcionalidad y k es el exponente de la potencia. Tanto a como k son constantes. [37]

2.1.2 Computación distribuida

Para definir la computación distribuida vamos a comenzar con una breve explicación de lo que es un sistema distribuido, el cual se define como una colección de computadores separados físicamente y conectados entre sí por una red de comunicaciones distribuida; cada máquina posee sus componentes de hardware y software que el usuario percibe como un solo sistema [23].

El usuario accede a los recursos remotos de la misma forma en que lo hace con los recursos locales, o a un grupo de computadores que usan un software para conseguir un objetivo común.

Una vez conocido este concepto, podemos definir la computación distribuida como un modelo para resolver problemas utilizando un gran número de computadoras organizadas en clústeres incrustados en una infraestructura de telecomunicaciones masiva [11].

La computación distribuida permite crear una abstracción al usuario de los componentes heterogéneos de la red. De esta forma, el operador no tiene que preocuparse de los detalles que involucran las diferentes plataformas, arquitecturas y lenguajes de programación.

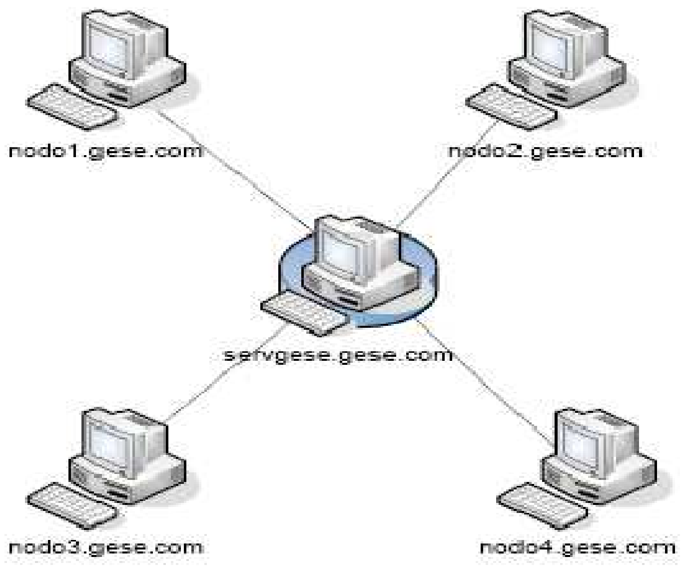


FIGURA 2.4. ESQUEMA DE COMPUTACIÓN DISTRIBUIDA [30].

2.2 Herramientas de desarrollo

2.2.1 Plataforma Hadoop y HDFS

Hadoop es una plataforma de procesamiento de datos masivos de código libre, desarrollada en Java. Esta plataforma fue construida pensando en la gran necesidad que existe en la actualidad de procesar grandes cantidades de información en el menor tiempo posible.

Hadoop está basado en un potente sistema de archivos distribuido llamado HDFS por sus siglas en ingles (Hadoop Distributed File System). Este sistema de archivos está diseñado para correr sobre máquinas de bajo costo y fácil acceso. Este sistema pese a tener similitudes con otros sistemas ya existentes se diferencia de ellos por ser altamente tolerante a fallos además de ser diseñado para ser desarrollado sobre un hardware de bajo costo.

2.2.2 Análisis de enlaces

Lucene y Nutch proporcionan herramientas útiles para el análisis de enlaces. Ellos trabajan como un motor de búsqueda de los sitios Web y a su vez realizan la indexación de los mismos. Para luego almacenarlos en el HDFS de Hadoop. Finalmente con la siguiente herramienta a describir estos se procesaran.

2.2.3 Map-Reduce

Es un paradigma de programación [21] similar a *Dividir y Vencer* pero que se aplica a grandes volúmenes de datos. Tiene dos componentes principales:

- **Map**, un proceso que se encarga de leer y procesar información generando tuplas clave, valor que después serán tomados por un proceso **Reducer**. Sobre una máquina pueden estar ejecutando varias instancias **Map**, cada uno de las cuales reciben y generan información. [22]
- **Reduce**, es la fase que se encarga de recopilar la información generada por los procesos **Map** y procesarla generando una salida final. Esta información a su vez puede ser procesada posteriormente por los procesos **Map** para ser procesado subsecuentemente generando otro tipo de información.[22]

2.2.4 Librería de visualización de enlaces

El desarrollo del presente proyecto incluye el análisis y representación gráfica de los datos obtenidos de las búsquedas de enlaces, por tanto fue necesario buscar herramienta que nos permitan la visualización de los mismos.

A continuación se detalla las herramientas encontradas para la representación gráfica de los datos: (ver sección 5.1)

Graphviz.- Su nombre proviene de Graph Visualization, es una herramienta de código abierto que nos permite representar las estructuras de información de grafos de redes. El problema de utilizar esta herramienta es capacidad limitada para procesar grandes cantidades de datos como el sitio de la ESPOL.
Para conocer más puede acceder a: http://www.graphviz.org/

LaNet-VI.- Esta herramienta provee imágenes de redes de grandes escalas en un plano bidimensional. El algoritmo está basado en el k-core decomposition [26]. Una descripción completa pueda ser encontrada en el artículo k-core decomposition: A tool for the visualization of the large scale networks [29]. Para conocer más puede acceder a: http://xavier.informatics.indiana.edu/lanet-vi/

Cytoscape.- Es una plataforma de código abierto diseñada para la visualización de redes de interacción molecular e integración con los perfiles de análisis de expresión genética y otros datos de estado. Aunque Cytoscape haya sido desarrollado para uso de investigaciones biológicas y sea ese el campo de mayor uso, puede ser utilizado para la visualización y el análisis de cualquier tipo de grafos de red que involucren nodos y hojas, por ejemplo las redes sociales. Un aspecto calve del diseño de la arquitectura es el uso de los plugin para funciones especializadas los cuales son creados por los desarrolladores de Cytoscape y por la comunidad de usuarios. Para conocer más puede acceder a: http://www.cytoscape.org/

2.2.5 Proyectos de Lucene

Lucene [12] es un API de código abierto desarrollado para recuperar información. Originalmente fue implementado en Java, pero ahora soporta varios lenguajes tales como Delphi, Perl, C#, C++, Python, Ruby y PHP.

Esta librería es útil para cualquier aplicación que requiera indexado y búsqueda a texto completo. Lucene ha sido utilizado ampliamente por su utilidad en la implementación de motores de búsqueda. Es por esto que fácilmente se lo confunde con un motor de búsqueda con funciones de crawling y análisis de documentos en HTML incorporadas.

El núcleo de la arquitectura se centra en el objeto Documento (Document), el cual está conformado por campos (Fields) de texto. De esta forma Lucene puede ser independiente del tipo de archivo, extrayendo información de los mismos sin importar si es un PDF, HTML, documento de WORD, etc. Los archivos pueden ser indexados siempre y cuando pueda extraerse información de ellos.

Esta librería forma parte de Nutch [13], el cual es un software que integra todo lo que hace falta para completar un motor de búsqueda de páginas Web.

CAPÍTULO 3

3. CONSTRUCCIÓN DE UN CLÚSTER HADOOP EN LA ESPOL

3.1 Análisis Preliminar y requerimientos para la implantación de un clúster Hadoop en la ESPOL

3.1.1 Requerimientos del proyecto.

El proyecto por ser de naturaleza de procesamiento masivo de datos ha requerido la utilización de herramientas que nos permitan realizar las tareas necesarias para recopilación, procesamiento y análisis de los datos.

A continuación se lista el tipo de herramientas necesarias para el desarrollo del proyecto:

- Distribución de Linux, por la alta disponibilidad de herramientas de código abierto disponible.
- Plataforma de procesamiento masivo de datos.

- Web crawler, para descargar íntegramente la Web de la ESPOL.
- Entorno de desarrollo integrado, idealmente multiplataforma y de código abierto.
- Lenguaje de programación de alto nivel que permite desarrollar código distribuido, utilizando la plataforma de procesamiento distribuida seleccionada.
- Lenguaje o herramienta para cálculos y análisis estadísticos (que permita determinar si una distribución es libre escala o no).

3.1.2 Análisis de las herramientas y selección de la más apropiada.

Ubuntu Studio Works.- Esta distribución de Linux fue elegida por tener características de ser estable, robusto y tener fácil administración.

Hadoop.- Debido a la alta demanda de recursos para procesamiento de datos fue necesaria la utilización de esta herramienta como plataforma de procesamiento. Elegimos esta herramienta porque nos ofrece un sistema de archivos distribuido diseñado para el procesamiento de grandes cantidades de información, tolerante a fallos, y de fácil instalación.

Nutch.- La necesidad de obtener cada enlace dentro del dominio www.espol.edu.ec, tanto los enlaces hacia él como desde él hacia sitios externos nos ha llevado a elegir esta herramienta de la Apache Software Foundation. El dominio www.espol.edu.ec además abarca los diversos sitios de la Escuela Superior Politécnica del Litoral (ESPOL) tales como el Instituto de Ciencias Matemáticas (ICM), Facultad de Ingeniería Eléctrica y Computación (FIEC) entre otras. Por lo expuesto anteriormente necesitábamos de una herramienta de búsqueda e indexación que nos permita obtener la información de cada enlace hacia y desde el dominio www.espol.edu.ec. Nutch por ser un potente motor de búsqueda e indexación, además de estar diseñada para integrase con Hadoop, fue utilizada para el presente proyecto.

ECLIPSE.- La plataforma Hadoop está desarrollada en diferentes lenguajes, incluyendo Java. Necesitábamos un editor que nos permita desarrollar código en dicho lenguaje y que sea fácil de utilizar e integrar con la plataforma escogida. Esta integración se la hace a través de un plug-in desarrollado para Eclipse por medio del cual es posible integrar los servicios de Hadoop.

JAVA.- Para minimizar esfuerzos nos vimos en la necesidad de un lenguaje que sea fácil de utilizar, potable, independiente del sistema operativo y compatible con la plataforma utilizada. Java es un lenguaje sencillo de aprender, además de que hemos adquirido cierto nivel de experiencia en su uso a través de nuestra vida académica. Además es soportado de manera nativa en Hadoop.

Tomcat.- Una de las herramientas utilizadas es Nutch, el cual nos provee de servicios de administración y seguimiento de los procesos que buscan e indexan la información de cada enlace del sitio www.espol.edu.ec. Para poder habilitar estos servicios es necesario mantenerlos bajo un servidor de Web.

R Project.- Una vez que los datos fueron obtenidos y procesados a través de Nutch y Hadoop respectivamente, hubo la necesidad de analizarlos. El análisis consistía en una serie de cálculos estadísticos. Debido a la gran cantidad de información que se estaba analizando se requería de una herramienta que nos permita procesar y graficar la información de forma adecuada.

3.2 Diseño e Implementación de un Clúster Hadoop en la ESPOL.

3.2.1 Requerimientos de hardware.

Los requerimientos de hardware propuestos para la instalación de Hadoop sobre el clúster son las siguientes[3] :

- Procesador Dual-Core Intel Xeon 2.0 GHZ
- 8GB de memoria RAM
- Discos SATA de 41TB
- Tarjeta de red Gigabit Ethernet

El siguiente software fue necesario instalarlo como parte de los pre-requisitos:

- Java 1.5.x o superior
- SSH y SSHD

3.2.2 Arquitectura de la plataforma del sistema de archivos distribuido (HDFS).

HDFS tiene una arquitectura maestro-esclavo la cual consiste en un nodo principal llamado NameNode, que sirve para administrar el sistema de archivos y regular el acceso de las sub-tareas de

---

[3] Requerimientos necesarios para asegurar una baja tasa de fallos debido al hardware. [22] Hadoo p Book – The Definitive Guide Chapter 9 – Setting Up a Hadoop Cluster – pag 245 - 246

procesamiento. Además hay procesos llamados DataNodes, los cuales son los encargados de administrar el almacenamiento en los nodos en los cuales están corriendo, generalmente existe un DataNode por cada nodo en el clúster. El HDFS crea localidades de memoria permitiendo a los datos del usuario ser almacenados en archivos dentro de él. El sistema de archivos divide estos en uno o más bloques que son pasados y almacenados en los DataNodes. El NameNode ejecuta operaciones de apertura, lectura, cierre y renombrado de archivos y directorios, determina el mapeo de bloques con los DataNode. Los DataNode son los responsables de atender las peticiones de lectura/escritura desde el sistema; son los responsables de la creación, eliminación y replicación de los bloques bajo las instrucciones del NameNode.

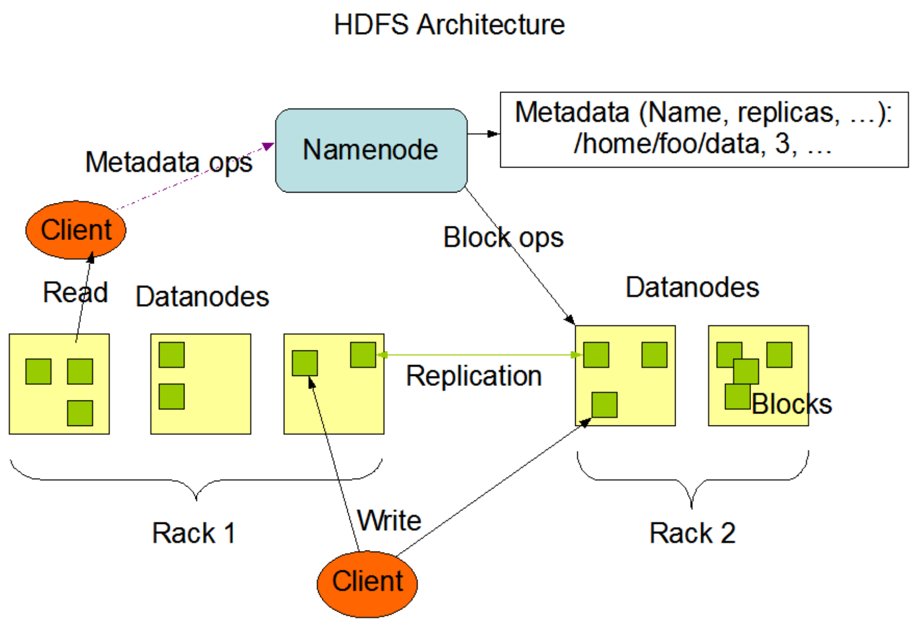


FIGURA 4.1 ARQUITECTURA DEL HDFS [31].

3.2.3 Instalación de los componentes necesarios para el clúster.

3.2.3.1 Linux.

La distribución que se instaló en las computadoras fue **Ubuntu Studio Works** por medio del respectivo asistente de instalación que se provee en el disco.

3.2.3.2 Nutch.

Para la instalación de esta herramienta fue necesario descargar los archivos fuentes del sitio oficial de Nutch[4]. La versión que utilizamos en el presente trabajo es Nutch 0.8.1. Una guía detallada de la instalación y configuración de esta herramienta se encuentra en el Anexo C.

3.2.3.3 Tomcat.

Para habilitar los servicios de visualización de los procesos de ejecución de tareas, fue necesaria la instalación de un servidor de aplicaciones que soporte Java. En este caso la versión que viene en la distribución de Linux es adecuada, y sólo fue necesario

[4] http://www.apache.org/dyn/closer.cgi/lucene/nutch/

configurarlo. El detalle de la configuración puede consultarse en el Anexo C.

3.2.3.4 Java.

La herramienta Hadoop necesita de la plataforma Java para su ejecución. El sito oficial de Java nos provee la máquina virtual que incluye el IDE Netbeans, una vez descargado el paquete lo instalamos siguiendo los pasos del instalador.

3.2.3.5 Configuración del clúster Hadoop.

El presente estudio toma como herramienta principal la versión Hadoop 0.19. La instalación de este paquete requiere de la instalación y configuración de Java y SSH. Contamos con tres computadores distribuidos de la siguiente forma:

- Un equipo encargado de la administración de los procesos, denominado máster, pero también podrá realizar la labor de esclavo.
- Dos equipos esclavos sobre los que se ejecutan los procesos Map-Reduce. Estos son denominados slave2 y slave3 respectivamente.

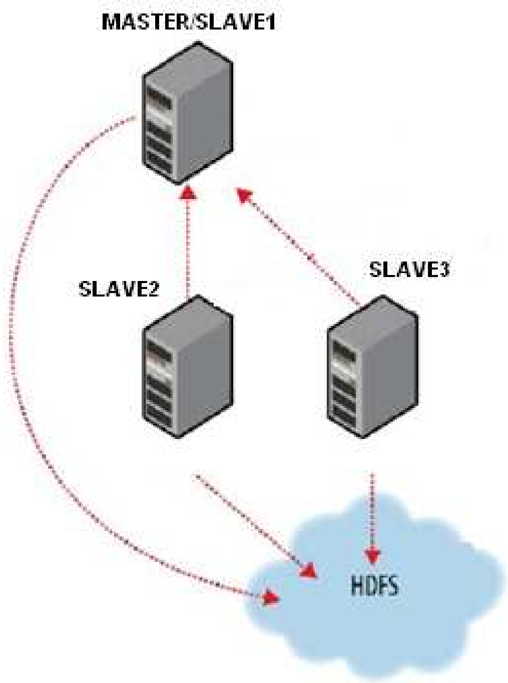


FIGURA 4.2 ESTRUCTURA DEL CLUSTER.

3.2.3.6 Pruebas.

Para comprobar el correcto funcionamiento del sistema ejecutamos un programa de ejemplo que viene con la herramienta Hadoop. La aplicación WordCount[5], el cual lo obtuvimos de la guía oficial de Hadoop [24], la misma recibe un archivo (o directorio con archivos) de texto plano y cuenta las palabras del mismo. Devuelve un archivo con las diferentes palabras encontradas y el número de repeticiones de la misma.

---

[5] http://wiki.apache.org/Hadoop/C%2B%2BWordCount

CAPÍTULO 4

4. DISEÑO E IMPLEMENTACIÓN DEL ALGORITMO MAP-REDUCE PARA EL PROCESAMIENTO MASIVO DE DATOS CON HADOOP

4.1 Map-Reduce.

MapReduce es un framework introducido por Google [21] para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras. Se han escrito implementaciones de MapReduce en C++, Java, Python y otros lenguajes.

Las funciones Map y Reduce están definidas ambas con respecto a datos estructurados en pares (clave, valor). Map toma uno de estos pares de datos con un tipo en un dominio de datos, y devuelve una lista de pares en un dominio diferente: [21]

Map (k1, v1) -> list (k2, v2)

La función de mapeo es aplicada en paralelo para cada ítem en la entrada de datos. Esto produce una lista de pares (k2, v2) por cada llamada.

Después de eso, el framework de MapReduce junta todos los pares con la misma clave de todas las listas y los agrupa, creando un grupo por cada una de las diferentes claves generadas. La función Reduce es aplicada en paralelo para cada grupo, produciendo una colección de valores para cada dominio:

Reduce (k2, list (v2)) -> list (v2)

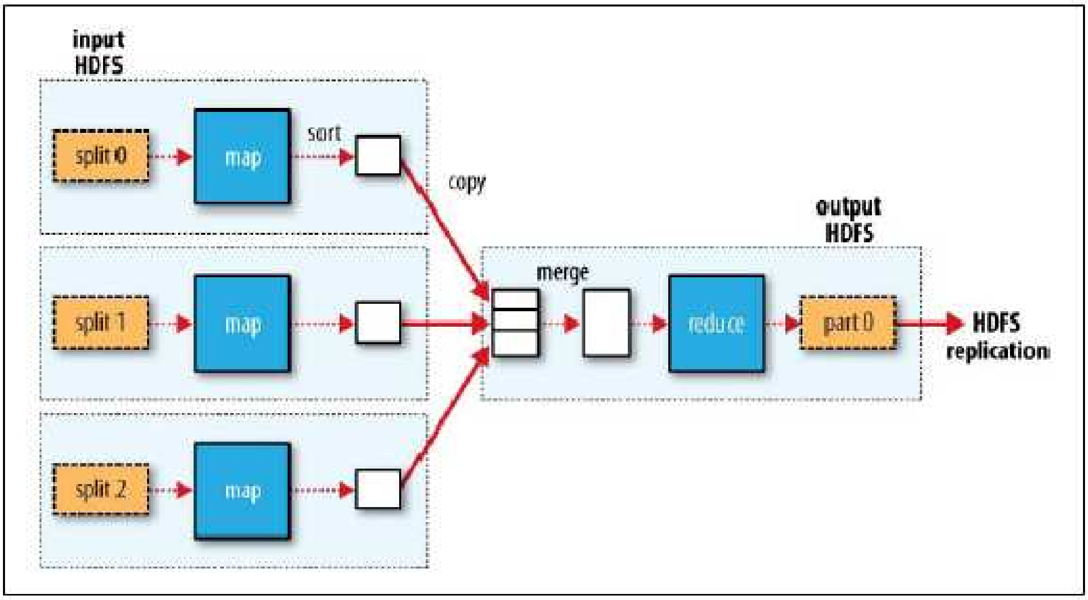


FIGURA 5.1. PROCESO MAP-REDUCE [22].

Cada llamada a Reduce típicamente produce un valor v2 o una llamada vacía, aunque una llamada puede retornar más de un valor. El retorno de todas esas llamadas se recoge como la lista de resultado deseado. Por lo tanto, el framework MapReduce transforma una lista de pares (clave, valor) en una lista de valores. Este comportamiento es diferente de la combinación "map and reduce" de programación funcional, que acepta una lista arbitraria de valores y devuelve un valor único que combina todos los valores devueltos por mapa.

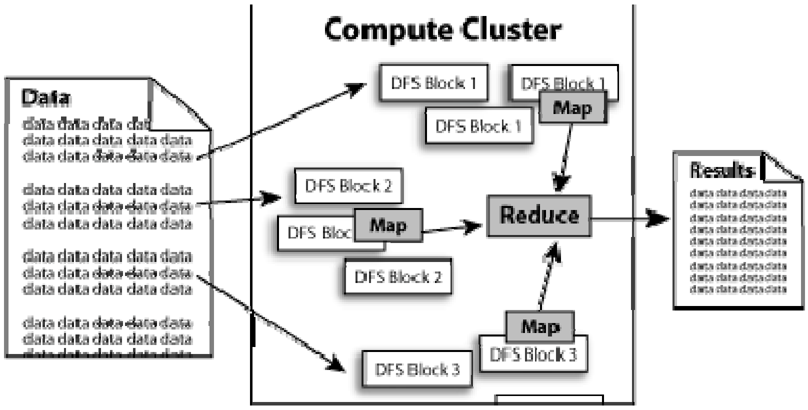


FIGURA 5.2. ESQUEMA MAP REDUCE [3].

4.1.1 Análisis de la herramienta.

Map-Reduce es una herramienta muy útil, ya que nos permite procesar grandes cantidades de información de manera rápida y eficiente, además de presentar la información obtenida en la forma que más nos convenga. En este análisis los datos a procesar son los enlaces que han sido previamente descargados e indexados por nuestro motor de búsqueda Nutch y depositados en el HDFS de Hadoop. [22]

Generamos un algoritmo MapReduce en lenguaje Java, que con los enlaces obtenidos nos presentaba una lista de los enlaces con sus enlaces de entrada y una lista del mismo con sus enlaces de salida.

4.1.2 Instalación de los componentes.

4.1.2.1 Eclipse.

El proceso de instalación de la herramienta Eclipse puede ser consultado en el Anexo A, sección Eclipse.

4.1.2.2 Plugin de Hadoop para Eclipse.

La comunidad Eclipse ha desarrollado un plugin que permite integrar los servicios de la plataforma Hadoop con esta potente herramienta de desarrollo. Para ver la instalación y configuración del plugin por favor revise el Anexo A, sección Plugin de Hadoop.

4.1.2.3 Pruebas.

Finalizada la instalación realizamos las pruebas correspondientes de las herramientas instaladas y así poder asegurarnos del correcto funcionamiento de las mismas, además de ir familiarizándonos con la misma para poder generar la solución. Nuevamente se corrió la aplicación WordCount mencionada en la sección 3.2.3.6.

Levantamos los servicios de Hadoop y a través del plugin pudimos integrar Eclipse y Hadoop, muestra de esto es que desde Eclipse conseguimos navegar a través del sistema de archivos distribuido

HDFS. Ejecutamos el ejemplo desde el IDE Eclipse y una vez más con el navegador provisto por el plugin logramos visualizar el archivo generado como resultado.

4.1.3 Diseño del algoritmo de procesamiento para la solución.

En esta etapa de la investigación empezamos por determinar los datos que eran necesarios para continuar con el proceso. En este caso, los datos de entrada eran los enlaces indexados. Gracias a la investigación previa, teníamos claro que se necesitaba obtener una lista de los enlaces de entrada y otra de los enlaces salida. Esta información sería vital para nuestro siguiente paso, la tabulación de los enlaces. El algoritmo y los resultados se pueden observar en el Anexo C.

4.1.4 Pruebas.

Ya establecido el algoritmo lo ejecutamos primero con una sección de enlaces recogida en una prueba de nuestro clúster para poder evitar errores en el proceso con gran cantidad de datos. Además de esta manera determinar si la forma de mostrar los datos era la correcta. Una vez finalizadas las pruebas procesamos los enlaces correspondientes y obtuvimos dos listas de enlaces ya procesadas en la cual se mostraban los enlaces, la cantidad total de enlaces de entrada y la lista de los mismos. De igual manera con los enlaces de salida.

Los datos obtenidos por estas pruebas se encuentran detallados en el Anexo C y analizados en el Capítulo 6.

CAPÍTULO 5

5. VISUALIZACIÓN DEL ESQUEMA DE RED DE LA ESPOL

5.1 Análisis y selección de la herramienta.

Con la lista de enlaces que obtuvimos de la indexación optamos por representar de forma gráfica la red de la ESPOL de manera que podamos apreciar en cierto grado su estructura. Para este efecto buscamos una herramienta apropiada. Mediante un proceso de investigación pudimos encontrar tres herramientas que podían cumplir con esta finalidad.

Inicialmente se analizó **Graphviz,** la cual resultó sencilla en su instalación y además el formato de archivos requeridos puede generarse sin complicaciones con un proceso MapReduce adicional. Esta herramienta funcionó correctamente en la fase de pruebas, con un número de enlaces limitado; pero el archivo de los enlaces de la Web de la ESPOL no pudo ser procesado debido a su gran tamaño.

Luego, **LaNet-VI** fue nuestra segunda opción, pero es una herramienta en línea que al parecer presenta problemas, ya que no nos envió el resultado a nuestra cuenta de correo electrónico, ni pudimos obtener información de los errores generados (en caso de existir).

Finalmente, utilizamos **Cytoscape** ya que con esta se pudo generar un grafico de la Web de la ESPOL de manera exitosa.

5.2 Instalación de la herramienta.

Como ya se mencionó inicialmente debido a que se contaba con pocos datos se consideró otras herramientas, estas fueron descartadas al momento de efectuar las pruebas con los datos reales.

Cytoscape es una herramienta diseñada para la visualización de redes, aunque fue desarrollada para el uso de investigaciones biológicas puede ser utilizado para la visualización y análisis de cualquier tipo de red. Para conocer más puede acceder a: http://cytoscape.org/.

Para poder observar de manera más detallada la instalación y configuración de las herramientas escogidas hemos considerado crear una sección de Anexos que nos ilustra este tópico. Para detalles de la instalación referirse al Anexo D.

5.3 Selección del tipo de gráfico a presentar.

Los gráficos estándar para redes bi-direccionales son los de enlaces entrantes, enlaces salientes y una combinación de ambos. Los tres tipos de gráficos, aplicados a la red de la Web de la ESPOL, se describen a continuación.

- En la figura 5.1 de la sección 5.5 podemos observar los enlaces de los sitios Web de la ESPOL con sus respectivos enlaces de entrada. Podemos ver claramente que todos los enlaces tienen al menos un enlace de entrada.
- En la figura 5.2 de la sección 5.5 se representa los enlaces con sus respectivos enlaces de salida, se puede inferir de esta gráfica que no se encuentran sitios aislados, que al menos existe un enlace de salida que comunica a cada enlace.
- La figura 5.3 de la sección 5.5 representa la red de la ESPOL, en esta se puede observar la gran cantidad de enlaces concentrados en nodos determinados. Existen nodos concentradores como espol.edu.ec., este presenta 1016 enlaces de salida.

De las figuras podemos apreciar que existen varios nodos concentradores como el blog de la Espol, el sitio del vicerrectorado, etc. Estos sitios sirven como enlace para poder generar caminos entre los nodos.

Resulta complejo definir la estructura con los gráficos obtenidos, por lo que es necesaria la aplicación de algún método matemático que demuestre los criterios que impulsaron el estudio. Para este efecto nosotros decidimos hacerlo con un programa de análisis estadístico: R-Project.

5.4 Selección del visor indicado.

Durante las pruebas iniciales cuando contábamos con un conjunto pequeño de enlaces la herramienta indicada era Graphviz, este necesitaba de un visor que modelaba la red. Luego de intensificar las investigaciones y contando con un archivo de enlaces mayor esta fue descartada. Por este motivo no fue necesario un visor.

5.5 Pruebas.

Finalmente, ya seleccionadas las herramientas nos enfocamos en determinar su idoneidad. Instalamos cada una de ellas y ejecutamos pruebas con archivos de ejemplo proporcionados por las herramientas, luego ya establecido el formato de archivo de entrada de cada una de ellas ejecutamos pequeños archivos generados con los datos reales, en este punto pudimos determinar el mejor de ellos.

En este caso el más idóneo para nuestro estudio fue **Cytoscape** el cual nos permitió graficar la red completa de la ESPOL, sin importar su tamaño. Las Figuras 5.1, 5.2 y 5.3, muestran los gráficos generados con esta herramienta, según lo detallado en la Sección 5.3.

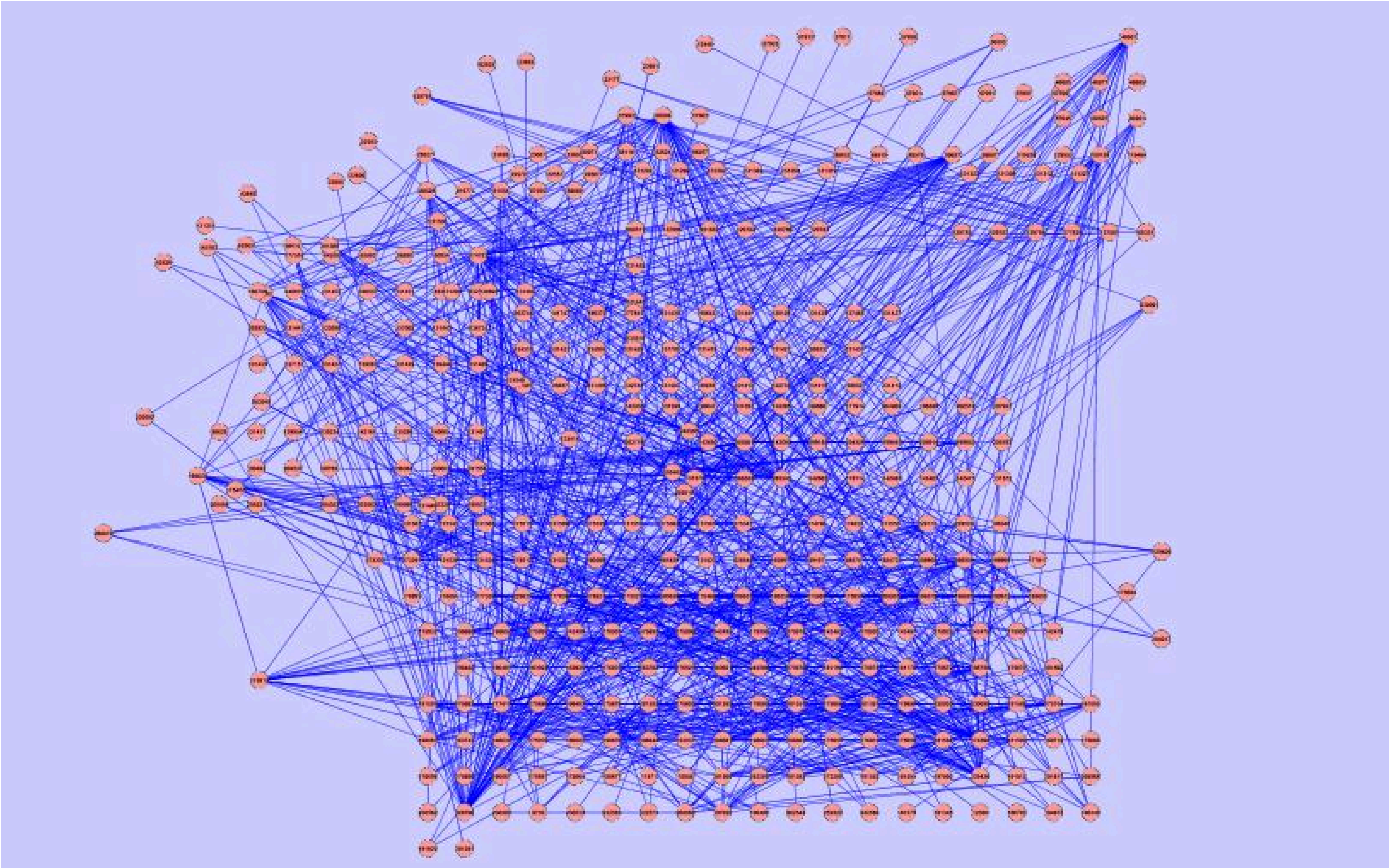


FIGURA 5.1 GRÁFICO DE ENLACES DE ENTRADA DE LA RED DE LA ESPOL.

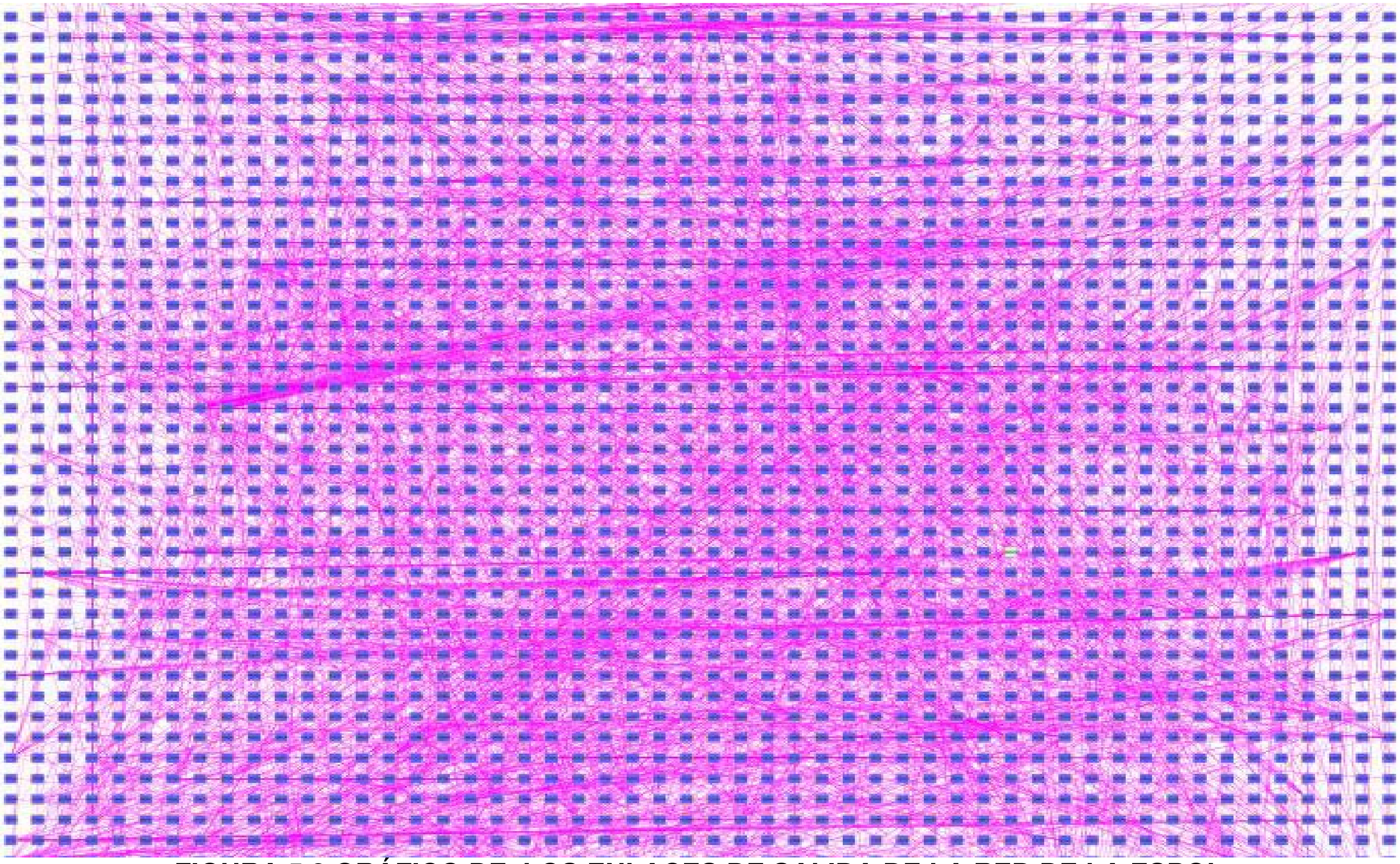


FIGURA 5.2 GRÁFICO DE LOS ENLACES DE SALIDA DE LA RED DE LA ESPOL.

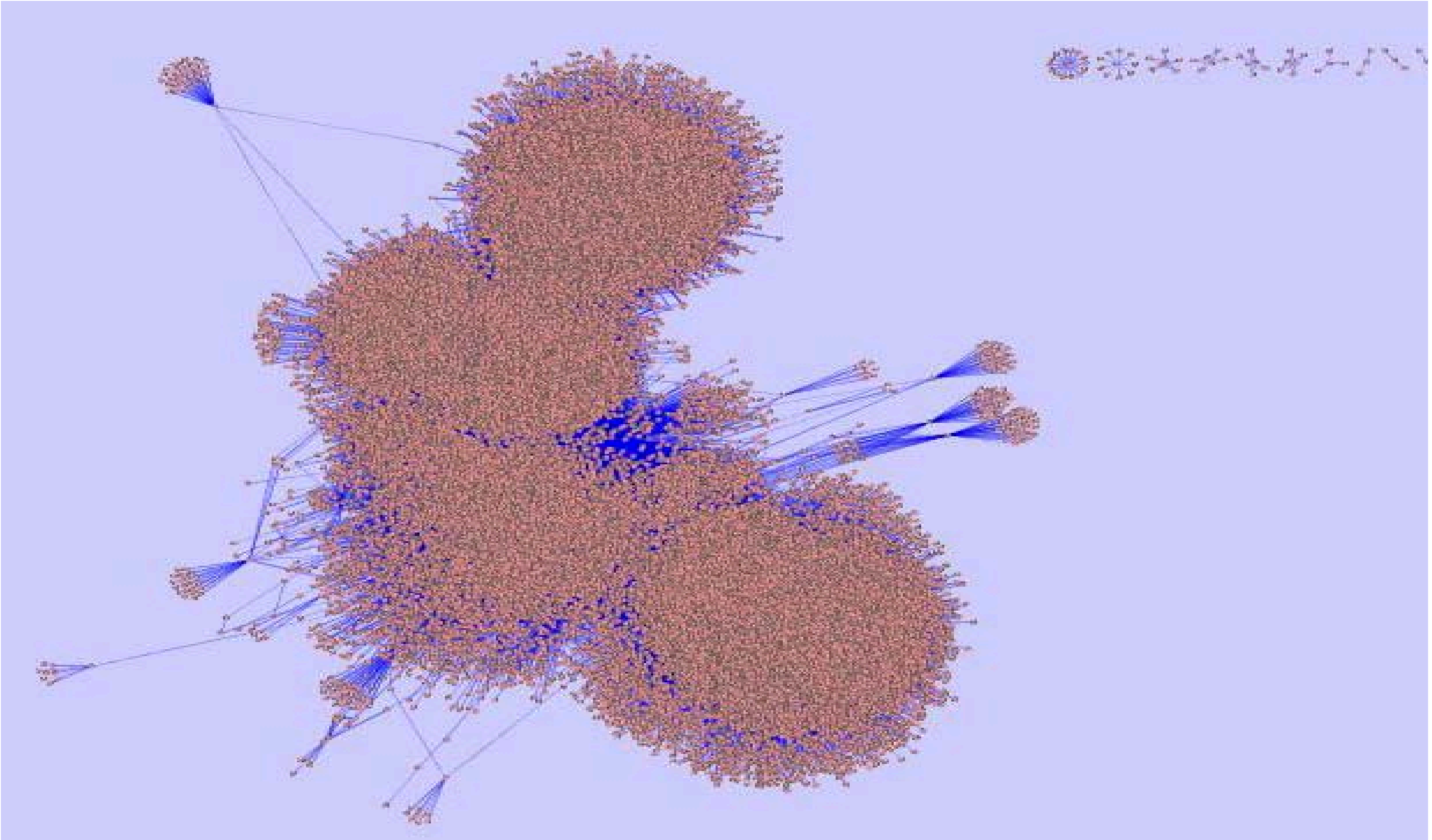


FIGURA 5.3 GRÁFICO DE LA RED DE LA ESPOL.

CAPÍTULO 6

6. PRUEBAS CON OTROS SITIOS Y ANÁLISIS.

6.1 Pruebas con otro sitio para validar el análisis.

La idea principal fue realizar el mismo estudio comparativo que se realizó con la ESPOL para otros sitios Web que se encuentren en un mejor puesto en el ranking de Webometric. Escogimos el MIT como objeto de análisis y comparación. Pero debido a complicaciones con la energía eléctrica de la ESPOL (durante varios meses de prueba, los crawls nunca pudieron terminar ya que las fallas y cortes planificados se dan al menos una vez por semana) y errores al alcanzar los enlaces, tuvimos una alta tasa de fallo en la indexación de enlaces de la misma. Esto hace que nuestro estudio presente un sesgo. Debido a esto nos vimos en la necesidad de utilizar como base comparativa, los resultados de un estudio doctoral [18] realizado sobre las universidades del Reino Unido.

6.1.1 Estructura del sitio.

El estudio que vamos a presentar a continuación muestra que los sitios Web de las universidades del Reino Unido sí tienen una estructura “Pequeño Mundo”.

6.1.1.1 Análisis del modelo de la red obtenido.

El estudio se centra en demostrar que la red obedece una estructura “Pequeño Mundo”, para lo cual se sirven de los datos obtenidos de una búsqueda Web a través de la red sobre los sitios de las universidades del Reino Unido. El trabajo utiliza las distribuciones de enlaces entrantes y salientes de los sitios Web que conforman esta red de universidades.

El estudio indicado analizó la recopilación de los 7669 sitios dentro de la red de universidades del Reino Unido y consideró las distribuciones de enlaces entrantes y salientes de los mismos.

Los sub-sitios en los componentes fuertemente conectados (SCC) en la red de universidades del Reino Unido fueron escogidos pensando ser utilizados en posteriores investigaciones. La decisión fue tomada en base a las siguientes condiciones:

(1) 100% de validez de nombres de dominios de SCC.

(2) La característica de que solo en la red SCC existen enlaces en ambas direcciones entre todos los sub-sitios, permitiendo de esta forma identificar fácilmente las propiedades “Small-World”.

(3) Un alto porcentaje o al menos el 85.5% de enlaces tienen en su camino nodos que pertenecen a la red SCC.

(4) Una gran parte, el 64.2% de todas las conexiones sub-sitio a sub-sitio se localizan dentro de la red SCC.

(5) La red SCC contienen sólo el 24.7% (1893) de todos los sub-sitios obtenidos.

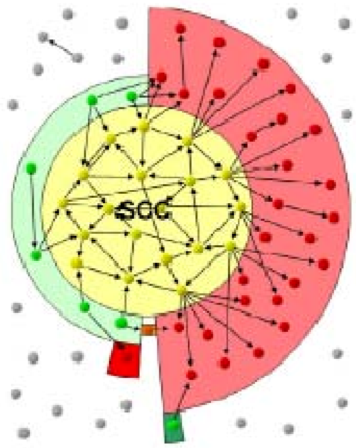


FIGURA 6.1 ANÁLISIS CENTRADO EN LA RED SCC [18].

Se encontró que las distribuciones de enlaces entrantes y salientes entre los nodos de la red de los 1893 sub-sitios elegidos en las Figuras 6.2 y 6.3 muestran una forma de ley potencias como la distribución para todos los sub-sitios (7669), condición necesaria en las redes pequeño mundo. Las Tablas 6.1 y 6.2 muestran los 15 enlaces con la mayor cantidad de enlaces entrantes y salientes.

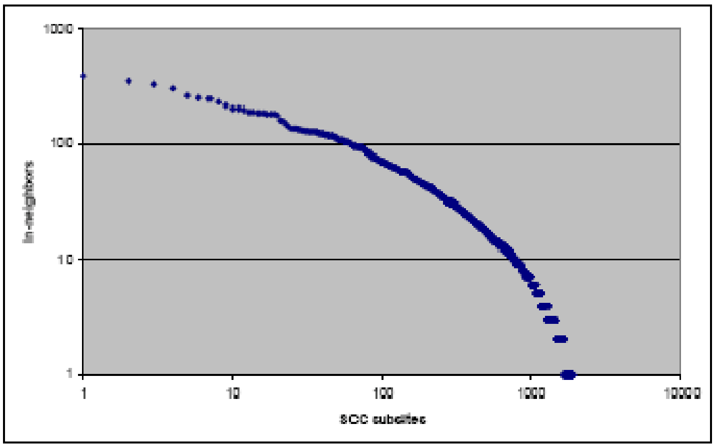


FIGURA 6.2 DISTRIBUCIONES DE ENLACES ENTRANTES PARA 1893 SUB-SITIOS DE LA RED SCC EN ESCALA LOG-LOG [18].

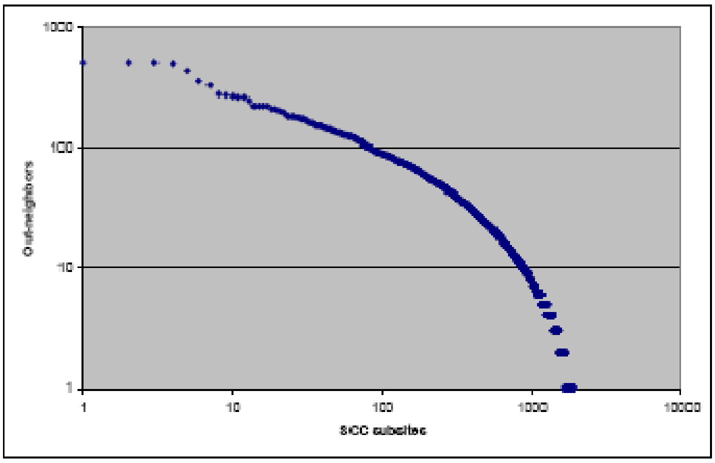


FIGURA 6.3 DISTRIBUCIONES DE ENLACES SALIENTES PARA 1893 SUB-SITIOS DE LA RED SCC EN ESCALA LOG-LOG [18].

rank	id	subsite	comp.	# in-neigh-bors	# out-neigh-bors	affiliation
1	4124	src.doc.ic.ac.uk	OUT	411	0	'Sun Site' mirror site (Usenet archive) at Imperial College, London
2	1357	cbl.leeds.ac.uk	SCC	387	91	Computer Based Learning Unit, Univ. of Leeds
3	3017	scit.wlv.ac.uk	SCC	349	434	School of Computing and Info.Technology, Univ. of Wolverhampton
4	4129	sunsite.doc.ic.ac.uk	OUT	331	0	'Sun Site' mirror site (Usenet archive) at Imperial College, London
5	1821	users.ox.ac.uk	SCC	330	507	Personal web pages at Univ. of Oxford
6	2760	cs.ucl.ac.uk	SCC	300	265	Dept. of Computer Science, Univ. College London
7	4928	comlab.ox.ac.uk	OUT	289	0	Computing Laboratory (CS dept.), Univ. of Oxford
8	1866	info.ox.ac.uk	SCC	259	120	Former server with official web pages of Univ. of Oxford
9	3020	scitsc.wlv.ac.uk	SCC	249	8	School of Computing and Info. echnology, Univ. of Wolverhampton
10	3339	cup.cam.ac.uk	OUT	249	0	Cambridge University Press
11	325	cl.cam.ac.uk	SCC	246	141	Computer Laboratory (CS dept.), Univ. of Cambridge
12	2642	cogs.susx.ac.uk	SCC	231	268	School of Cognitive and Computing Sciences, Univ. of Sussex
13	1466	cs.man.ac.uk	SCC	218	224	Dept. of Computer Science, Univ. of Manchester
14	925	dcs.gla.ac.uk	SCC	203	511	Dept. of Computing Science, Univ. of Glasgow
15	3010	csv.warwick.ac.uk	SCC	202	354	Univ. of Warwick Information Service

TABLA 6.1. LOS 15 SUB-SITIOS CON LA MAYOR CANTIDAD DE ENLACES ENTRANTES HACIA SUS VECINOS EN LA RED [18].

rank	id	subsite	comp.	# in-neigh-bors	# out-neigh-bors	affiliation
1	1068	cee.hw.ac.uk	SCC	146	518	Dept. of Computing and Electrical Engineering, Heriot-Watt Univ.
2	1572	doc.mmu.ac.uk	SCC	127	514	Dept. of Computing and Mathematics, Manchester Metropolitan Univ
3	925	dcs.gla.ac.uk	SCC	203	511	Dept. of Computing Science, Univ. of Glasgow
4	1821	users.ox.ac.uk	SCC	330	507	Personal web pages at Univ. of Oxford
5	3017	scit.wlv.ac.uk	SCC	349	434	School of Computing and Info. Technology, Univ. of Wolverhampton
6	3010	csv.warwick.ac.uk	SCC	202	354	Univ. of Warwick Information Service
7	2367	ecs.soton.ac.uk	SCC	117	327	Dept. of Electronics and Computer Science, Univ. of Southampton
8	791	dai.ed.ac.uk	SCC	137	280	Department of Artificial Intelligence, Univ. of Edinburgh
9	2291	afm.sbu.ac.uk	SCC	2	277	'Virtual library', Centre for Applied Formal Methods, South Bank Univ
10	2642	cogs.susx.ac.uk	SCC	231	268	School of Cognitive and Computing Sciences, Univ. of Sussex
11	1268	comp.lancs.ac.uk	SCC	183	265	Computing Dept., Univ. of Lancaster
12	2760	cs.ucl.ac.uk	SCC	300	265	Dept. of Computer Science, Univ. College London
13	19	users.aber.ac.uk	SCC	34	250	Personal web pages at Univ. of Wales, Aberystwyth
14	3042	www-users.york.ac.uk	SCC	94	226	Personal web pages at Univ. of York
15	1597	dcs.napier.ac.uk	SCC	127	226	School of Computing, Napier Univ.

TABLA 6.2. LOS 15 SUB-SITIOS CON LA MAYOR CANTIDAD DE ENLACES SALIENTES HACIA SUS VECINOS EN LA RED [18].

Un sub-sitio dentro de la red SCC recibe un promedio de 18.1 enlaces de entradas de otros sub-sitio y provee un promedio de 23.6 enlaces de salida hacia otro sub-sitio.

Adicionalmente, el estudio presenta un análisis de 10 caminos entre sub-sitios que forman parte de la red SCC. Se realizó una prueba piloto con el fin de extraer todos los caminos cortos entre 10 pares de nodos seleccionados aleatoriamente. Los nodos de inicio fueron tomados de los componentes de entrada y los nodos destino fueron tomados de los componentes de salida. Esta prueba reveló que los caminos de enlaces resultantes están formados solo por un sub-sitio de entrada y un sub-sitio de salida. Todos los sub-sitios que forman parte del camino están localizados dentro de la red SCC. Además ningún enlace transversal fue identificado en el primer o en el último camino de los enlaces de muestra. Los caminos de los enlaces dentro de la red SCC contienen todos los links transversales o “topic drift”. Esta observación es de especial interés ya que la disertación está relacionada con el fenómeno “pequeño mundo”. [18]

La prueba piloto, además reveló que algunos pares de sub-sitios podrían ser conectados por muchos caminos cortos de la misma longitud. Una muestra de 10 pares de sub-sitios fue considerada a ser registrada con el objetivo de analizar todos los caminos cortos entre los sub-sitios acercándonos a la inspección de las paginas Web fuente y destino y los enlaces de nivel de pagina, los cuales fueron analizados posteriormente.

Las figuras 6.4 y 6.5 muestran dos de los 10 grafos resultantes que contienen los caminos cortos entre los pares de sub-sitios con tópicos diferentes.

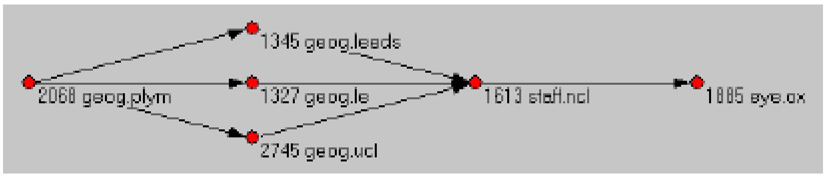


FIGURA 6.4. PATH NET HN05. TODOS LOS CAMINOS CORTOS ENTRE geog.plym.ac.uk Y eye.ex.ec.uk [18].

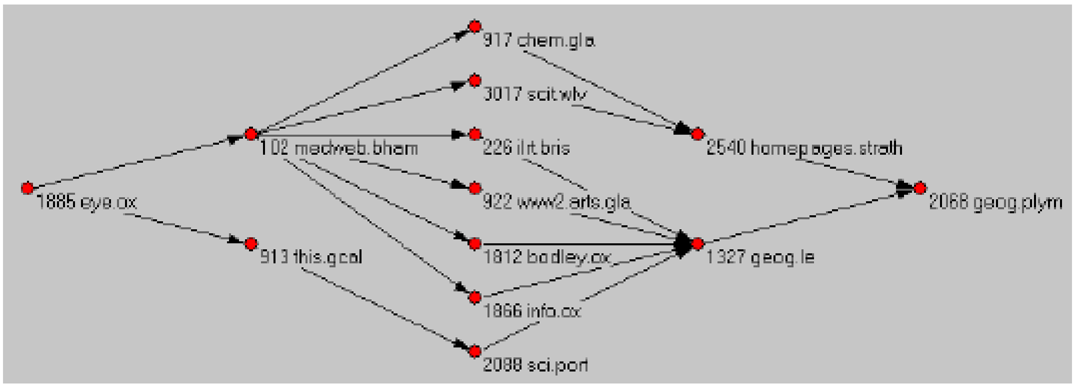


FIGURA 6.5. PATH NET NH05. TODOS LOS CAMINOS CORTOS ENTRE eye.ex.ec.uk Y geog.plym.ac.uk [18].

Hay que notar que cuando tanto el nodo de inicio como el nodo destino sobre una ruta pertenecen a la red SCC, todos los nodos intermedios también pertenecerán a dicha red.

Cabe recalcar que los problemas de redes cambiantes, enlaces rotos y páginas con cambios durante el crawl, fueron considerados en el estudio y resueltos gracias a la ayuda de uno o más “Internet

Archives” utilizados como una herramienta de “Web Arqueológica” o “Web Archaelogical”.

El estudio consideró también qué tipos de enlaces, páginas Web y sitios Web proveen de caminos cortos transversales a través de dominios de diferentes tópicos en una Web académica “pequeño mundo”.

El término transversal es usado para denotar enlaces de tópico cruzado, es decir de enlaces que dirigen de un tópico determinado hacia otro tópico diferente. Los enlaces de tópico cruzado contribuyen a la formación de propiedades “pequeño-mundo” en la forma de enlaces cortos sobre la Web.

De los 81 caminos seguidos, 17 contenían sub-sitios de tipo general y 58 (71.6%) sub-sitios relacionados con ciencias computacionales, incluyendo dos enlaces que tenían sub-sitios tanto generales como de ciencias computacionales. Sólo 8 (9.9%) de los 81 enlaces seguidos no contenían ninguno de los sub-sitios (general y ciencias computacionales). Por ejemplo sub-sitios que en su contenido combinan ciencias computacionales con ingeniería eléctrica, información de ciencia o matemáticas.

6.1.1.2 Estadísticas.

Las tablas presentadas a continuación muestran las estadísticas obtenidas para los enlaces de entrada y salidas para los sub-sitios dentro de la red SCC. Estos datos fueron obtenidos de [18].

	# sub-sites	# with inlinks	# in-neigh-bors	mean # in-neighbors / subsite	median # in-neighbors / subsite	range # in-neighbors / subsite	std. dev.
IN	626	36	43	0.07	0	0 – 3	0.3
SCC	1893	1893	34315	18.1	7	1 – 387	32.8
OUT	2660	2660	14421	5.4	2	1 – 411	16.2
IN-Tendrils	96	96	99	1.0	1	1 – 2	0.2
Tube	7	6	11	1.6	2	0 – 3	1.0
OUT-Tendrils	55	3	3	0.05	0	0 – 1	0.2
Disconnected	2332	10	10	0.004	0	0 – 1	0.06
	7669	**4704**	**48902**	**6.4**	**1**	**0 – 411**	**20.2**

TABLA 6.3. ESTADISTICAS DE ENLACES DE ENTRADA POR SUB-SITIO [18].

	# sub-sites	# with outlinks	# out-neigh-bors	mean # out-neighbors / subsite	median # out-neighbors / subsite	range # out-neighbors / subsite	std. dev.
IN	626	626	3953	6	3	1 – 163	13.4
SCC	1893	1893	44754	23.6	9	1 – 518	43.9
OUT	2660	83	111	0.04	0	0 – 7	0.3
IN-Tendrils	96	1	1	0.01	0	0 – 1	0.1
Tube	7	5	10	1.4	1	0 – 4	1.4
OUT-Tendrils	55	53	63	1.1	1	0 – 3	0.5
Disconnected	2332	10	10	0.004	0	0 – 1	0.06
	7669	**2671**	**48902**	**6.4**	**0**	**0 – 518**	**24.3**

TABLA 6.4. ESTADISTICAS DE ENLACES DE SALIDA POR SUB-SITIO [18].

6.1.1.3 Conclusión.

Entre las conclusiones de [18] más relevantes para nuestro estudio encontramos:

- La característica de la longitud del camino y el coeficiente de agrupamiento de la red de universidades del Reino Unido (UK) cumplen los requisitos de una red pequeño mundo. La longitud de camino fue 3.5 y el diámetro (distancia máxima entre enlaces) fue 10 entre los sub-sitios alcanzables.

- La red de universidades del Reino Unido presenta una escasa conectividad de enlaces, presentando un promedio de 11.6 enlaces de salida, incluyendo 10.1 de páginas que apuntan a otras páginas dentro del mismo sitio y 1.5 de páginas que apuntan a otras páginas de otros sitios. De estos últimos, solo el 7.7% fueron paginas enlazadas desde otros sitios de las 108 universidades y su sub-sitios.
- Se observó que las distribuciones de enlaces entrantes y salientes de los sub-sitios de la red de universidades del Reino Unido y los sub-sitios dentro de los 10 caminos seleccionados aleatoriamente poseen propiedades pequeño mundo. Esta observación concuerda con estudios anteriores que indican que los sub-sitios que son parte de una red, muestran las mismas propiedades de grafo que la red completa.
- Se encontró que la red de nodos centrales (SCC) poseían una distribución pequeño mundo.

Los nodos con alto grado de conectividad no tienden a conectar a otros nodos con muchas conexiones.

6.2 Análisis de los resultados.

6.2.1 Resultados por actividad.

6.2.1.1 Datos generados por la búsqueda.

Los datos obtenidos nos revelan la cantidad de enlaces que posee el sitio de la ESPOL en este caso podemos determinar que el sitios consta de 273294 enlaces aproximadamente, este archivo generado en formato de texto plano tiene un tamaño de 27 MB, en el visualizamos los enlaces sin filtro alguno. Finalmente para el procesamiento los datos se filtraron para obtener dos archivos contenían los enlaces con sus respectivos enlaces de entrada y salida.

6.2.1.2 Datos procesados por el algoritmo Map-Reduce.

Luego de procesados los datos nos presentan mayor información ya que es posible visualizar todas las ramificaciones del dominio y sus enlaces entrantes y salientes, además determinar los nodos concentradores. También podemos determinar la relación entre nodos principales, saber si ellos están conectados y si existen sitios sin enlaces salientes o entrantes ya que estos representan un problema para la distribución de contenido y conectividad entre enlaces.

6.2.1.3 Análisis del modelo de la red obtenido.

Inicialmente se conoce que para determinar que una red sigue una estructura “Pequeño Mundo” debe ser “Libre Escala” y por lo tanto, la distribución del grado de sus enlaces (entrantes y salientes) debe seguir una distribución de “ley de potencias”. Para este efecto tomamos los datos obtenidos de los enlaces y los tabulamos de manera que obtengamos el grado de enlaces y la cantidad de nodos con ese grado. Una vez establecida la tabulación utilizamos el proyecto para desarrollo de estadísticas llamado R-project, para averiguar si nuestra Web es “libre escala”.

Finalmente realizamos la ejecución de los comandos correspondientes para determinar si la distribución es la que buscamos. A continuación detallaremos el proceso seguido para determinar si nuestra red es libre escala.

- La primera fase de este proceso es realizar la tabulación de los datos.
- Convertirlo en un archivo en formato csv para poder cargarlo en el R-project.
- Luego utilizamos el R-project para generar la distribución de los enlaces.

- Comparar los valores y concluir

Detalle del proceso de demostración con R-project

Rproject utiliza comandos para generar las distribuciones. El primer comando es el siguiente: **X<- read.table(`archivo.csv`, sep=”,”)**

Con este comando cargamos el archivo para poder generar la distribución. Luego ejecutamos el comando **“do.power.law”** a los datos tabulados obteniendo el valor 0.8630296 para el alfa de la distribución y el siguiente gráfico de la distribución acumulada:

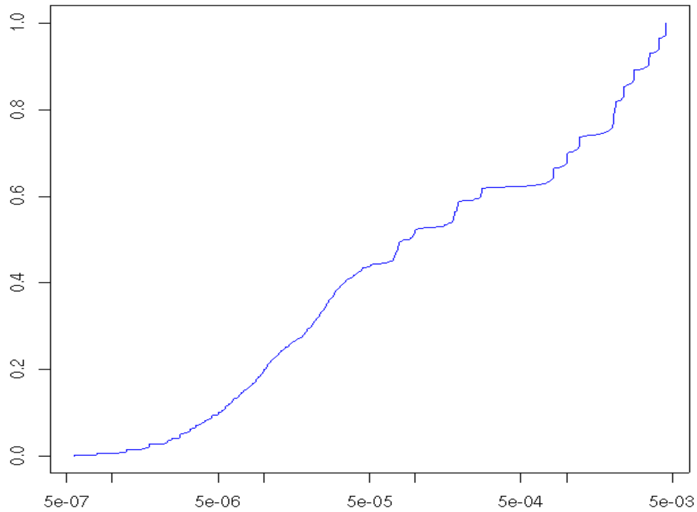


FIGURA 6.6 DISTRIBUCIÓN ACUMULADA DE ENLACES ENTRANTES.

Observemos que el valor del alfa de esta distribución es muy bajo, ya que se establece que para ser una red bien formada debe estar entre 2 y 3.Este proceso fue realizado con los enlaces de entrada y de salida de la red de la ESPOL de manera que podamos observar la tendencia de las distribuciones en ambas tabulaciones y de esta forma darle mayor credibilidad a este estudio.

La tabulación de enlaces de salida mostró un valor de alfa diferente, pero la distribución siguió la misma forma. El valor de alfa para esta distribución fue de: 0.76365. A continuación observaremos el gráfico de la distribución de los enlaces salientes.

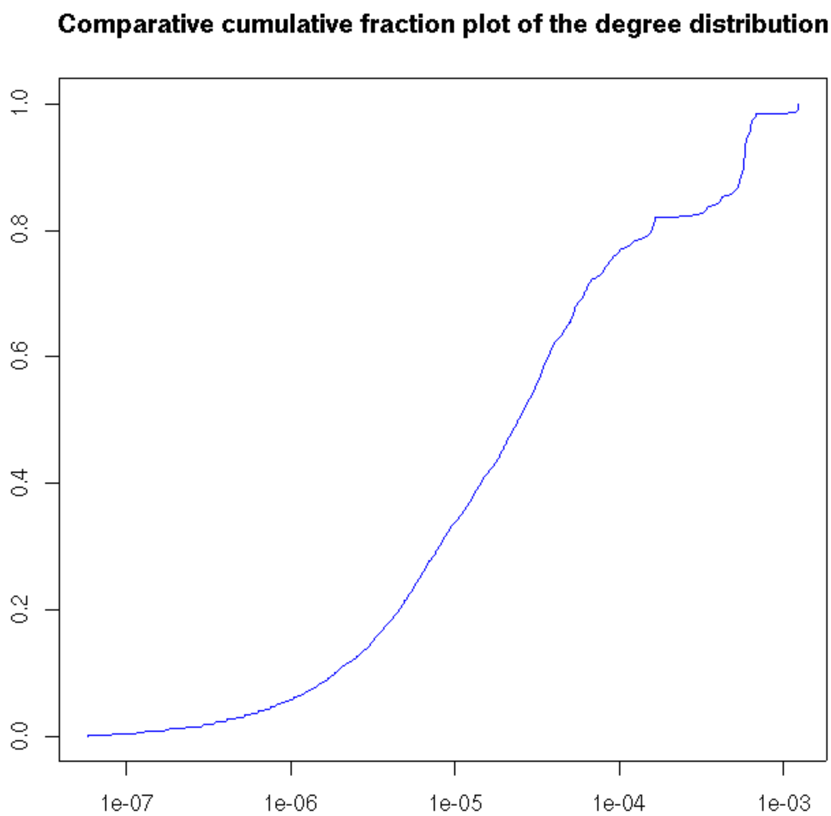


FIGURA 6.7 DISTRIBUCIÓN ACUMULADA DE ENLACES SALIENTES.

Con estas gráficas podemos demostrar que la Web de la ESPOL no es “libre de escala” y por lo tanto se puede inferir que no tiene una estructura “pequeño mundo”.

6.2.1.4 Comparaciones con el estudio de otro sitio.

El presente proyecto escogió como referencia el estudio de las universidades del Reino Unido para validar los resultados obtenidos así como las conclusiones.

Este estudio tiene como finalidad estudiar la estructura de la red de universidades y sus sub-sitios con el fin de identificar las propiedades pequeño mundo.

A continuación se detallarán las similitudes y diferencias entre ambos estudios:

- El estudio de la red de universidades del Reino Unido utilizó el Internet Archivo para realizar un estudio Arqueológico de la Web para analizar posibles enlaces rotos y contenido que hay sido modificado. En la red de la Espol no fue necesario tal análisis ya que al momento de la toma de datos la red en sí estaba siendo modificada en vías a mejorar la calidad de la navegabilidad de la misma y esta información sería relevante para identificar si los cambios que ha sufrido la red de la Espol han tenido algún impacto real en la navegabilidad del sitio.

- La red de universidades de Reino Unido y sus sub-sitios de marea independiente (es decir, también las redes de universidades independientes) tienen una forma pequeño mundo. La Web de la ESPOL no tienen una forma pequeño mundo.
- Los sub-sitios de las facultades de Ciencias Computacionales en la red de universidades de Reino Unido son concentradores de enlaces entrantes y salientes entre sub-sitios y de/hacia otras redes. La Web de la FIEC de la ESPOL no representa un concentrador de enlaces dentro ni fuera de ESPOL.

6.2.1.5 Otras actividades.

En la etapa de investigación fue necesario la exploración manual de la Web de la ESPOL para observar cual era el estado inicial de la misma y emitir nuestros criterios. Además para poder determinar si existían sitios aislados en la red que no pudieran ser alcanzados de manera alguna. También debemos considerar gracias a esto pudimos observar las mejoras realizadas a la Web.

CONCLUSIONES Y RECOMENDACIONES

Con los resultados obtenidos podemos concluir de la siguiente manera:

1. En este estudio se pudo determinar de manera específica que la red de la ESPOL no posee una forma pequeño mundo en este momento, pero que gracias a su interés en el mejoramiento de su información la estructura ha mejorado.
2. Es importante resaltar que en meses anteriores a la presente fecha, varias unidades de investigación de la ESPOL se encontraban aisladas y era casi imposible encontrarlas en la Web. Solamente era posible llegar a ellas si conocíamos previamente el enlace. En la actualidad es posible acceder desde diferentes sitios a muchos de estos grupos gracias al trabajo de re-estructuración y aplicación de mejoras prácticas en la Web de la ESPOL.
3. Hemos podido observar una positiva reestructuración en los sitios de la ESPOL. Aunque esto tuvo un impacto negativo en el presente estudio ya que produjo inconvenientes en el proceso de indexación. Esto ha mejorado la navegabilidad de la ESPOL, ya que ahora existe

al menos un enlace de entrada y salida en cada sitio de la Espol aunque aún no existe una correcta navegabilidad en muchos sitios, pues estos poseen problemas en su interacción con el usuario y no manejan un correcto enlazado de sitios.

4. Sabemos que el ranking de universidades que utiliza la ESPOL como referencia mide la usabilidad de los sitios Web y la cantidad de información investigativa que este proporcione a los usuarios. Por esto debemos coordinar una correcta estructura, de manera que se priorice la visibilidad de las investigaciones realizadas dentro de la institución y que esto sirvan para el desarrollo de actividades similares en otras instituciones.

Gracias a los resultados obtenidos en este estudio, podemos hacer algunas recomendaciones que podrán ser de ayuda al momento de reestructurar los sitios de la ESPOL, de manera que podamos obtener una estructura que posea una mejor navegabilidad.

1. Debido a la política de mejoramiento del sitio Web de la ESPOL es recomendable realizar un nuevo estudio en un periodo no mayor a un año, ya que considerando la reestructuración de algunos sitios de las diferentes unidades, la importancia que se le otorga a los papers científicos y a los diferentes centros de desarrollo e investigación, se

dará una nueva percepción de las mejoras en la estructura y la navegabilidad alcanzadas por estos cambios hasta esa fecha.

2. En todos los sub-sitios de la ESPOL debería agregarse en su página de inicio o Home un enlace hacia el sitio de la ESPOL y a otro sitio que muestre una lista de enlaces de sitios de similares características.
3. Debemos evitar en todo momento la creación de sitios Web que no estén enlazados directamente con algún sito representativo de la universidad, en este caso debería estar conectada a una unidad, facultad, centro de investigación, etc. que tenga alguna característica común con el sitio que se pretende crear.
4. Hacer un mantenimiento de los sitios de las unidades u otras representaciones dentro del dominio de la ESPOL de manera que no nos encontremos con sitios que no están activos o que carezcan de enlaces de entrada o salida.

ANEXOS

Anexo A: Instalaciones y Configuraciones

La instalación de las siguientes herramientas fue realizada en cada uno de los nodos del clúster.

Instalación de Java.

La instalación de Java se la hace mediante el siguiente comando:

apt-get install sun-java6-jdk

Edite el archivo /etc/profile y añada las líneas siguientes:

NUTCH_JAVA_HOME=/usr/lib/jvm/java1.6.0sun1.5.0.08

export NUTCH_JAVA_HOME

Verifique que NUTCH_JAVA_HOME enlaza con el directorio de instalación Java de nuestro sistema.

Instalación de Tomcat 5.

La instalación de Tomcat se lo hace a través del siguiente comando:

apt-get install tomcat5

Después de la instalación, debemos verificar que el servicio este funcionando correctamente, para lo cual abrimos un browser y digitamos:

http://localhost:8080

Si la instalación ha sido exitosa, el browser nos mostrara una página de bienvenida al servicio Tomcat.

Instalación de Hadoop.

Las fuentes de Hadoop pueden ser obtenidas desde este enlace.

La estructura de directorios que se definió en cada uno de los nodos del clúster es la siguiente:

/Nutch

/Nutch/filesystem (Hadoop)

/Nutch/search (Nutch)

Las máquinas que forman parte del clúster son las siguientes:

192.168.46.232 master (slave)

192.168.46.233 slave2

192.168.46.234 slave3

A continuación se descomprime el paquete y se mueve los archivos hacia el directorio /Nutch/filesystem:

tar -xzf Hadoop.0.19.tar.gz

cp Hadoop.0.19/* /Nutch/filesystem

chown Hadoop.Hadoop /Nutch/ -R

chmod 775 /Nutch/ -R

Archivos de configuraciones.

Hadoop.env

La única variable que necesita Hadoop es la de la variable de ambiente de java, esta puede ser especificada en /Nutch/filesystem/conf/Hadoop-env.sh por medio de las siguientes líneas:

The java implementation to use. Required.

export JAVA_HOME=/usr/lib/jvm/java1.6.0sun1.5.0.08

Hadoop-site.xml

Este archivo permite sobrescribir las configuraciones del archivo /Nutch/filesystem/conf/Hadoop-default.xml en el cual se encuentran las configuraciones por defecto. Edite el archivo /Nutch/filesystem/conf/Hadoop-site.xml para agregar las configuraciones deseadas. En este estudio se realizaron las siguientes configuraciones:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>
  <name>Hadoop.tmp.dir</name>
  <value/Nutch/filesystem//tmp/dir/Hadoop-${user.name}</value>
```

```
  <description>A base for other temporary directories.</description>
</property>

<property>
  <name>fs.default.name</name>
  <value>hdfs://master:54310</value>
  <description>The name of the default file system.  A URI whose
  scheme and authority determine the FileSystem implementation.  The
  uri's scheme determines the config property (fs.SCHEME.impl) naming
  the FileSystem implementation class.  The uri's authority is used to
  determine the host, port, etc. for a FileSystem.</description>
</property>

<property>
  <name>mapred.job.tracker</name>
  <value>master:54311</value>
  <description>The host and port that the MapReduce job tracker runs
  at.  If "local", then jobs are run in-process as a single map
  and reduce task.
  </description>
</property>

<property>
  <name>dfs.replication</name>
  <value>3</value>
  <description>Default block replication.
  The actual number of replications can be specified when the file is created.
  The default is used if replication is not specified in create time.
  </description>
</property>

</configuration>
```

Instalación y configuración de openSSH.

Hadoop requiere que los procesos puedan comunicarse de forma segura entre si con los diferentes equipos en la red, para esto es necesaria la instalación de SSH que permite los mecanismos de autenticación, esto se lo consigue a través del paquete openSSH. El paquete openSSH esta habilitado para la distribución de Ubuntu por lo cual solo es necesario el siguiente comando:

apt-get install openSSH

Se debe crear una cuenta de usuario bajo la cual podrán correr todos los procesos de Hadoop, por lo cual el nombre de la cuenta asignada para los procesos se llamo “Hadoop”, esto se lo hizo con el siguiente comando:

$ adduser Hadoop

A continuación se procede a realizar las configuraciones para permitir la comunicación entre los procesos.

noll@ubuntu:~$ su - Hadoop

Hadoop@ubuntu:~$ ssh-keygen -t rsa -P ""

Generating public/private rsa key pair.

Enter file in which to save the key (/home/Hadoop/.ssh/id_rsa):

Created directory '/home/Hadoop/.ssh'.

Your identification has been saved in /home/Hadoop/.ssh/id_rsa.

Your public key has been saved in /home/Hadoop/.ssh/id_rsa.pub.

The key fingerprint is: 9d:47:ab:d7:22:54:f0:f9:b9:3b:64:93:12:75:81:27 Hadoop@ubuntu

Hadoop@ubuntu:~$

Una vez generada la clave pública es necesario guardarla en el archivo de claves autorizadas para el usuario actual en cada una de las máquinas que forman parte del clúster:

Hadoop@ubuntu:~$ cat $HOME/.ssh/id_rsa.pub >>

$HOME/.ssh/authorized_keys

Finalmente, es necesario almacenar en cada máquina la clave personal asociada a su instalación de cada nodo del clúster. Esto se consigue accesando desde cada nodo hacia los demás nodos en el cluster:

Hadoop@ubuntu:~$ ssh slave2

The authenticity of host slave (192.168.46.233)' can't be established.

RSA key fingerprint is 76:d7:61:86:ea:86:8f:31:89:9f:68:b0:75:88:52:72.

Are you sure you want to continue connecting (yes/no)? yes

Warning: Permanently added ' slave2' (RSA) to the list of known hosts.

Ubuntu 7.04

...

Hadoop@ubuntu:~$

Se utilizo como guías para el proceso de instalación y configuración de Hadoop los siguientes manuales:

- http://wiki.apache.org/Hadoop/Running_Hadoop_On_Ubuntu_Linux_%28Single-Node_Cluster%29

- http://www.michael-noll.com/wiki/Running_Hadoop_On_Ubuntu_Linux_%28Multi-Node_Cluster%29

Instalación y configuración de Nutch.

Como habíamos mencionado anteriormente, hemos utilizado el paquete nutch-0.9.0 que puede ser descargado desde aquí.

El paquete fue descargado y descomprimido en el directorio /Nutch/search:

tar -xzf nutch-0.9.0.targ.gz

cp nutch-0.9.0/* /Nutch/search/

Archivos de configuraciones.

Hadoop.env

El archivo se puede encontrar en la siguiente ruta /Nutch/search/conf/Hadoop-env.sh. En este archivo debemos definir las dos variables de entorno que necesita Nutch para funcionar, Java y Hadoop.

The java implementation to use. Required.

export JAVA_HOME=/usr/lib/jvm/java1.6.0sun1.5.0.08

export HADOOP_HOME=/Nutch/filesystem/

craw-urlfilter.txt

Edite el archivo /Nutch/search/conf/crawl-urlfilter.txt.

En la línea de extensiones ignoradas, suprima las extensiones de los archivos que desea indexar. Las extensiones ejecutables ignoradas por defecto son Microsoft Powerpoint (ppt) y Excel (xls). Puede igualmente quitar los archivos de imágen de tipo jpg, gif, de todos modos sus nombres serán indexados.

 # skip image and other suffixes we can't yet parse \.(

gif|GIF|jpg|JPG|ico|ICO|css|sit|eps|wmf|zip|xls|

ppt|mpg|gz|rpm|tgz|mov|MOV|exe|png)$

En la línea # accept hosts in MY.DOMAIN.NAME, modifique y sustituya MYDOMAINNAME por su nombre de dominio. Para quitar la limitación al nombre de dominio, suprima el texto MYDOMAINNAME.

 # accept hosts in MY.DOMAIN.NAME

+^http://([az09]\.)*MYDOMAINAME/*

En este caso reemplazamos MYDOMAINAME por el dominio espol.edu.ec, de este modo debería verse de la siguiente forma:

 # accept hosts in MY.DOMAIN.NAME

+^http://([az09]\.)*espol.edu.ec/*

Por fin, en la última línea, si desea suprimir la limitación al nombre de dominio de la línea anterior, realice la modificación siguiente:

skip everything else

.

Sustituir por:

accept everything else

+.

nutch-default.xml

Edite el archivo /usr/lib/nutch/conf/nutch-default.xml, busque el parámetro **http.agent.name** e introduzca un mobre:

<name>http.agent.name</name>

<value>NUTCHCRAWLER</value>

Para eliminar la limitación de tamaño de los archivos analizados en http, busque el parámetro **http.content.limit** y sustituya el valor por defecto 65536 por 1.

<property>

<name>http.content.limit</name>

<value>1</

value>

Para suprimir la limitación del número de símbolos del indexador (que pueden truncar los documentos burocráticos de tipo Word o Excel), busque el parámetro **indexer.max.tokens**. La documentación de Nutch aconseja utilizar la variable *Integer.MAX_VALUE*. Pero después de numerosos ensayos, creemos que esta variable no debe ser tenida en cuenta. Para evitar este problema, se puede utilizar el valor máximo de un entero Java de 32 bits con signo, o sea, 2147483647.

<property>

<name>indexer.max.tokens</name>

<value>2147483647</value>

Para activar todos los plugins que permiten indexar los documentos burocráticos del tipo Microsoft Word, Excel, Powerpoint y los archivos PDF, busque el apartado plugin.includes, y añada pdf|msword|mspowerpoint|msexcel así como analysis-(fr) en la línea.

<property>

<name>plugin.includes</name>

<value> protocolhttp|urlfilterregex|parse-(text|html|js|pdf|msword|mspowerpoint|msexcel)|indexbasic|query-(basic|site|url)|summarybasic|scoringopic|analysis(fr)</value>

Archivos masters y slaves.

En el nodo master debemos editar el archivo /Nutch/filesystem/conf/masters. Este archivo contiene el nombre del host que será el nodo maestro; agregamos la siguiente línea:

master

En los tres nodos debemos indicarle a Nutch cuales serán los esclavos, para esto editamos el archivo /Nutch/filesystem/conf/slaves agregando las siguientes líneas:

master

slave2

slave3

Con las líneas anteriores, estamos indicando al HDFS que va a tener un maestro que será master y tres esclavos que serán master, slave2 y slave3. Anteriormente mencionamos que el nodo principal o maestro actuaría a la vez como esclavo. Esta configuración fue necesaria para aprovechar al máximo los recursos, ya que las tareas como maestro consumen menos recursos que las tareas como esclavos.

Levantando el HDFS.

Ahora vamos a levantar el sistema de archivos distribuido (HDFS) para lo cual desde la consola de comandos debemos ejecutar el siguiente comando:

$ /Nutch/search/bin/Hadoop namenode –format

Advertencia: Este comando borrara todos los datos dentro del sistema de archivos.

Esto nos permitirá formatear el sistema de archivos, lo cual es necesario para inicializarlo. Después de esto podemos ejecutar el siguiente comando:

$ /Nutch/search/bin start-all.sh

El comando arriba mencionado permite levantar todos los servicios relacionados con el sistema de archivos distribuido. A continuación se puede utilizar el comando jps en la consola:

$ jps

Este comando permite listar los procesos del HDFS que se están ejecutando y que deben ser los siguientes:

- NameNode
- SecondaryNameNode
- TaskTracker
- JobTracker
- DataNode

Los procesos Namenode, SecondaryNameNode y JobTracker son procesos que corren en el master, mientras que los procesos TaskTracker y DataNode deben de estar ejecutando en los esclavos. En el master se presentan los cinco procesos debido a que este nodo funciona tanto como maestro como esclavo.

Una vez levantados los procesos, el sistema de archivos distribuidos se encuentra habilitado para poder realizar el proceso de crawl.

Ejecutando un ejemplo: wordcount.

El ejemplo viene incluido como parte de los ejemplos de la herramienta Hadoop. El ejemplo tiene como objetivo leer un archivo de texto y devolver un archivo único con cada una de las palabras encontradas y su frecuencia en el archivo.

En este ejemplo utilizaremos los siguientes archivos (disponibles en la página del Proyecto Gutenberg):

- The Outline of Science, Vol. 1 (of 4) by J. Arthur Thomson
- The Notebooks of Leonardo Da Vinci
- Ulysses by James Joyce

Cada uno de estos archivos posee gran cantidad de información lo cual servirá para demostrar la rapidez del procesamiento del algoritmo Map Reduce.

Lo primero que debemos hacer es bajar los archivos y ponerlos en un directorio por ejemplo /tmp/gutenberg.

El siguiente paso es copiar estos archivos dentro del sistema de archivos distribuido, lo cual hacemos por medio del siguiente comando:

```
$ /Nutch/search/bin/Hadoop dfs -copyFromLocal /tmp/gutenberg gutenberg
```

Después podemos ejecutar los siguientes comandos para poder examinar los archivos en el sistema de archivos:

```
$ /Nutch/search/bin/Hadoop dfs -ls
  Found 1 items
  /user/Hadoop/gutenberg  <dir>
```

$ /Nutch/search/bin/Hadoop dfs -ls gutenberg

Found 3 items

/user/Hadoop/gutenberg/20417-8.txt <r 1> 674425

/user/Hadoop/gutenberg/7ldvc10.txt <r 1> 1423808

/user/Hadoop/gutenberg/ulyss12.txt <r 1> 1561677

Finalmente podemos ejecutar el ejemplo wordcount con el siguiente comando:

$ /Nutch/search/bin/Hadoop jar Hadoop-0.14.2-examples.jar wordcount gutenberg gutenberg-output

Una vez que el proceso haya concluido podemos ver los resultados con los comandos:

$ /Nutch/search/bin/Hadoop dfs -ls

Found 2 items

/user/Hadoop/gutenberg <dir>

/user/Hadoop/gutenberg-output <dir>

$ /Nutch/search/bin/Hadoop dfs -ls gutenberg-output

Found 1 items

/user/Hadoop/gutenberg-output/part-00000 <r 1> 903193

Utiice el siguiente comando para revisar el archive:

$ /Nutch/search/bin/Hadoop dfs -cat gutenberg-output/part-00000

El archivo part-00000 que es el resultado del proceso wordcount se encuentra guardado en el sistema de archivos distribuido y no podemos accesarlo localmente, para hacerlo debemos de copiarlo a nuestro sistema local, esto lo podemos hacer con la ayuda de los siguientes comandos:

```
$ mkdir /tmp/gutenberg-output
$ /Nutch/search/bin/Hadoop dfs -copyToLocal gutenberg-output/part-00000
/tmp/gutenberg-output
$ head /tmp/gutenberg-output/part-00000
 "(Lo)cra"      1
 "1490   1
 "1498," 1
 "35"    1
…..
```

Finalmente, obtenidos los resultados del proceso en nuestro sistema local, podemos detener los procesos para asi deshabilitar el HDFS. Esto lo conseguimos con el siguiente comando:

```
$ /Nutch/search/bin stop-all.sh
```

En la consola podemos ejecutar el comando jps para verificar que los procesos no se estén ejecutando.

Hadoop provee interfaces Web para visualizar el estado de los nodos y de los procesos que se están ejecutando, las direcciones son las siguientes:

- http://master:50030/ - MapReduce job tracker(s)
- http://master:50060/ - task tracker(s)
- http://master:50070/ - HDFS name node(s)

El ejemplo anterior nos ilustra la forma de trabajar y utilizar el HDFS. HDFS es un solo sistema de archivos pese a que está distribuido en los tres nodos del clúster. Esta es la razón por la cual si deseamos que el HDFS este habilitado necesitamos que estén corriendo todos los procesos mencionados arriba, en los tres nodos de la red. Las operaciones o tareas como wordcount se ejecutan sobre archivos que se encuentran dentro del HDFS por ello es necesario copiar nuestros archivos de datos dentro del sistema distribuido (copyFromLocal). Si los procesos no se ejecutan, dejan de ejecutarse o se los detiene, el HDFS deja de estar habilitado incapacitando la lectura o ejecución de procesos sobre los archivos que estén dentro de él, por ello la necesidad de copiar los archivos de resultados de nuestros procesos al sistema local (copyToLocal) antes de detener o apagar los procesos.

Ejecutando el Crawl.

Ahora veremos cómo ejecutar el crawl para hacer la indexación de los sitios dentro del dominio de la ESPOL, para lo cual debemos seguir los siguientes pasos:

Debemos levantar el sistema de archivos distribuido (HDFS) con el comando:

$ /Nutch/search/bin start-all.sh

Crear un archivo de texto /tmp/urls/urls.txt con los enlaces que el crawl debe buscar e indexar:

http://www.espol.edu.ec

En este caso vamos a realizar la búsqueda e indexación de todos los sitios dentro del dominio de la ESPOL.

Copiamos este archivo dentro del HDFS:

$ /Nutch/search/bin/Hadoop dfs -copyFromLocal /tmp/urls urls

Una vez que el directorio que contiene las urls a ser indexadas se encuentra dentro del HDFS podemos ejecutar la indexación:

$ /Nutch/search/bin/nutch crawl **urls** -dir **crawldir** -depth **500**

donde:

- **urls** es la carpeta que contiene el archivo urls.txt
- **crawldir** es una carpeta que va a crear Nutch para almacenar su base de datos con los índices de la búsqueda.
- 500 es la profundidad de indexación.

El comando mencionado arriba inicia el proceso de búsqueda e indexación de los enlaces dentro del dominio de la ESPOL. Podemos utilizar las interfaces Web que nos provee Hadoop para observar y monitorizar los procesos Map Reduce y el estado del HDFS y los nodos.

Nutch almacenara la información que obtenga de la búsqueda en indexación de los sitios en el dominio de la ESPOL en una base de datos, la cual estará disponible una vez concluido el proceso del crawl en el directorio crawldir.
Una vez que el proceso haya concluido podemos copiar el archivo en nuestro sistema local:
$ /Nutch/search/bin/Hadoop dfs -copyToLocal crawldir /tmp/crawldir

Ahora ya tenemos los datos necesarios de los índices para realizar los procesos y análisis de los enlaces del sitio de la ESPOL.

Opcionalmente, podemos utilizar los índices generados por Nutch para configurar la página Web de búsqueda de Nutch. Para esto es necesario configurar y levantar el servicio Tomcat5, esto lo hacemos de la siguiente forma:

Detener el servicio de tomcat5, si se encuentra levantado:

/usr/share/tomcat5/bin/catalina.sh stop

Elimine la carpeta ROOT del directorio de publicaciones de Tomcat5, por defecto este directorio se encuentra en /usr/share/tomcat5/webapps:

$ rm –rf /usr/share/tomcat5/webapps/ROOT/

Advertencia: Debe tener cuidado con el comando rm -rf ya que borra todo un árbol de directorios.

El siguiente paso es copiar el archivo nutch-0.9.war dentro de la carpeta webapps de Tomcat5 y renombrarla como ROOT.war, de este modo cuando iniciemos el servicio de Tomcat, este descomprimirá el archivo en una carpeta ROOT con todo el árbol Web de Nutch:

$ cp /Nutch/search/nutch-0.9.war /usr/share/tomcat5/webapps/

$ cd /usr/share/tomcat5/webapps/

$ mv nutch-0.9.war ROOT.war

Debemos configurar algunos parámetros de seguridad del servidor Tomcat para que Nutch funcione correctamente. La configuración se realiza en el archivo /usr/share/tomcat5/conf/catalina.policy. Al principio del archivo busque la sección **grant{** y agregue las siguientes lineas:

```
grant {
// Modifications pour Nutch
permission java.util.logging.LoggingPermission "control", "";
permission java.io.FilePermission "./*", "read,write,execute,delete";
permission java.util.PropertyPermission "user.dir", "read";
permission java.util.PropertyPermission "disableLuceneLocks", "read";
permission java.util.PropertyPermission "java.io.tmpdir", "read";
permission java.util.PropertyPermission "org.apache.*", "read";
permission java.io.FilePermission "/",
"read,write,execute,delete";
permission java.lang.RuntimePermission "createClassLoader", "";
```

Sitúese en el dir /tmp/crawldir y levante el servicio de Tomcat y después de unos segundos vuelva a detenerlo:

```
$ cd /tmp/crawldir
$ /usr/share/tomcat5/bin/catalina.sh start
$ /usr/share/tomcat5/bin/catalina.sh stop
```

Al ejecutarse el servicio de Tomcat se creo el directorio /usr/share/tomcat5/webapps/ROOT, ahora ya puede borrar el archivo /usr/share/tomcat5/webapps/ROOR.war

rm –f /usr/share/tomcat5/webapps/ROOT.war

Ahora vaya al directorio /usr/share/tomcat5/webapps/ROOT/WEB-INF/classes/nutch-default.xml y busque la sección searcher.dir y reemplace el valor crawldir por la ruta hacia el archivo de resultados de la indexación de Nutch (/tmp/crawldir):

<property>

<name>searcher.dir</name>

<value>/usr/lib/nutch/crawldir</value>

Haga lo mismo en el archivo /Nutch/search/conf/nutch-default.xml.

Ahora inicie el servicio de Tomcat 5:

/usr/share/tomcat5/bin/catalina.sh start

En el browser puede ir a la siguiente dirección http://master:8181. Ejecute una búsqueda y vea los resultados.

ECLIPSE

El desarrollo del codigo de la solucion MAP REDUCE fue escrito con la ayuda del compilador para Java Eclipse, el cual puede ser descargado desde su sitio oficial.

Una vez descargado el paquete es necesario descomprimirlo y ejecutarlo, ya que este no necesita ningún tipo de instalación.

Plugin de Hadoop para ECLIPSE

La instalación del plugin de Hadoop e integración con el entorno Eclipse es muy sencilla:

- Una vez descargado el plugin de Hadoop, debe ir a la carpeta de instalación de Eclipse y copiar el plugin dentro del directorio *plugins* dentro de la misma.
- Iniciar Eclipse e ir al menú Windows -> Show View -> Map Reduce Tools y seleccione Map Reduce Servers y acepte la selección.

El plugin de Hadoop mostrara una interface en la cual puede configurar los datos de conexión hacia un servidor Hadoop habilitado.

Anexo B: listado de sitios de la ESPOL

www.espol.edu.ec
www.admisión.espol.edu.ec
www.académico.espol.edu.ec
www.mail.espol.edu.ec
www.icm.espol.edu.ec
www.fiec.espol.edu.ec
www.blog.espol.edu.ec
www.cenaim.espol.edu.ec
www.espae.espol.edu.ec
www.dspace.espol.edu.ec
www.fimcm.espol.edu.ec
www.laclo.espol.edu.ec
www.kokoa.espol.edu.ec
www.rte.espol.edu.ec
www.cvr.espol.edu.ec
www.fimcp.espol.edu.ec
www.fen.espol.edu.ec
www.ceemp.espol.edu.ec
www.cibe.espol.edu.ec
www.celex.espol.edu.ec
www.iepse.espol.edu.ec
www.iptv.cti.espol.edu.ec
www.proyectos.ssw.espol.edu.ec
www.ceproem.espol.edu.ec
www.msig.espol.edu.ec
www.ccma.espol.edu.ec
www.icf.espol.edu.ec
www.icm.espol.edu.ec
www.espae.espol.edu.ec
www.cdts.espol.edu.ec
www.aefimcp.espol.edu.ec
www.encuestas.ceproem.espol.edu.ec
www.cec.espol.edu.ec
www.vicerrectorado.espol.edu.ec
www.vlir.espol.edu.ec
www.icqa.espol.edu.ec
www.cise.espol.edu.ec
www.calidadyevaluacion.espol.edu.ec
www.aja.espol.edu.ec
www.ceap.espol.edu.ec
www.csi.espol.edu.ec

www.cib.espol.edu.ec
www.cti.espol.edu.ec
www.cenacad.espol.edu.ec
www.sidweb.espol.edu.ec
www.intranet.espol.edu.ec
www.abet.espol.edu.ec
www.espolciencia.espol.edu.ec
www.iyd.espol.edu.ec
www.focus.espol.edu.ec
www.cicyt.espol.edu.ec
www.edcom.espol.edu.ec
www.fiec.nogal.espol.edu.ec
www.calendario.espol.edu.ec
www.transparencia.espol.edu.ec
www.ctt.espol.edu.ec
www.jntt.espol.edu.ec
www.msia.espol.edu.ec
www.lictur.espol.edu.ec
www.finanzas.espol.edu.ec
www.espolinforma.espol.edu.ec
www.proyectoancon.espol.edu.ec
www.protmec.espol.edu.ec
www.visid.espol.edu.ec
www.relex.espol.edu.ec
www.wiki.espol.edu.ec
www.fundespol.espol.edu.ec
www.ecoproyectos.espol.edu.ec
www.vlir8.espol.edu.ec
www.protal.espol.edu.ec
www.rte.espol.edu.ec
www.protep.espol.edu.ec
www.protel.espol.edu.ec
www.taws.espol.edu.ec

Anexo C: Tabulaciones de enlaces de entrada y salida

Aquí se muestra la tabulación de los enlaces de entrada de manera que podamos observar la cantidad de nodos con un grado determinado de enlaces.

CANTIDAD NODOS	GRADO
2172	1
4271	2
8131	3
7296	4
4981	5
1803	6
3650	7
3746	8
1467	9
6281	10
344	11
492	12
138	13
135	14
178	15
76	16
325	17
48	18
59	19
44	20
28	21
7	22
17	23
6	24
6	25
11	26
10	27
5	28
8	29
45	30
8	31
4	32
1	33
3	34
5	35
4	38
1	39
1	40
1	42
1	43
1	44
4	47

2	48
2	58
1	59
12	79

TABLA C.1 TABULACIÓN DE ENLACES ENTRANTES.

Aquí se muestra la tabulación de los enlaces de salida de manera que podamos observar la cantidad de nodos con un grado determinado de enlaces.

CANTIDAD NODOS	GRADO
2	0
411	1
95	2
36	3
3	4
5	5
5	6
3	7
3	8
4	9
4	11
3	12
5	13
3	14
2	15
2	17
1	18
2	19
2	20
2	21
1	22
3	25
1	26
1	29
1	30
1	32
1	34
2	35
1	36
1	38
2	39
1	40
2	41
1	47

1	60
1	61
1	63
2	67
2	70
1	71
1	72
1	81
2	83
1	85
1	90
1	91
2	93
1	96
2	102
2	113
1	114
2	115
1	118
2	119
2	120
2	122
1	128
1	144
1	149
1	153
2	154
1	156
1	161
2	167
3	179
1	180
1	198
1	222
1	227
2	228
1	233
1	246
2	260
1	261

2	276
1	313
1	343
1	359
2	360
1	374
1	380
1	392
1	417
1	475
1	476
1	542
1	545
1	571
1	596
1	599
1	613
2	638
1	657
1	670
1	690
1	702
1	797
1	924
1	928
1	934
1	941
1	1071
1	1120
1	1127
1	1169
1	1418
1	1508
1	1595
1	1762
1	2091
1	2781
1	2852
1	2854
1	5937
1	7211

1	9253
1	9940
1	9999
10	10000
1	10012
1	10306
1	21287

TABLA C.2 TABULACIÓN DE ENLACES SALIENTES.

Anexo D: Visualizadores

Graphviz

Instalación

Descargar el paquete de Graphviz:
http://www.graphviz.org/Download..php

Ejecutar el comando en la terminal:

#apt-get install graphviz graphviz-dev graphviz-doc

Los archivos .dot tienen la siguiente sintaxis:

```
/*Esto es un comentario*/
graph nombre_del_grafo {
   "idNodo1";
   "idNodo2"; /*estos son identificadores de nodos*/
   "idNodo3";
   "idNodo1" -- "idNodo2";
   "idNodo1" -- "idNodo3"; /*estas son relaciones entre nodos*/
   "idNodo3" -- "idNodo1";
}
```

Para compilar el siguiente código escribimos en la terminal
$dot ejemplo1.dot -o ejemplo1.png -Tpng -Gcharset=latin1

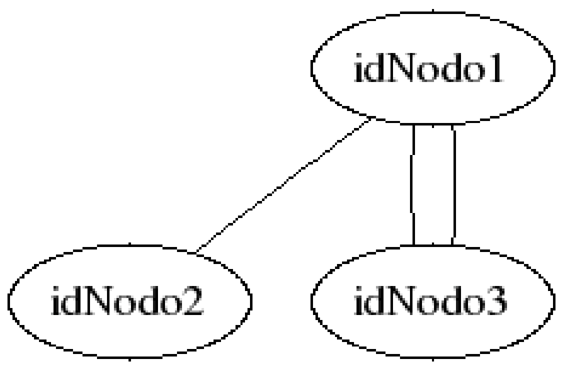


Figura D.1. Resultado de compilación del ejemplo [32].

Ejemplo:

```
Graphviz Example Invoking Circo =
{{{
#!graphviz.circo
digraph G {
   Hello->World
   Hello->Goodbye
   World->Graphviz
   Graphviz->Rules
}
}}}
```

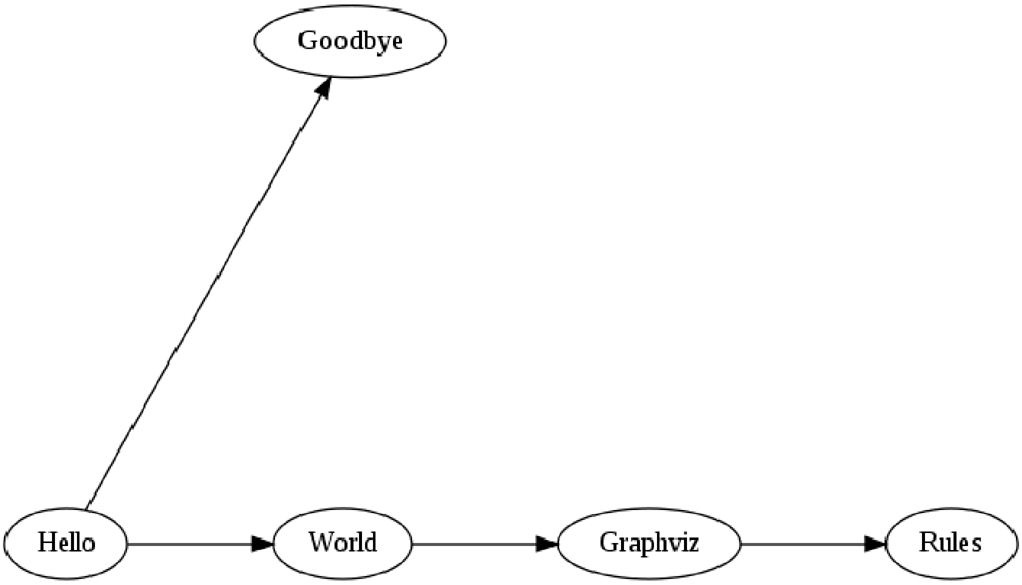


Figura D.2. Grafica de ejemplo Invorking Circo [32].

Archivo de la Red Espol

```
digraph G {

        URL1 -> URL3
        URL1 -> URL9
        URL1 -> URL12
        URL1 -> URL10
        URL1 -> URL2
        URL1 -> URL13
        URL1 -> URL11
        URL1 -> URL14
        URL1 -> URL15
        URL1 -> URL8
        URL2 -> URL3
        URL2 -> URL9
        URL2 -> URL12
        URL2 -> URL10
        URL2 -> URL13
        URL2 -> URL11
        URL2 -> URL14
        URL2 -> URL15
        URL2 -> URL8
        URL3 -> URL9
        URL3 -> URL12
        URL3 -> URL10
        URL3 -> URL2
        URL3 -> URL13
        URL3 -> URL11
        URL3 -> URL14
        URL3 -> URL535
        URL3 -> URL15
        URL3 -> URL8
        URL9 -> URL3
        URL9 -> URL12
        URL9 -> URL10
        URL9 -> URL2
        URL9 -> URL13
        URL9 -> URL11
        URL9 -> URL14
        URL9 -> URL15
        URL9 -> URL8
        URL10 -> URL3
        URL10 -> URL9
        URL10 -> URL12
        URL10 -> URL2
        URL10 -> URL13
        URL10 -> URL11
        URL10 -> URL14
        URL10 -> URL15
        URL10 -> URL8
        ………………………
        URL746 -> URL718

}
```

LaNet-vi

Instalación

En este caso en particular no necesitamos instalarlo, ya que cuando la utilizamos solo presentaba la opción de subir los archivos de redes para que ellos nos envíen los gráficos resultantes por vía mail. Esto resultó complicado a medida que los archivos resultaban más grandes por lo cual tuvimos que desistir del uso de la misma.

Ejemplo:

1 2
2 3
2 4
3 5
4 5
3 4
2 5
4 6
6 7
7 8
10 11
......

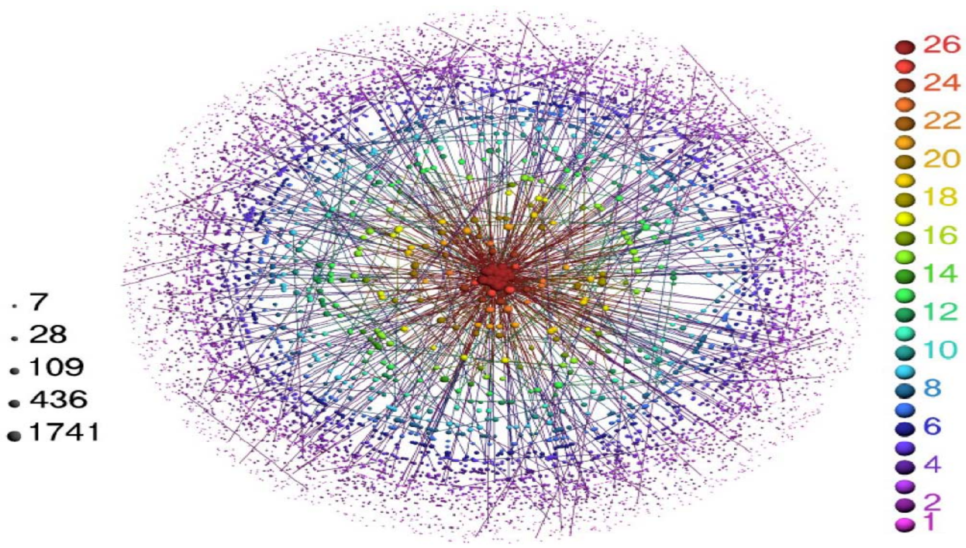


Figura D.3. Grafica de ejemplo LaNet-vi [33].

Archivo de la Red Espol

242594	177841
98897	196372
98897	196372
98897	196372
98897	196372
98897	196372
198756	196372
196340	196372
62032	141747
163743	163744
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
196340	232580
181832	83240
29627	83240
65365	83240
65365	83240
65365	83240
65365	83240
50693	83240
242595	83240
177820	180537
50693	180537
50693	180537
50693	180537
60848	180537
242595	180537
242595	180537
198756	180537
65365	248879
43848	248879
163767	248879
48601	248879

……………………………..

Cytoscape

Instalación

Requerimientos de Hardware:

	Visualización de redes Pequeñas	Analisis/visualización de Grandes Redes
Procesador	1GHz	As fast as possible
Memoria	512MB	2GB+
Graphics Card	On board Video	Highend Graphics Card
Monitor	XGA (1024X768)	Wide or Dual Monitor

Tabla D.1. Requerimientos de Hardware de Cytoscape. Tomado de: http://www.cytoscape.org/cgi-bin/moin.cgi/Cytoscape_User_Manual/Launching_Cytoscape

Este necesita versiones de java SE 5 ó 6. El paquete cytoscape.sh nos permite la instalación en Linux, solo necesitamos ejecutar el script para utilizarlo.

Para mayor información de instalación y ejemplos consultar el manual:
http://www.cytoscape.org/manual/Cytoscape2_6Manual.html#Launching%20Cytoscape

Ejemplo:

42429 181397
42429 180799

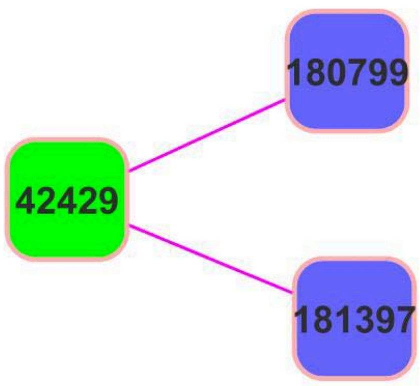


Figura D.4. Gráfica de Enlaces de Entrada de la Espol.

Archivo de la Red Espol

177841 178823
177841 273275
177841 178825
177841 178826
177841 179017
177841 179019
177841 179020
177841 179022
177841 179021
177841 179023
177841 179952
177841 179025
177841 178834
177841 179028
177841 179030
177841 179031
177841 178835
177841 179032
177841 179035
177841 242596
177841 178837
177841 178838
177841 179038
177841 273271
177841 179041
177841 179042
177841 178842
177841 179043
177841 178844
177841 179047
177841 179048
177841 179049
177841 178847
177841 178849
……………………………………………
177841 179074

Formato de archivos Originales de Enlaces de la Red de la Espol obtenidos del Map-Reduce.

```
http://200.10.148.7/plan
            http://www.espae.espol.edu.ec/
(1 links)
http://200.10.150.109:5080/demos/simpleSubscriberIPTV.swf
            http://iptv.cti.espol.edu.ec/index.htm
      http://iptv.cti.espol.edu.ec/index.htm
      http://iptv.cti.espol.edu.ec/
      http://iptv.cti.espol.edu.ec/
(4 links)
http://academias.espol.edu.ec/
            http://www.fiec.espol.edu.ec/
(1 links)
http://caos.cti.espol.edu.ec/recursos/reservacion.nsf/Reservaciones?
openView
            http://www.cti.espol.edu.ec/historia.php
      http://www.cti.espol.edu.ec/login.php
      http://www.cti.espol.edu.ec/objetivos.php
      http://www.cti.espol.edu.ec/preguntas.php
      http://www.cti.espol.edu.ec/normas_informacion.php
      http://www.cti.espol.edu.ec/politica.php
      http://www.cti.espol.edu.ec/proyectos.php
      http://www.cti.espol.edu.ec/servicios.php
      http://www.cti.espol.edu.ec/
      http://www.cti.espol.edu.ec/contactenos.php
      http://www.cti.espol.edu.ec/suscripcion.php
      http://www.cti.espol.edu.ec/index.php
      http://www.cti.espol.edu.ec/preguntas.php
      http://www.cti.espol.edu.ec/estadisticas.php
(14 links)
http://ccma.cti.espol.edu.ec/

      http://www.vicerrectorado.espol.edu.ec/index.php/static/index/p
age/cti
(1 links)
http://ccma.cti.espol.edu.ec/actividades.php
            http://ccma.cti.espol.edu.ec/
(1 links)
http://ccma.cti.espol.edu.ec/contactenos.php
            http://ccma.cti.espol.edu.ec/
(1 links)
http://ccma.cti.espol.edu.ec/css/main.css
            http://ccma.cti.espol.edu.ec/
(1 links)
http://ccma.cti.espol.edu.ec/iflateng.htm
            http://ccma.cti.espol.edu.ec/
(1 links)…………………………………………………………….
……………………………………………………………
```

Anexo E: Herramienta Estadística

La herramienta R-Project fue el instrumento que nos permitió demostrar que la red de la ESPOL no posee una estructura pequeño mundo, ya que para este proceso debíamos realizar la tabulación de los enlaces de entrada y salida y con ellos verificar si siguen la distribución de ley de potencias. Como podemos observar en el Anexo anterior las tabulaciones de enlaces son archivos extensos de datos, así que procedimos a la instalación de R-Project.

La instalación del mismo es sencilla, el paquete lo podemos encontrar en la página http://cran.patan.com.ar/ . Obtenido el paquete tenemos que descomprimirlo en algún directorio de Linux para luego ejecutarlo. También es posible hacerlo de manera automática descargándolo directamente de un repositorio de Ubuntu. Para esto vamos a seguir los pasos a continuación.

Entrando en un terminal vamos a modificar el archivo de configuración de sources.list para poder obtener las claves necesarias para la instalación.

$sudo gedit /etc/apt/sources.list

De esta manera vamos a permitir que R se instale desde un repositorio oficial de Ubuntu, copiamos en el archivo sources.list que está abierto cualquiera de los repositorios para luego guardar los cambios y salir.

Para Dapper:

#REPOS R-CRAN

deb http://trapananda.homelinux.org/r-project/bin/linux/ubuntu dapper/

Para Edgy:

#REPOS R-CRAN

deb http://trapananda.homelinux.org/r-project/bin/linux/ubuntu edgy/

El siguiente paso es obtener la llave pública del repositorio. Te recomendamos intentarlo varias veces si te retorna en consola el mensaje: No fue posible tener acceso al lugar.

gpg —keyserver subkeys.pgp.net —recv-key E2A11821

gpg -a —export E2A11821 | sudo apt-key add -

Terminado este paso ya estamos listos para la instalación del programa, vamos a descargar la herramienta.

sudo apt-get update

sudo apt-get install r-base r-recommended

La descarga puede tardarse un poco, en el caso de que lo descargues directo del enlace anterior debes seguir los pasos a continuación.

$./configure —enable-R-shlib

$make

$sudo make install

Finalmente colocamos R en el terminal y está listo para desarrollar cualquier tipo de cálculo estadístico. Las librerías que tiene esta herramientas son muchas cada una de ellas para una determinada tarea. En este estudio nos servimos de la librería **netmodels** la cuál necesitaba de **igraph**, por esto fue necesaria la instalación de ambos paquetes de librerías.

La librería **netmodels** nos permitía el estudio de las redes, específicamente de los sitios Web, nos prestaba como herramienta principal comandos como.

Do.power.law: el mismo que nos realizaba una aproximación a una distribución power law de manera que podíamos determinar si una red tenia propiedades libre escala.

Además, nos permitía calcular valores como el alfa de las distribuciones, también presentaba diversas opciones para el estudio específico de enlaces. De manera que podamos conocer la estructura en mayor medida.

Esta es una herramienta muy completa y presenta una diversidad de opciones en el estudio de estadística y el análisis de redes extraordinario. A continuación presentamos algunos de los valores obtenidos con esta herramienta.

c->plot(tabulacion): nos devuelve el grafico de la tabulación inicial.

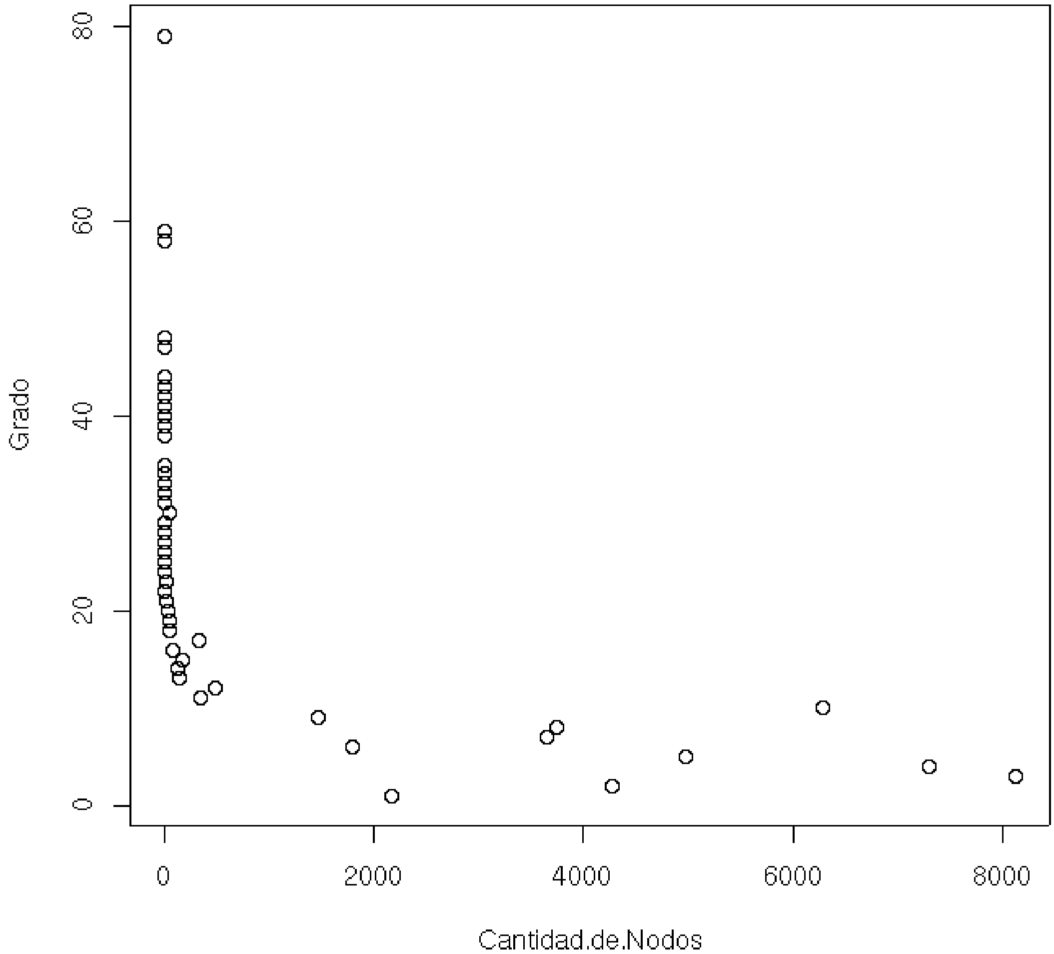
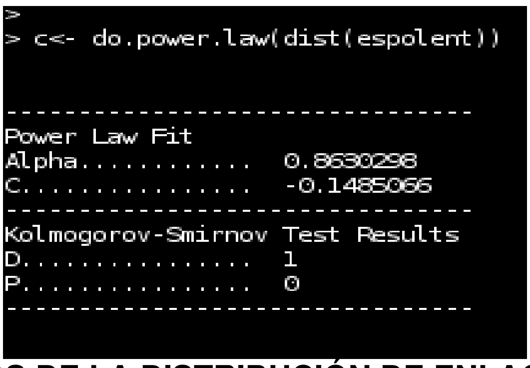


FIGURA E.1 GRAFICA DE LA TABULACION DE ENLACES ENTRANTES.



```
>
> c<- do.power.law(dist(espolent))


--------------------------------
Power Law Fit
Alpha............  0.8630298
C................  -0.1485066
--------------------------------
Kolmogorov-Smirnov Test Results
D................  1
P................  0
--------------------------------
```

FIGURA E.2 DATOS DE LA DISTRIBUCIÓN DE ENLACES ENTRANTES.

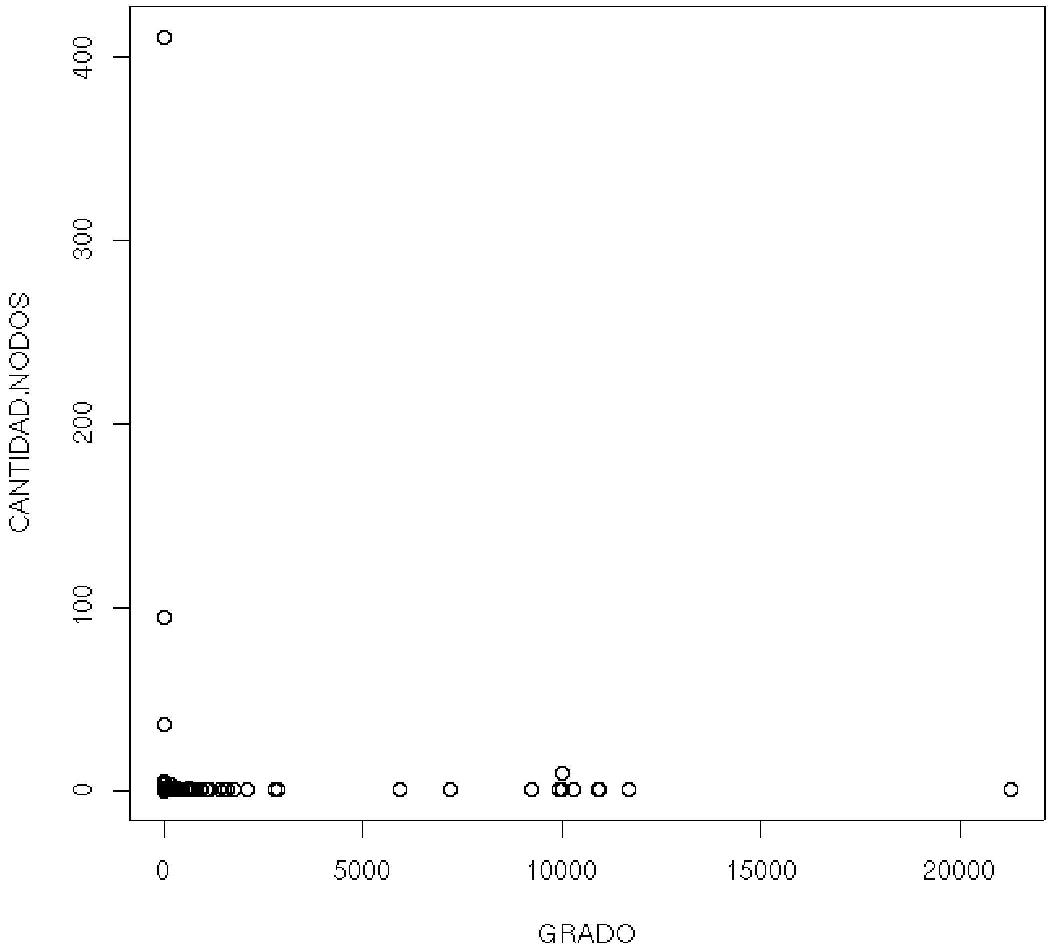


FIGURA E.3 GRAFICA DE LA TABULACION DE ENLACES SALIENTES.

```
> evaluacion<- do.power.law(dist(espolsalidas))


--------------------------------
Power Law Fit
Alpha............  0.7636595
C................  -0.2723724
--------------------------------
Kolmogorov-Smirnov Test Results
D................  1
P................  0
--------------------------------
```

FIGURA E.4 DATOS DE LA DISTRIBUCIÓN DE ENLACES SALIENTES.

BIBLIOGRAFÍA

[1] Aguillo, Isidro F. "World Universities' ranking on the Web". http://www.webometrics.info/. Fecha de última visita: Octubre 8 de 2007.

[2] Aguillo, I. F.; Granadino, B.; Ortega, J. L.; Prieto, J. A. "Scientific research activity and communication measured with cybermetric indicators". Journal of the American Society for the Information Science and Technology, 57(10): 1296 - 1302.

[3] Proyecto Lucene. "Hadoop". http://lucene.apache.org/Hadoop/. Fecha de última visita: Octubre 8 de 2007.

[4] V. Zlatic, M. Bozicevic, H. Stefancic and M. Domazet, "Wikipedias: Collaborative Web-Based Encyclopedias as Complex Networks", Physical Review E, vol. 74, Issue 1, 2006.

[5] A. Capocci, V. Servidio, F. Colaiori, L. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, 2006. "Preferential Attachment in the Growth of Social Networks: The Case of Wikipedia," Physical Review E, vol. 74, Issue 1, 2006.

[6] S. Spek, "Wikipedia: organisation from a bottom-up approach". In Proceedings of Research in Wikipedia workshop (WikiSym 2006). 2006.

[7] Wikipedia: The Free Encyclopedia. "Wikipedia: Six degrees of Wikipedia". Disponible en línea en: http://en.wikipedia.org/wiki/Wikipedia:Six_degrees_of_Wikipedia. Fecha de última visita: Septiembre 22 de 2008.

[8] Abad, C. "El Fenómeno Pequeño Mundo en la Naturaleza, Redes Sociales y Redes de Computadoras". Charla magistral a ser dictada en las Jornadas de Estadística e Informática, organizadas por el ICM, ESPOL. Resumen de charla adjunto. Octubre, 2007.

[9] Varabais, A-L. "Linked: The new science of Networks". Perseus Books Group. 1era Edición. Mayo, 2002.

[10] Referencia hacia la página del proyecto Graphviz en: http://www.graphviz.org/. Fecha de última visita: Septiembre 22 de 2008.

[11] Wikipedia: The Free Encyclopedia. "Concepto de computación Distribuida" en: http://es.wikipedia.org/wiki/Computacion_distribuida. Fecha de última visita: Septiembre 22 de 2008.

[12] Referencia hacia el sitio oficial del proyecto Lucene en: http://lucene.apache.org/. Fecha de última visita: Septiembre 22 de 2008.

[13] Referencia hacia el sitio oficial del proyecto Nutch en: http://lucene.apache.org/nutch/about.html. Fecha de última visita: Septiembre 22 de 2008.

[14] Tutorial de configuración de la plataforma Hadoop con el motor de búsqueda Nutch en: http://wiki.apache.org/nutch/NutchHadoopTutorial. Fecha de última visita: Septiembre 22 de 2008.

[15] Wikipedia: The Free Encyclopedia. "Referencia del IDE Eclipse" en: http://es.wikipedia.org/wiki/Eclipse_(software). Fecha de última visita: Septiembre 22 de 2008.

[16] Plugin de Eclipse para la herramienta Map Reduce de Hadoop en: http://www.alphaworks.ibm.com/tech/mapreducetools. Fecha de última visita: Septiembre 22 de 2008.

[17] Nwagwu, Williams E., & Agarin, Omoverere (2008). "Nigerian University Websites: A Webometric Analysis" *Webology*, **5**(4), Artículo 65. Disponible en: http://www.webology.ir/2008/v5n4/a65.html. Fecha de última visita: Septiembre 22 de 2009.

[18] Björneborn, Lennart. "Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach" Tesis Doctoral, del Departamento de Estudios de la Información, Royal School of Library and Information Science, Dinamarca. Disponible en línea en: http://vip.db.dk/lb/phd/. Fecha de última visita: Septiembre 22 de 2009.

[19] Paper "The structure and function of complex networks" en: http://www-personal.umich.edu/~mejn/courses/2004/cscs535/review.pdf. Fecha de última visita: Septiembre 22 de 2009

[20] Barabasi, Albert-Laszlo. “Linked: How Everything Is Connected to Everything Else and What It Means”. Plume, reissue edition, April 2003.

[21] J. Dean y S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in OSDI'04: Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation. Berkeley, CA, USA: USENIX Association, 2004, p. 10. [Online]. Available: http://portal.acm.org/citation.cfm?id=1251254.1251264. Fecha de última visita: Septiembre 22 de 2009.

[22] Tom White. “Hadoop: The Definitive Guide“. 1era edición Junio 2009

[23] Coulouris, George; Dollimore, Jean; y Kindberg, Tim. “Distributed Systems: Concepts and Design”. Addison-Wesley. 4ta edición. 2005.

[24] Ejemplo de ejecución de Map-Reduce (WordCount) en: http://wiki.apache.org/Hadoop/C%2B%2BWordCount. Fecha de última visita: Septiembre 22 de 2009

[25] Smith, Andrew. “SMALL WORLD, BIG COSMOS: The Role of Scale-free and Other Networks in Hierarchical Organization”. Ensayo. Disponible en línea en: http://www.geocities.com/andybalik/network.html. Fecha de último acceso: Octubre 15 de 2009.

[26] Batagelj, V. y Zaversnik, M. “Generalized Cores”. Disponible en línea en: http://arxiv.org/PS_cache/cs/pdf/0202/0202039v1.pdf. Fecha de último acceso: Octubre 15 de 2009.

[27] J. Voss, "Measuring wikipedia," in Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics, July 2005.

[28] "Redes libres de escala". Disponible en línea en: http://www.wikilearning.com/curso_gratis/redes_libres_de_escala_un_tipo_de_red_con_muchos_ejemplos_y_aplicaciones-redes_libres_de_escala/6072-1. Fecha de último acceso: Octubre 15 de 2009.

[29] J. I. Alvarez-Hamelin, L. Dall'asta, A. Barrat, and A. Vespignani, "k-core decomposition: a tool for the visualization of large scale networks," arXiv, Oct 2005. Disponible en línea en: http://arxiv.org/abs/cs/0504107.

[30] Esquema de Computación Distribuida "Computación Distribuida". Disponible en línea en: http://sig.utpl.edu.ec/download/data/computacion%20distribuida.PDF. Fecha de último acceso: Octubre 15 de 2009.

[31] "HDFS Architecture". Disponible en línea en: http://Hadoop.apache.org/common/docs/r0.19.2/hdfs_design.html. Fecha de último acceso: Octubre 15 de 2009.

[32] "Ejemplos de Graphviz". Disponible en línea en: http://plataforma.cenditel.gob.ve/wiki/EjemplosGraphviz/03. Fecha de último acceso: Octubre 15 de 2009.

[33] "Lanetvi". Disponible en línea en:

[34] http://ortegaalfredo.googlepages.com/lanetvi.pdf. Fecha de último acceso: Octubre 15 de 2009.

[35] Referencia de "Small World". Disponible en línea en: http://en.wikipedia.org/wiki/Small_world_experiment. Fecha último acceso: 21 Noviembre de 2009.

[36] "Redes libres de escala". Disponible en línea en: http://es.wikipedia.org/wiki/Red_libre_de_escala. Fecha de último acceso: Lunes 11 de Enero 2010.

[37] "Power Law". Disponible en línea en: http://en.wikipedia.org/wiki/Power_law. Fecha de último acceso: Lunes 11 de Enero 2010.