Universidad de las Ciencias Informáticas.

Facultad 10.

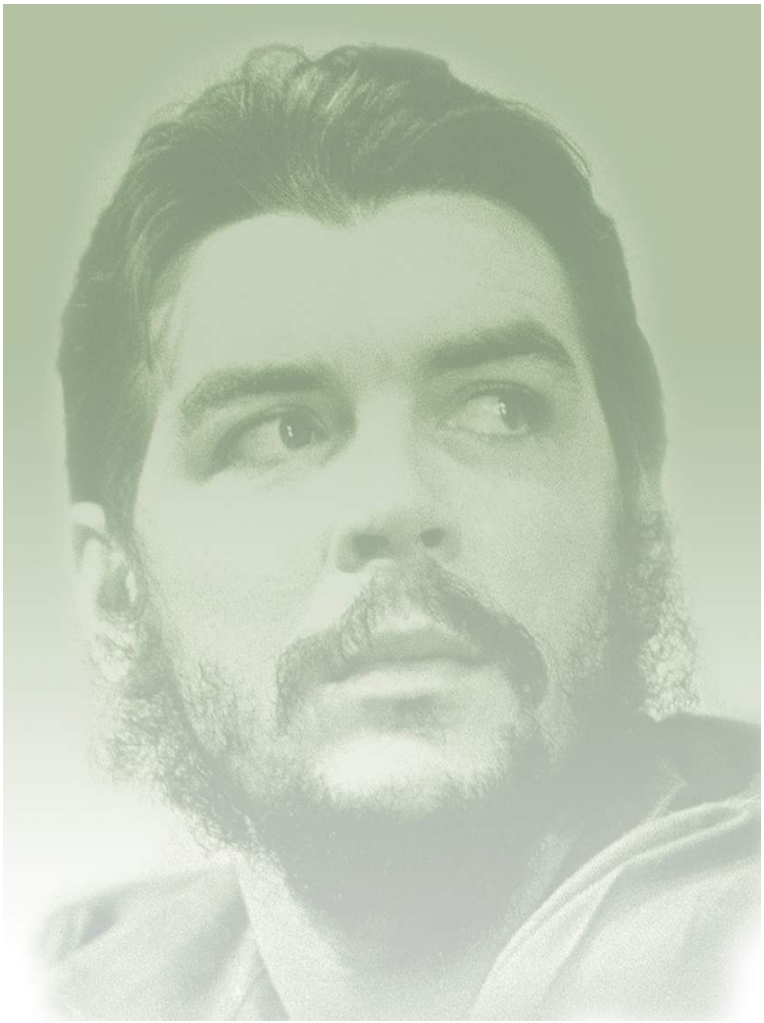
Título: Características de la Web de la Universidad de las Ciencias Informáticas.

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.

***Autor* (es):**
Yonny Mondelo Hernández
Yuley Díaz Madruga

***Tutor(es)*:**
Ing. Dovier Antonio Ripoll Méndez

Ciudad de La Habana, 2009.



“El revolucionario verdadero está guiado por grandes sentimientos de amor”.
“Seamos realistas, y hagamos lo imposible”.
Ernesto Guevara de la Serna.

DECLARACIÓN DE AUTORÍA.

Declaran ser los únicos autores del trabajo titulado:

Características de la Web de la Universidad de las Ciencias Informáticas

y reconocen a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firman la presente a los ____ días del mes de _____ del año 2009.

_____	_____	_____
Firma del Autor	Firma del Autor	Firma del Tutor
Yonny Mondelo Hernández	Yuley Díaz Madruga	Dovier Antonio Ripoll Méndez

OPINIÓN DEL TUTOR DEL TRABAJO DE DIPLOMA.

Título: Características de la Web de la Universidad de las Ciencias Informáticas.

Autores: Yonny Mondelo Hernández y Yuley Díaz Madruga.

Tutor: Dovier Antonio Ripoll Méndez.

Por todo lo anteriormente expresado considero que los estudiantes están aptos para ejercer como Ingenieros Informáticos; y propongo que se le otorgue al Trabajo de Diploma la calificación de ___ puntos.

_____ ____ de _____ de 2009.

Firma

OPINIÓN DEL OPONENTE DEL TRABAJO DE DIPLOMA.

Título: Características de la Web de la Universidad de las Ciencias Informáticas.

Autores: Yonny Mondelo Hernández y Yuley Díaz Madruga.

Oponente: Liudmila Amat Reyes.

Por todo lo anteriormente expresado considero que los estudiantes están aptos para ejercer como Ingenieros Informáticos; y propongo que se le otorgue al Trabajo de Diploma la calificación de ___ puntos.

_____ ____ de _____ de 2009.

Firma

Agradecimientos

Agradezco antes que nada a mis padres, por haberme traído a este mundo. Por estar a mi lado desde siempre, por ser mis guías, y mostrarme los mejores caminos de la vida. Ustedes son mi principal razón de ser, por brindarme ese amor incondicional que solo los padres saben dar. Los llevo en el corazón. Agradezco a mi pequeña hermanita Yudith, que espero llegue muy lejos, sigue así querida. Agradezco a quienes considero más que amistades; pues son como hermanas que tuve de grandes. Aly, sabes que tú no puedes faltar; qué paso importante en mi vida he dado sin contar contigo? Siempre tendrás en mí, un hermano mayor. A ti Mayté, que siempre me has brindado tu apoyo, que siempre confiaste en mí. Que me has dado más dolores de cabeza que yo mismo, y siempre te ocupaste de ocupar hasta mis momentos libres. Por ser personas imprescindibles en mi vida, me es imposible pasarlas por alto. A mi novia Aniuskita, mi mayor tesoro. Por haberse convertido en todo lo que siempre he soñado, y ser todo lo que yo necesito. Tú eres mi universo; eres todo mi mundo. Nos queda toda una vida por delante amor, te lo aseguro. Yo sé cuánto deseabas este momento. A Vilma Reyes, My Friend del Alma, siempre has sido mi inspiración. Tú lo sabes mejor que nadie. A mi primo Danni, quien me ha servido de espejo en múltiples ocasiones. También te agradezco el enseñarme a no rendirme ante las dificultades, siendo ejemplo. Y a mi prima del alma, Yunia, que tantos momentos de alegría me ha dado, no me cansaré de agradecer nunca tu buena compañía. A Milady y Mailín, que ojalá lleguen muy lejos. A Enrique, uno de mis pocos primos informáticos; por compartir conmigo interesantes conversaciones y ser tan diferente de todos. A Danaisy, y Darien; que siempre me han acogido con alegría, y nunca dudaron de que lo lograría. A mi querida amiga Elsita, mis Padrinos Ángel y Tere, también Célida, por ayudarme siempre cuando lo necesité, por brindarme hoy sus invaluables amistades. Juntos vivimos momentos inolvidables. Y a todo el resto de mi familia, que de una forma u otra han estado conmigo siempre. Agradezco además, a mis amistades del barrio, Alexander(Alejo), Yoendri(El Negro), Onelio(Cocuyo), y demás. Que incluso entre momentos de diversión, han buscado el espacio para incentivarme a seguir adelante, y enorgullecerme de mi carrera. Sí que saben darle el justo valor a las cosas. A mi profe de la Secundaria Gabriel, por toda la matemática de la vida que he aprendido en su compañía. Se necesitan potentes pilares para sostener un edificio, creo que sabes bien como construirlos. A mi profesor Edgardo, por confiar siempre en mis capacidades y enseñarme lo importante que es ser perseverante. A mi Tutor, Dovier Antonio Ripoll Méndez, por estar ahí en todo momento. Por su ayuda incondicional, y sus consejos. Al Comandante Fidel Castro por crear esta oportunidad que hoy estamos viviendo, gracias desde lo más profundo. Siempre me consideraré un informático de Fidel. A mis amigos de manera general, y a los que me han ayudado y dado fuerzas a lo largo de la carrera.

Yonny Mondelo Hernández

Agradecimientos

Agradezco antes que todo a nuestro Comandante en Jefe Fidel Castro por haberme dado la oportunidad de estudiar en la Universidad del Futuro. A mis padres por haberme dado todo lo necesario para hacerme Ingeniero, siempre estaré orgullos de ellos. A mi tía Anastasia que ha sido durante toda la carrera y más, mi apoyo, guía y casi una madre para mí. Hoy soy mejor persona gracias a ella, gracias por todo tía. A mi hermana Yolaisy por ser mi ejemplo durante toda la vida, gracias a eso he llegado hasta aquí. A mi hermana yuderkys que es lo máximo, no querría una mejor hermana, al igual que Liudmila y Maryuri que me han ayudado mucho. A mis sobrinos Cristian, Ángel Adrián y Frank Alejandro, que se que llegaran muy lejos en la vida. No podría dejar de agradecer a mis primas, que mas bien son mis hermanas, Enisleidys, Marisleydis, Daily, yenisy , yenisbel,Damaris, Maide, Mayelin, Santica, Vanessa , a mis primos, Duny, Reinier, Enrique, Luis, Marco Antonio, a mi tia Maritza, Lety, Leony, en fin, toda la familia. Por supuesto, no podía dejar de agradecer a Adonis, que más que mi amigo lo considero mi hermano, siempre serás el hermano varón que nunca tuve, estoy muy orgulloso de ti. A mi amigo Leitniz, a quien considero mi primito más chiquito y sé que llegara muy lejos en la vida. A Rigoberto a quien considero un tremendo amigo y fue mi primer profesor de programación. A mi gran amigo Leodan, quien es hoy por hoy el mejor en ingles de la Universidad y es ejemplo a seguir. A mi vecina Vita, que siempre ha sido como mi abuela, para ella también es este título. También quisiera agradecerle a Paquita, por haberme ayudado cuando más lo necesite y a mis vecinas Zulaimy, Mary y Belky. A los socios del barrio yosvany, el kiki, el toti, yoandry, el Turco, Amaury, que ha sido mi consejero en muchas oportunidades., en fin, a toda la gente del barrio. No podría dejar de mencionar a Yulietsy quien me soportó durante 3 años en mi carrera. A mis amigos de la Universidad Adrian y koty(los habaneros del oriente), Tito, Ailet, Daney, Kirenia, Yolaine, yodelis, el Jose, en fin, todo el grupo 10502. No podría dejar de mencionar a Zuleidy, mi gran amiga, no cambies nunca. A mi tutor Pedro, que más que mi tutor de Ingles, es mi gran amigo. A mi amiga Yisel, por confiar siempre en mí. A Beatriz y Xiomara, por ser tan buenas amigas. A Yonny Mondelo Hernández, a quien más que mi compañero de tesis, lo considero mi amigo. A mi tutor Dovier, por ser ejemplo siempre y por estar ahí siempre cuando se le necesita. A todos los que de una forma u otra me ayudaron a obtener este logro, gracias.

Yuley Díaz Madruga

Dedicatoria

De manera especial dedico este Trabajo de Diploma a mi amiga del alma ("My Friend del Alma", como cariñosamente suelo llamarla), Vilmita Reyes. Quien aunque no puede estar hoy aquí, es como si lo estuviera para mi, pues este resultado es también en parte suyo.

Por ser la persona que ha estado siempre conmigo en los buenos y los malos momentos.

Por darme fuerzas constantemente, e incentivarme a salir adelante.

Por ayudarme a luchar para obtener buenos resultados y ser una mejor persona.

Por brindarme incondicionalmente su amor, su confianza y su fuerza.

Por ser una voz digna de escuchar, que me ha guiado en muchísimas más ocasiones, de las que puedo recordar. Y por darme más de un millón de buenas razones para recordar y convertirse en un torbellino de energía positiva a mi alrededor. Por todo eso, muchas gracias.

Lo dedico además, a mis padres; Manuel Mondelo Martínez y Zaida Hernández Feria. Por no haber dejado de confiar en mi ni un solo segundo. Por ser mucho más que importantes en mi vida; sin ustedes nada hubiera sido igual. Este logro, es para ustedes, más que para mi.

Yonny Mondelo Hernández

Dedico este trabajo a mi papa Ramón Díaz Martínez y Delia Madruga Trujillo por haberme dado siempre la seguridad de que llegaría muy lejos.

A mi Tía Anastasia Pérez Martínez que es como una madre para mí.

A mi hermana Yolaisy por ser mi guía, ejemplo y el espejo donde siempre quiero mirarme.

A mis hermanas Yude, Liudmi y Mayi por ser las mejores hermanas del mundo.

Y a mi gran amigo Adonis por estar ahí siempre que lo necesito, para el también es este logro.

Yuley Díaz Madruga

Resumen

En el presente trabajo se muestran los principales resultados de los tres Estudios Webmétricos realizados en la Universidad de las Ciencias Informáticas (UCI), que permitirán analizar datos exhaustivos obtenidos de los sitios web de dicho centro. Se ofrecen además varios datos comparativos entre cada uno de los estudios, lo cual constituye la base fundamental para medir la evolución de la Web durante los últimos meses, tomando como punto de partida los estudios webmétricos realizados con anterioridad.

El último estudio presentado fue realizado entre el 1ro de abril de 2009 y el 8 de abril de 2009. Al igual que los dos anteriores, fue desarrollado por el Proyecto Generador de Estudios Webmétricos (GEWEB) del Grupo de Proyectos de Cibermetría Aplicada (CIBA), perteneciente al Polo Productivo de Soluciones Informáticas para Internet (SINI). Se utilizó en todos los casos como *Spider*, el Sistema *WIRE*, desarrollado en el Centro de Investigación de la Web (CIW) de Chile, y una lista de partida de 240 sitios web oficialmente acogidos en la UCI en el último recorrido.

Se pueden mencionar, por ejemplo, los siguientes resultados obtenidos:

- La Web de la UCI está compuesta por aproximadamente 160 sitios web, con más de 700000 páginas web, aunque sólo se han logrado descargar cerca de 462000.
- El 86 % de las referencias a dominios externos al dominio de la UCI, son al .CU, con más de 7700000 enlaces encontrados.
- Más del 90% de las páginas web analizadas se encuentran en idioma español.
- El sitio web con mayor contenido es el Portal del Proyecto ERP-CUBA (Planificación de Recursos Empresariales de CUBA), con más de 120000 páginas web.
- La Intranet de la UCI, presenta el mayor Grado Interno y Externo, de todos los sitios web.
- Más de un 97% de las páginas web analizadas, están entre las profundidades 1 y 10.
- Más de un 93% de las páginas web analizadas, fueron creadas o modificadas en el último año.
- En la Web de la UCI predominan las extensiones .PHP, .GIF, .PNG, .RPM, .CSS y .JPG.
- Predomina el Servidor de Aplicaciones Web Apache (en distintas variantes), y el Sistema Operativo GNU/Linux, del cual se encontraron cuatro distribuciones distintas.

Un estudio webmétrico es como una fotografía tomada a la Web, algo similar a lo que hace un astrónomo cuando observa las estrellas en el Universo [6]: lo que ve es la luz que viaja desde las estrellas, que en ese momento ya pudieron dejar de existir. De ese mismo modo, pueden haber dejado de existir muchas de las páginas analizadas en estos estudios. Sin embargo, al medir los resultados en conjunto con otros estudios, se pueden establecer pronósticos y tendencias que ayuden a tener un control de las tecnologías que actualmente son utilizadas y establecer líneas de trabajo en función de mejorar el uso de las mismas y acelerar este proceso.

Palabras Clave: Estudio webmétrico, Web de la UCI, herramienta *WIRE*, sitios web, páginas web.

Abstract

This work is about the main results of three webmetric studies conducted at the University of Informatics Sciences (UCI by its acronym in Spanish), which will analyze data obtained from comprehensive websites that center. It also offers several comparative data from several individual studies, which is the fundamental basis for measuring the evolution of the Web in recent months, taking as a starting webmetric studies conducted previously.

The last study was conducted between April 1, 2009 and April 8, 2009. Like the previous two, was developed by the Project Generator of Webmetric Studies (GEWEB by its acronym in Spanish), Group Projects of Aplicated Cybermetrics (CIBA by its acronym in Spanish) belonging to the Productive Area of Informatics Solutions for the Internet (SINI by its acronym in Spanish). Was used in all cases as Spider, the WIRE System, developed at the Center for Web Research (CIW) from Chile, and a 240-item list of sites officially hosted in the University of Informatics Sciences in the last travel. We can mention, for example, the following results:

- The Web of the University of Informatics Sciences is comprised of approximately 160 Websites, with more than 700000 web pages, but have managed to download only about 462000.
- The 86 percent of the references to domains outside the domain of the UCI, are at .CU, with more than 7700000 links found.
- Over 90 percent of the web pages analyzed were found in Spanish language.
- The website with more content is the Portal of the Project ERP-CUBA (Enterprise Resource Planning of CUBA), with more than 120000 web pages.
- The Intranet Site of the UCI, have the highest internal and external grade, of all sites analyzed.
- On more than 97 percent of the web pages analyzed, the depths are between 1 and 10.
- More than 93 percent of the web pages analyzed, were created or modified in the last year.
- In the Web of the University of Informatics Sciences were found more the extensions .PHP, .GIF, .PNG, .RPM, .CSS and .JPG.
- Dominates the Apache Web Server Application (in different variants), and the operating system GNU/Linux, which were found four different distributions.

A webmetric study is like a photograph taken on the Web, something similar to what an astronomer when he observes the stars in the universe [6]: what you see is the light that travels from the star at that time and could leave to exist. In the same way may have left to exist many of the web pages analyzed in these studies. However, when measuring the results in conjunction with other studies, we can establish forecasts and trends to help take control of the technologies currently used and to establish lines of work in terms of improving the use, and speed up this process.

Keywords: Webmetric Study, UCI's Web, WIRE tool, websites, web pages.

Índice General

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

INTRODUCCIÓN

La Web es una colección de billones de documentos escritos de tal forma que pueden ser citados usando hiperenlaces, conformando el denominado hipertexto. Tales documentos, o páginas web, tienen múltiples caracteres escritos en gran cantidad de idiomas y cubren esencialmente todas las materias del saber humano [1].

El desarrollo acelerado de los procesos sociales en la actualidad, implica no sólo un cambio de paradigmas económicos, políticos e ideológicos, también presupone un creciente desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC). En el desarrollo tecnológico global, la Web constituye premisa, condición y consecuencia de este inevitable progreso de la sociedad.

La Web se presenta como un espacio público utilizado por millones de usuarios con propósitos diferentes. Aún cuando su objetivo esencial es propiciar el intercambio de información, en la actualidad ha devenido en un importante medio que provee servicios para múltiples usos como publicidad, entretenimiento, educación, comercio electrónico, entre otros. Por encontrarse en constante crecimiento y ser tan dinámica, el estudio de sus características brinda interesante información para entender su estructura; por lo que ha sido objeto de constantes estudios.

La Web es uno de los principales medios de difusión de la información en estos momentos. En la actualidad sus páginas son contenedoras de programas, archivos comprimidos, documentos de texto, archivos multimediales como pueden ser imágenes, videos, animaciones, sonidos, entre otros. Son construidas utilizando diferentes lenguajes de programación y son sustentadas por diversos tipos de servidores web. No sería un error afirmar que la Web es un reflejo de las tendencias que son seguidas por la mayoría de los usuarios en cuanto a la utilización de las tecnologías de la Informática.

A nivel mundial se han realizado diferentes caracterizaciones, a través de indicadores cibermétricos que ayudan a conocer el comportamiento de la Web; específicamente, por citar algunos ejemplos, en las siguientes regiones: Argentina, Austria, África, Chile, Brasil, Corea, Cuba, España, Hungría, Perú y Portugal. Uno de los países con más experiencia en el estudio de las características de su Web es Chile [2], que ha sido objeto de estudio durante los años 2000 [3], 2002 [4], 2004 [5] y, más recientemente, 2006 [6]. Por su parte, en Cuba se han realizado acercamientos teóricos que constituyen la base para los estudios del comportamiento de su Web, fundamentalmente el confeccionado por la Empresa de Tecnologías de la Información y Servicios Telemáticos Avanzados (CITMATEL)[1]; dicho estudio tiene como título “Estudio de las Estadísticas Web de accesos y visitas del Portal Cuba.cu” [7] el cual, a pesar de ser el primero de su tipo, aportó significativos datos.

La Universidad de las Ciencias Informáticas (UCI)[2] no dispone de suficiente información sobre las características de su Web. La ausencia de una caracterización actualizada implica también la carencia de estudios de indicadores que permitan medirla en cifras para la obtención de estadísticas rigurosas, pertinentes y actualizadas sobre el impacto de la red en distintas actividades de interés científico-tecnológico, económico y social y el análisis de los patrones de comunicación científica a través de la Web y su trascendencia para el proceso de informatización de la sociedad cubana.

Mediante una caracterización se pueden construir indicadores, que permiten el estudio de aspectos tales como la evolución del tamaño de la Web, y la forma en que la misma pudiera hacer algún tipo de cambio aunque no necesariamente en el tamaño sino en ámbitos como la estructura y el contenido. Se pueden realizar labores de mantenimiento de la Web, tales como detección y corrección de *links* o vínculos rotos. Se puede realizar clasificaciones de tipo de documentos, distribución de sitios por países, o por dominios, incluso se pueden realizar distribuciones de

[1] http://www.citmatel.cu

[2] http://www.uci.cu

páginas por el tipo de idioma utilizado en las mismas, o la manera de mostrar la información, o sencillamente por la edad de cada una de ellas. También es muy común la realización de estudios webmétricos con el objetivo de analizar algún contenido específico en la Web, su nivel de difusión o impacto, o ambos.

Se puede inferir como **problema científico** de la investigación: ¿Cómo realizar un análisis cualitativo y cuantitativo de la Web de la Universidad de las Ciencias Informáticas para fortalecer la información acerca del estado real de determinados indicadores webmétricos, como punto de referencia de la toma de decisiones?

De este **problema científico** se derivan las siguientes **preguntas de investigación:**

¿Carece la Universidad de las Ciencias Informáticas de información sobre el estado real de algunos indicadores webmétricos fundamentales de su Web?

¿Qué indicadores webmétricos son comúnmente usados para caracterizar una Web?

¿Influye de manera positiva en la toma de decisiones la realización de un estudio cualitativo y cuantitativo de estos indicadores?

¿Qué potencialidades brinda el *WIRE (Web Information Retrieval Environment)* [8] para la recolección de datos de la Web, cómo se utiliza la aplicación?

El **objeto de estudio** son los estudios webmétricos.

El **campo de acción** está dado por el proceso de control y estudio de la información de la Web.

La **idea a defender**, que se establece como base de la investigación efectuada, se puede formular de la siguiente manera: Es posible realizar un estudio cualitativo y cuantitativo de la Web de la Universidad de las Ciencias Informáticas, para contribuir a la toma de decisiones sobre la misma.

A partir de lo anterior, se persigue como **objetivo general:**

Caracterizar integralmente la Web de la Universidad de las Ciencias Informáticas, para contribuir a la toma de decisiones sobre la Web analizada.

Para dar solución al problema se definen los siguientes **objetivos específicos**:

1. Analizar caracterizaciones realizadas de la Web, para definir los indicadores webmétricos de interés para la caracterización de la Web de la UCI.
2. Utilizar el *WIRE* para recuperar la información de la Web de la UCI durante un periodo de tiempo determinado.
3. Valorar las potencialidades y funcionalidades del WIRE.
4. Analizar cualitativa y cuantitativamente la información recuperada sobre la Web de la UCI.

Para dar solución a los objetivos de la investigación se plantean las siguientes **tareas de investigación:**

- Revisar caracterizaciones de la Web realizadas en otros países y Universidades Tecnológicas, y analizar los indicadores cibermétricos utilizados en las mismas.
- Definir los indicadores de interés para la caracterización de la Web de la UCI.
- Estudiar el modo de funcionamiento del *WIRE*, así como su configuración.
- Realizar una caracterización integral de la Web de la UCI.

Una vez concluido el trabajo se contará con los siguientes **resultados concretos:**

- Artículo científico sobre la caracterización cuantitativa y cualitativa de la Web de la Universidad de las Ciencias Informáticas.

Un **método de investigación** provee estrategias elementales para ahorrar esfuerzo y tiempo en una investigación científica. Para guiar las tareas de investigación se utilizan principalmente los métodos de investigación siguientes:

Métodos Teóricos:

- Análisis - Síntesis.

Monitoreando durante un tiempo determinado la Web de UCI podremos descomponer en sus características el objeto de estudio, para lograr una valoración cuantitativa y cualitativa que nos permita sintetizar en el pensamiento sus propiedades como fenómeno más multifactorial e integral.

- Inducción - Deducción

Analizando de manera particular las caracterizaciones realizadas en diferentes países podremos llegar a la generalización de indicadores cibermétricos de interés y su aplicación en condiciones singulares de la Web de la UCI.

- Lógico - Histórico

La caracterización cuantitativa y cualitativa de la Web de la UCI nos aportará la lógica de su desarrollo y el avance real a través de diferentes etapas.

El contenido de la investigación se estructurará de la siguiente manera:

Primer Capítulo: Fundamentación Teórica del tema, estado del arte acerca de los estudios webmétricos. Valoración de las caracterizaciones realizadas en otros países.

Segundo Capítulo: Definición de los principales indicadores cibermétricos para caracterizar la Web en la UCI, así como la herramienta que permitirá realizar el estudio.

Tercer Capítulo: Se realizará el estudio del estado real de la Web de la UCI, mediante un análisis cuantitativo y cualitativo de los resultados obtenidos

Finalmente se presentan las **Conclusiones**, **Recomendaciones**, **Referencias Bibliográficas** y el **Glosario de Términos**.

CAPÍTULO I - Fundamentación Teórica.

En este capítulo, se expone una visión general del estado del arte; tanto a nivel nacional, como internacional. Además, se profundiza de manera objetiva en temas, tales como la Webmetría, los *Spiders*, Internet, etc. De esta manera se pretende lograr un mayor entendimiento de la investigación realizada.

1.1 La Web. Su composición.

La Web es el universo de información accesible a través de Internet y su componente más usado. Es una fuente inagotable del conocimiento humano. La misma se puede conceptualizar de la siguiente manera:

La Web es un conjunto de documentos entrelazados en un sistema de hipertexto. El usuario entra en la Web a través de una página de inicio [9].

La misma forma el conjunto total de documentos de hipertexto con enlaces entre ellos que residen en servidores HTTP (Protocolo de transferencia de hipertexto) alrededor de todo el mundo. Los documentos en el World Wide Web (www), llamados páginas o páginas web, se escriben en HTML (lenguaje de marcas hipertextuales), identificados por una URL (localizador uniforme de recursos) que especifican la máquina y camino particulares por los que se pueden acceder a un archivo, y transmitirse de nodo a nodo al usuario final bajo HTTP. Los códigos, llamados etiquetas, incrustados en un documento HTML relacionan las palabras asociadas e imágenes particulares en el documento con URL para que un usuario pueda acceder a otro archivo, que puede estar en cualquier parte del mundo, en el momento de presionar una tecla o hacer clic a un ratón. Estos archivos pueden contener textos (con variedad de fuentes y estilos), imágenes de gráficos, archivos de películas y sonido así como applets o subprogramas de Java, controles ActiveX u otros pequeños programas de software incrustados que se ejecutan cuando el usuario los activa pulsando sobre un enlace. Un usuario que visite una página web puede también descargar archivos de un sitio FTP y enviar mensajes a otros usuarios por vía e-mail, utilizando enlaces en las páginas web [9].

Usando la Web, se tiene acceso a millones de páginas de información agrupadas en sitios. La apariencia de un sitio web puede variar ligeramente dependiendo del explorador que se use. Así mismo, las versiones más recientes disponen de una funcionalidad mucho mayor tal como animación, realidad virtual, sonido y música. Estos documentos, o páginas web, tienen unos pocos cientos de caracteres escritos en múltiples idiomas y que cubren esencialmente todas las materias del saber humano. Estas páginas web se encuentran instaladas en un servidor web y son servidas ante las peticiones del cliente empleando el protocolo *HTTP* y presentadas al usuario por los visores web [1].

El hipervínculo presenta muchas ventajas, tanto para los usuarios, a la hora de buscar información, como para los programas que recorren la Web, a la hora de buscar contenido para recolectar (probablemente para un motor de búsqueda). Debido a esto se plantea que la Web sigue un modelo de grafo dirigido, en el que cada página es un nodo y cada arco representa un enlace entre dos páginas. En general las páginas enlazan a páginas similares, de modo que es posible reconocer páginas mejores que las demás, es decir, páginas que reciben un número mayor de referencias que lo normal [6].

Composición de La Web.

Hoy en día, "*cuando se navega en Internet"* es impresionante constatar la existencia de gran cantidad de sitios web, que a su vez están compuestos por páginas web, que han llegado para cambiar la forma de ver y hacer las cosas, para mostrar que se puede ir más allá del aspecto informativo.

Los sitios web**:** son un conjunto de archivos electrónicos y páginas web referentes a un tema en particular, que incluye una página inicial de bienvenida, generalmente denominada página de inicio o *home page*. Que presentan un nombre de dominio y una dirección en Internet especifica, empleada por instituciones, organizaciones e individuos para comunicarse en el mundo.

Un ejemplo de la forma de utilización es en el caso particular de las empresas, que a través de este servicio presentan mensajes que tienen que ver con ofertas de sus bienes y servicios a través de Internet, y en general para dar eficiencia a sus funciones de mercadotecnia.

Los sitios web pueden ser de diversos géneros, destacando los sitios de negocios, servicios, comercio electrónico en línea, imagen corporativa, entretenimiento e informativos.

Las páginas web son en su mayoría un documento electrónico que contiene información específica de un tema en particular y que es almacenado en algún sistema de cómputo que se encuentre conectado a la red mundial de información denominada Internet, de tal forma que este documento pueda ser consultado por cualquier persona que se conecte a esta red mundial de comunicaciones y que cuente con los permisos apropiados para hacerlo. Una página web es la unidad básica del *WWW*.

Los sitios web, así como sus páginas, tienen la característica peculiar de que se pueden combinar para hacer que la Web pase del estado estático al dinámico, así como al colaborativo.

1.1.1 Tipos de Web.

La evolución de la Web comienza con el surgimiento de la Web 1.0 (Web estática), luego surge la Web 1.5 (Web dinámica), y por último la Web 2.0 (Web colaborativa). Todo se debe a los avances tecnológicos y al empuje de muchas compañías que se aprovechan de todas las potencialidades que presenta la Web, e implementan un valor agregado a sus clientes.

La Web Estática: *Son aquellos sitios enfocados principalmente a mostrar una información permanente, donde el navegante se limita a obtener dicha información, sin que pueda interactuar con la página web visitada, las Web estáticas están construidas principalmente con hipervínculos o enlaces (links) entre las páginas web que conforman el sitio, este tipo de Web son incapaces de soportar aplicaciones web como gestores de bases de datos, foros, consultas online, e-mails inteligente [9].*

Esta es una opción más que suficiente para aquellos sitios web que simplemente ofrecen una descripción de su empresa, quiénes somos, dónde estamos, servicios. Ideal para empresas que no quieren muchas pretensiones con su sitio web, simplemente informar a sus clientes de sus productos y su perfil de empresa.

La Web Dinámica: *Son aquellos sitios que permiten crear aplicaciones dentro de la propia Web, otorgando una mayor interactividad con el navegante. Aplicaciones dinámicas como encuestas y votaciones, foros de soporte, libros de visita, envío de e-mails inteligentes, reserva de productos, pedidos on-line, atención al cliente personalizada [9].*

El desarrollo de este tipo de Web es más complicado, pues requiere conocimientos específicos de lenguajes de programación así como creación y gestión de bases de datos, pero la enorme potencia y servicios que otorgan este tipo de páginas hace que merezca la pena la inversión y esfuerzo realizados respecto a los resultados obtenidos.

La Web Colaborativa: *Es la transición que se ha dado de aplicaciones tradicionales hacia aplicaciones que funcionan a través de la Web, enfocándola al usuario final. Se trata de aplicaciones que generen colaboración y de servicios que reemplacen las aplicaciones de*

escritorio. Es una etapa que ha definido nuevos proyectos en Internet y está preocupándose por brindar mejores soluciones para el usuario final. Muchos aseguran que hemos reinventado lo que era el Internet, otros hablan de burbujas e inversiones, pero la realidad es que la evolución natural del medio realmente ha propuesto cosas más interesantes [9].

1.2 Cibermetría y Webmetría. Surgimiento y objetivo.

Dentro de las características que han hecho de la Web el mayor repositorio de información de la humanidad se encuentran el fácil acceso, bajo costo y la libertad de publicación.

Gran parte de la información está disponible a través de simples mecanismos de interacción para y entre las personas, las que producen datos en un formato adecuado para el entendimiento de cada una de ellas, sin embargo a esta información la mayoría de las veces no es fácil acceder y resulta un poco engorrosa su interpretación mediante un procesamiento automático, debido a que pueden existir sitios que no estén referenciados por otros sitios, que simplemente se pueda llegar a ellos mediante su dirección de *URL*, por lo que no son indexados ya que no existe un camino para llegar a ellos, además no se debe descartar la posibilidad de la ocurrencia de algún error en el lado del servidor.

La abundancia de contenidos en la Web ha llevado a que prevalezca la falsa creencia de que los recursos disponibles en la Red son ya accesibles por el mundo y así cubre no solo todas las áreas del conocimiento, sino también que reflejan la mayoría de las posiciones e idiosincrasias que la diversidad mundial ofrece. Lo que se encuentra muy lejos de la realidad, por la diferencia de desarrollo que hay en cuanto a las tecnologías en los diferentes países, tanto desarrollados como subdesarrollados.

Aunque el tamaño de la Web es relativamente grande, superando los 10000 millones de páginas, existe mucha información que no se ha llegado a representar aún. Se está viviendo en una época de explosión informática donde la información se mide en exabytes, se estima que las nuevas informaciones almacenadas crecieron alrededor del 30% al año entre 1999 y 2002. Los flujos de información a través de canales electrónicos - teléfono, radio, TV, y la Internet - representaron casi 18 exabytes de información en 2002, el *WWW (World Wide Web)* contiene alrededor de 170 terabytes de información sobre su superficie, haciéndose mayormente visible y palpable en idiomas que no sean el inglés, aunque no se puede decir que está completo, todo lo contrario aún en este existen importantes lagunas que “aunque el ritmo del crecimiento de la Web es explosivo” tardará en rellenar.

Todo lo anterior condicionó la necesidad de una ciencia que estudiara detalladamente el comportamiento cuantitativo y cualitativo de la Web y del ciberespacio en general, lo que propició el origen a la Cibermetría. La misma, con todas sus variantes terminológicas, estudia la aplicación de las técnicas informétricas a cualquier tipo de información disponible en la Red Internet. A su vez, formando parte de esta, la Webmetría se basa en la aplicación de la Informetría, y otras técnicas nuevas de medida, específicamente a la información disponible a través de la *WWW*, estudiando en profundidad a la Web [12].

El origen de la Cibermetría puede situarse a mediados de los noventa, en sus inicios fueron propuestos varios términos para designar la nueva disciplina, aunque finalmente se adoptaron dos que, sin llegar a serlo, en momentos se emplean como sinónimos *Cybermetrics* y *Webometrics*. Para su traducción en español ambos fueron adaptados literalmente del inglés, dando así lugar a las expresiones Cibermetría y Webmetría [10].

Entre los principales objetivos de la Cibermetría y la Webmetría se encuentran la construcción de indicadores, los que permiten el estudio de aspectos tales como la evolución del tamaño de la Web, y la forma en que la misma pudiera hacer algún tipo de cambio aunque no necesariamente en el tamaño sino en ámbitos como la estructura y el contenido.

Cibermetría: *Es la disciplina dedicada a la descripción cuantitativa de los contenidos y procesos de comunicación que se producen en el ciberespacio [11]. Es el estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías en Internet, desde perspectivas bibliométricas e informétricas [12].*

Webmetría: *Es el estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías de una parte concreta de Internet, por regla general a una Web o portal* [12].

1.2.1 Aplicaciones de la Cibermetría y la Webmetría.

Internet, como enorme autopista de la información, ha proporcionado argumentos para que se le realice un estudio profundo a la Web, tomando como punto de partida que es su componente más usado. Desde un punto de vista webmétrico se considera que las técnicas de medición pueden aplicarse fundamentalmente a las siguientes categorías del *WWW:*

- El número de sedes web y de páginas de inicio en el mundo y también su distribución por países.
- Clasificación de las páginas web por tipos de documentos.
- Número de sitios web por dominios.
- Clasificación de páginas web por el idioma de los documentos y por los modos de representación de la información.
- Estadísticas de uso y usuarios de las páginas web en un periodo de tiempo determinado.
- El número de citas recibidas por cada página web.
- Ordenar los sitios web más visitadas y páginas personales según el tipo de documento.
- Los tipos de colecciones electrónicas disponibles en cada sede web.
- Factor de impacto de la Web y productividad de los autores.
- Análisis del contenido de las páginas web.
- Identificar la variedad de publicaciones electrónicas por el tipo, el idioma y la distribución geográfica.

La Cibermetría es una ciencia relativamente reciente con un carácter multidisciplinario, lo que se puede observar por medio del análisis de las múltiples aplicaciones que presenta. Los estudios de visibilidad de la Web, de su densidad, los análisis de citas y la investigación sobre el diámetro de la misma, se basan en la utilización de los indicadores cibermétricos.

1.3 Indicadores cibermétricos.

Uno de los frentes abiertos en el campo de la Cibermetría es el estudio de aspectos importantes de la Web, a través de indicadores cibermétricos por ser *una medida de relevancia en la toma de decisiones, que cuantifica aspectos de creación, difusión y aplicación de la ciencia y la tecnología en la medida en que están representadas en Internet o el World Wide Web [10]*.

Los indicadores cibermétricos ya han sido incorporados a los estudios de descripción y evaluación de la actividad científica, como por ejemplo en las caracterizaciones de diversos entornos, como países, o sencillamente instituciones.

De forma general, los indicadores cibermétricos pueden agruparse en tres grandes grupos o tipos de medida:

- Medidas descriptivas.
- Medidas de visibilidad e impacto.
- Medidas de popularidad.

Lo que ayuda a que los mismos estén organizados a la hora de ser escogidos para la medición de un espacio en específico.

Medidas descriptivas: Como su nombre indica, miden fundamentalmente el tamaño o número de objetos encontrados en cada sede, la riqueza de las páginas, ficheros medianos o ricos en contenido, densidad de enlaces totales y únicos. Se utilizan para medir la penetración de Internet desde el punto de vista de los contenidos que por países, regiones, organizaciones o grupos de individuos pueden presentar.

Dentro de esta medida se encuentran los indicadores descriptivos o de recuento. En la Webmetría, los indicadores descriptivos, además del conteo de los artículos producidos y publicados por la comunidad científica en el entorno electrónico de Internet, también incluye el recuento de diferentes aspectos de los recursos en la red.

Entre los que se encuentran:

- El tamaño medio de los documentos analizados.
- Los protocolos utilizados por las direcciones *URL* de los documentos *HTML* analizados.
- Los tipos de ficheros.
- La tipología documental de las páginas Web.
- Los recursos: página Web con datos textuales o audiovisuales.
- El número medio de enlaces por página.
- El tamaño documental.
- El tamaño informático.
- La densidad hipertextual.
- La densidad multimedia.
- La profundidad.

A éstos, pueden añadirse otros elementos como: el número de sitios según su temática e idioma, el número de páginas por sitios, la distribución de recursos electrónicos por tipo, país e institución, así como la productividad científica en el entorno electrónico.

Estos últimos elementos apuntarían sobre todo, a la medición de la comunicación científica en el Web y como puede observarse constituyen sólo adaptaciones al entorno digital, porque se utilizan también en los estudios métricos tradicionales.

Medidas de visibilidad e impacto: Se basan en el carácter hipertextual de la Web y exploran los patrones de enlace entre páginas y sedes de distintas procedencias. El número y diversidad de enlaces externos recibidos, su volumen respecto a los contenidos y objetos de enlace (llamado Factor de Impacto Web) o índices que se construyen de acuerdo con el peso relativo de las sedes

de origen de los enlaces, como el *PageRank* de Google, aunque en otras instituciones se puede visualizar como *Ranking*. Este indicador ayuda a establecer valores de utilidad y calidad en cuanto al contenido o estructura, del sitio o de la página web, al haber creado una buena impresión al visitante. Esta medida permite además establecer listas ordenadas, según la jerarquía numérica de estos indicadores. Entre estos indicadores se encuentra el Factor de Impacto (FI). El Factor de Impacto Web (FIW) es uno de los primeros indicadores examinados en los trabajos webmétricos.

Existen una serie de problemas relacionados con el FIW que se centran fundamentalmente en los métodos de recopilación de citas y páginas web y en la propia naturaleza de ambos elementos. Con respecto al método de recopilación, no todos los motores de búsqueda ofrecen iguales posibilidades para realizar un estudio webmétrico, y particularmente para calcular el FIW. Por lo que en muchos estudios de los ya realizados omiten este indicador.

Medidas de popularidad: El consumo de información medido en términos de número y características de las visitas desde la Web resulta un atractivo, aunque extremadamente complejo de implementar, es un método de evaluación, que ayuda a saber la popularidad que puede presentar la información o la forma de representarla. Es ciertamente interesante para estudios temporales, en los que la medida de la evolución resulta prioritaria para los correspondientes informes. Como se indica es complicado obtener valores absolutos, pero ciertos valores relativos con valores importantes pueden, no obstante, utilizarse en análisis comparativos.

Las dos primeras medidas se corresponden con los indicadores métricos tradicionales, entre los que se distinguen dos tipos fundamentales: indicadores de actividad e indicadores de impacto, dentro de los de actividad se encuentran el número y distribución de publicaciones, productividad de autores y formando parte de los de impacto el número de citas recibidas y el factor de impacto.

Ventajas de los indicadores cibermétricos.

Los indicadores son más utilizados para realizar estudios cibermétricos proporcionan una serie de beneficios, como es el caso de las siguientes ventajas:

1. Mayor potencia crítica para la realización de análisis de patrones globales y sectoriales, dentro de las que están:
    - Mejor tratamiento con técnicas estadísticas.
    - Nuevas unidades: Mayor finura en el análisis.
    - Perspectiva cuantitativa y objetiva.
2. Mejores resultados esperados:
    - Presentando una batería de indicadores más amplia.
    - Las medidas combinadas.
    - Visualización más espectacular.
    - Seguimiento individualizado.
    - Medidas directas e indirectas.
    - Comparación con descriptores “*offline*”.
3. Ventajas políticas:
    - Medida de la producción del conocimiento.
    - Incremento de los contenidos.

1.4 Herramientas cibermétricas.

En la actualidad la Web es objeto de interesantes análisis cibermétricos con ayuda de métodos indirectos, basados en las potencialidades de los motores de búsqueda, los que a su vez utilizan para la indexación de páginas a los *robots* de búsqueda. Los robots eran utilizados originalmente para estudiar el posicionamiento de sedes web, estos métodos pueden ser útiles para la evaluación comparativa de muestras homogéneas.

Unas de las herramientas en la que se basa la Cibermetría son los propios *robots* de búsqueda para la extracción de datos cuantitativos de las sedes web previamente identificadas, permitiendo obtener datos complementarios para la realización de estudios de este tipo. Los *robots* de búsqueda forman parte de los motores de búsqueda. Para la tarea de realizar recolección de datos para un estudio webmétrico es recomendable utilizar los robots de búsqueda.

Antiguamente cuando las páginas web sólo se contaban por miles, los robots de búsqueda no se necesitaban, puesto que la mayoría de los directorios de la Web incluían un botón o módulo, para añadir una nueva página web. Hoy en día, las *URL* de páginas nuevas ya no son un recurso escaso, puesto que hay miles de millones de páginas web.

El principal problema de los robots de búsqueda es que tienen que hacer frente al tamaño de la Web y al tipo de cambio existente en la misma; sin la indexación de los *robots* de búsqueda, más de un tercio de la Web a disposición del público estaría perdida. Pues a medida que el número de páginas crece, será cada vez más importante centrarse en lo indispensable que son las páginas, y la forma en que un *robot* de búsqueda sea capaz de indexar la Web, puesto que existe un incremento muy acelerado de la misma, lo que provoca que tienda al infinito.

1.5 Robots de búsqueda. Principio de su funcionamiento.

Un robot es un programa que atraviesa una estructura de hipertexto recuperando ese enlace y el resto de los que están referenciados allí. Los robots son usualmente llamados "*Web Wanderers*", "*Web Crawlers*" o "*Spiders*" (arañas de búsqueda) y se suele imaginar que se mueven entre los sitios como si fuesen virus, lo que no es el caso, un robot simplemente visita los sitios y extrae los enlaces que están incluidos dentro de éstos [13].

Son programas que inspeccionan las páginas del *WWW* de forma metódica y automatizada, pueden llegar a ser reiterativos, si los que lo realizaron crearon el mecanismo para que fueran recursivos. Los *robots* se utilizan para crear una copia de todas las páginas web visitadas para su procesado posterior por un motor de búsqueda que indexa las páginas, proporcionando un sistema de búsquedas rápido.

El típico diseño de los *robots* de búsqueda es una cascada que se retroalimenta de ella misma, en la que un rastreador web crea una colección que es indexada y buscada. La mayoría de los diseños de los *robots* de búsqueda cumplen la función de examinar, para eso realiza una primera etapa de búsqueda en la Web, con poca retroalimentación de los algoritmos de clasificación para el proceso de rastreo. Se trata de una cascada modelo, en el que las operaciones se llevan a cabo en estricto orden: en primer lugar se rastrea, luego se realiza la indexación, y posteriormente la búsqueda, aunque en algunos casos realizan además la creación de estadísticas y reportes.

1.5.1 Principio de funcionamiento.

Los *robots* son programas que simulan el funcionamiento de nuestros Navegadores ("*Explorer*" o "*Netscape*", u otros...), estos programas comúnmente denominados "*Robots*" o "*Web-Crawlers*" pueden estar hechos en varios lenguajes (Perl, C, entre otros...) pero su funcionamiento básico depende del protocolo *HTTP* [14].

Los *robots* comienzan visitando una lista de direcciones *URL*, donde identifica los hiperenlaces de dichas páginas y los añade a la lista de direcciones *URL* a visitar de manera recurrente de acuerdo a determinado conjunto de reglas. La operación normal es que se le proporciona al programa un grupo de direcciones iniciales, descarga estas direcciones, analiza las páginas y busca enlaces a páginas nuevas. Luego descarga estas páginas nuevas, analiza sus enlaces, y así sucesivamente [15].

Entre las tareas más comunes de los Web *Crawlers* tenemos:

- Crear el índice de una máquina de búsqueda.
- Analizar los enlaces de un sitio para buscar enlaces rotos.
- Recolectar información de un cierto tipo, como precios de productos para recopilar un catálogo.

1.5.2 Cómo decide un robot qué visitar.

Cómo decidir qué visitar depende del robot. Cada uno usa diferentes estrategias. En general comienzan a trabajar desde una lista histórica de direcciones *URL*. Especialmente con documentos con muchos links, tales como una lista de servidores "*what's New*"(qué hay de nuevo) y desde los sitios más populares en la Web.

Muchos indexan servicios que le permiten dar de alta a uno o más sitios manualmente, los cuales harán cola para ser indexados por el robot. Son usados a veces otros recursos también como listas de correo, grupos de discusión, entre otros. Esto les proporciona un punto de partida para comenzar a seleccionar las direcciones *URL* que ha de visitar, analizarlas y usarlas como recurso para incluirlas dentro de su base de datos.

1.5.3 Cómo decide un robot qué Indexar.

Depende del robot lo que se va a indexar, pero generalmente usan los títulos de HTML o los primeros párrafos, o selecciona la página HTML completa e indexa las palabras contenidas, excluyendo las de uso común (pronombres, adverbios y palabras como "Web", "página") dependiendo de las construcciones de las propias páginas HTML.

Algunos *robots* seleccionan las etiquetas o tags, u otros tipos especiales de etiquetas ocultas. Una práctica muy común es indexar también los textos alternativos de los gráficos. En el tiempo en que el robot este haciendo su trabajo es necesario que se le preste especial atención pues, en caso de indexarse, son palabras que contarán con un gran peso sobre la relevancia final en el documento.

1.6 Estudios internacionales y nacionales de la Web.

La caracterización de espacios web es una tarea compleja a escala global, a la que se le ha dedicado tiempo por parte de países tanto desarrollados como subdesarrollados. Los países que han realizado estudios de la Web, previamente realizan un análisis de los indicadores cibermétricos para poder seleccionar los más convenientes para su investigación, así también de las herramientas a utilizar. Entre las regiones que han desarrollados estas investigaciones se encuentran: Argentina, Austria, África, Chile, Corea, Cuba, España, Hungría, Perú y Portugal.

- **Argentina.**

El estudio de la Web de Argentina [16] se desarrolló durante los meses de marzo y abril del 2006, obteniendo que por cada página descargada se almacenaran como máximo 100 KB, utilizando el *crawler WIRE* para la recolección de las páginas.

Para el estudio del contenido de su Web se dividió en diferentes niveles teniendo como el primer nivel:

Contenido: tamaño de la página, términos más utilizados, términos en nombres de sitios y páginas por sitios.

Enlaces: grado entrante y saliente de páginas, *PageRank* (*Ranking* de las páginas), grado entrante y saliente del *Hostgraph*, componentes fuertemente conectados (SCC, en inglés Strong Connected Components) y la estructura microscópica.

Tecnologías: códigos de respuestas *HTTP*, longitud de las *URL*, profundidad de los documentos, documentos estáticos y dinámicos, y distribución de sitios por país.

Dando como resultado que la cantidad media de páginas por sitio es 65. Existen un total de 66.021 componentes fuertemente conectados. Dentro de la distribución de los códigos *HTTP* para dar respuesta el que predomina es el OK con un 78.66%. Se observa una longitud promedio de 68 bytes sin incluir la parte correspondiente al protocolo, lo que la incrementaría en 7 bytes. El mayor porciento dentro de la profundidad de los documentos está dado por el número 4 que la presentan 4964279 documentos, para un 40,44%. Dentro de la distribución de sitios por país se encuentra Argentina con 18177 sitios para un 75,87%; siguiéndole en la escala Estados Unidos con 4700 sitios, para un 19.62%.

- **Austria [17].**

Su estudio se dirigió hacia la Embajada de Austria, estableciendo un robot de búsqueda denominado *Data Warehause* y el modelo de procesamiento analítico en línea.

Los indicadores utilizados para hacer la caracterización fueron los siguientes:

Páginas: tipos de archivo, tamaño de los archivos, enlaces externos, direcciones de correos electrónicos y la fecha de la última actualización.

Dominios: direcciones IP, tipos de Red, sistema operativos y software del servidor web.

Lo que dio como resultado encontrar más de 200000 tipos diferentes de archivos sobre la base de sus extensiones, y más de 200 tipos de información. Se encontraron diferencias significativas en las extensiones de vídeo, se proporcionó información en relación con el tipo de servidor web utilizado. Contaban con servidores web Apache, Netscape, Stronghold. De 10 dominios 9 están estrechamente relacionados entre sí. Los distintos estudios realizados han mostrado estadísticamente, por ejemplo, los primeros rastros de documentos XML a principios de 1999 y revelan que la forma en documentos *XML* han ido aumentando en número y como parte de la Web documental que ha estado disponible a partir de 1999 y hasta la fecha. Otro fascinante ejemplo de ello es la sorprendente victoria repentina del PDF sobre el archivo PostScript.

- **África.**

Los datos para el estudio de la Web africana [18] se recolectaron con el *robot* de búsqueda *UbiCrawler*. Para la realización de la caracterización se tomaron en cuenta como puntos de partida los datos obtenidos del análisis sintáctico como nivel de páginas *HTML*, los tipos de cabeceras *HTTP*, tipos de servidores, ultima fecha de modificación, tamaño de la página, lenguaje natural, lenguaje del scripting, extensiones de archivos, protocolos en las *URL* y los gráficos de la Web de África, donde está la distribución, el derecho del poder, la estructura de componentes fuertemente conectados y las interconexiones.

Con la guía de estos indicadores se llegó a la conclusión de que la mayoría de las páginas en el dominio de África no tienen un tipo de documento definido, pues la mayoría de los documentos están realizados en *HTML* 4 y es el 7.71% de toda la Web. La distribución de las cabeceras de las páginas no parece diferir significativamente de los datos que se conocen de manera general las

investigaciones en la Web. La mayoría de los sitios en el dominio de África usan de tecnología de Microsoft-IIS, Apache y Netscape-Enterprise. Más de 600000 páginas están en los 10KB. Los resultados del lenguaje refieren a un total de 7 idiomas prevaleciendo el inglés con un 74.68%.

Dentro de las ocurrencias que tiene un lenguaje Script, predomina el JavaScript con un 32.37%. Entre los protocolos predomina el *HTTP* con un 86.02%, seguido por el mailto con un 29.97%. Los componentes están conectados con uno que viene siendo la cabeza de todos, el denominado gigante.

- **Chile. Análisis de tendencia.**

El presente país es el que más se ha desarrollado en el estudio de su Web, habiendo realizado cuatro caracterizaciones, lo que permite hacer no sólo un estudio cuantitativo sino que permite comparar su contenido con aquellas realizadas anteriormente. Lo que permito realizar un análisis de tendencia y poseer un significativo desarrollo en este sentido.

- **Chile 2000 [3].**

Durante los meses de mayo y junio del 2000 se realizó el primer estudio sobre las características de la Web Chilena, basado en datos obtenidos con el recolector de páginas del buscador *TodoCL*, desarrollado en el Departamento de Ciencias de la Computación de la Universidad de Chile. El análisis de la Web se dividió en cuatro niveles:

Colección: cifras globales y estudio de vocabulario.

Página: tamaño, tipo de documento e idioma.

Sitio: profundidad de la página, número de páginas por sitios, y contenido de texto por sitios.

Dominio: número de referencia hacia y desde un dominio, representación de la estructura global de hipervínculos entre dominios y preferencias de los usuarios.

La colección descargada contaba con 730673 páginas distribuidas en 10352 sitios pertenecientes a 9102 dominios, la misma utilizó 2.3 GB. Se observó que la mayoría de las páginas tienen poco texto siendo el promedio de texto de 3.4 KB mientras para la página en su totalidad es de 15.3 KB. Además que el tipo de documento con mayor empleo es el HTML con más del 95%.

Se observó que el 52% de los sitios poseía una sola página y que prácticamente todos los sitios tenían menos de 100 páginas, lo que dice bastante de la tendencia al estar en Internet de las empresas y organizaciones más que hacer cosas en Internet.

- **Chile 2001-2002 [4].**

En el caso de este estudio se analizó la Web Chilena, a través de los datos recopilados por el buscador chileno *TodoCL*. En el análisis realizado se estudiaron los contenidos de la Web Chilena, principalmente un número de elementos encontrados a nivel de páginas, sitios y dominios.

Una porción importante de los dominios inscritos no se utilizaban, y de aquellos utilizados más de la mitad tenían sólo una página. La de presencia en la Web, el 56% de los dominios y el 54% de los sitios tienen sólo una página. En comparación con el 2000, en que un 45% de los sitios tenía sólo una página, produciéndose un aumento porcentual y absoluto en el número de sitios con una sola página. El tamaño promedio de una página era de 11562 bytes, considerando sólo el texto y tags *HTML*. Sólo el 4% de las páginas contenían más de 40 KB de texto. Además del *HTML* en la Web existen contenidos de diversos tipos, los que también son interesantes de indexar y recuperar. Estos documentos de tipo distinto a *HTML* se dividieron en:

Multimedios: Documentos no indexables por el buscador, a su vez se divide en imágenes, video

y audio.

Texto: Documentos de texto en formato distinto a *HTML*, con filtros pueden ser indexados en la mayoría de los casos.

Servidores de aplicación: Son páginas cuyo resultado es *HTML*, pero son generadas dinámicamente.

Cerca de un 85% del total de documentos incluyendo multimedios son *HTML* o páginas dinámicas que generan *HTML*. Dentro de los documentos de texto el HTML es un 97% del total.

- **Chile 2004 [5].**

En diciembre del 2004 se recorrió la Web chilena usando el sistema *WIRE*, desarrollado en el Centro de Investigación para la Web (CIW). En el presente estudio se analiza La Web a través aspectos como características de las páginas, de los sitios, enlaces entre sitios web, cada uno de ellos con subíndices que le van dando forma al estudio.

En todos los experimentos, usualmente se obtuvo entre 75% y 85% de las transferencias exitosas. La proporción de enlaces rotos, sobre 6 %, es significativa. Para evitar saturar excesivamente el ancho de banda, se descargaron solamente los primeros 200 KB de cada página. Se observó que en un 83% de los casos los sitios web retornan fechas de última modificación válidas.

En el estudio, se limitó al recolector para que descargara solamente 5 niveles para páginas dinámicas, y 15 niveles para páginas estáticas, apreciándose que la cantidad de páginas dinámicas crece exponencialmente en cada nivel. Cerca del 38% de las páginas descargadas eran páginas dinámicas. La aplicación más usada para generarlas es *PHP*, seguida por *ASP* y páginas generadas usando Java. De los 370000 enlaces a archivos que no eran *HTML*, pero que tenían extensiones que son comúnmente usadas para documentos, el formato Adobe PDF es el más ampliamente usado y el estándar de facto, seguido de texto plano y Microsoft Word.

Hay muchos enlaces a archivos multimedia, incluyendo más de 80000000 de enlaces a imágenes, 50000 enlaces a archivos de audio, y 8000 enlaces a archivos de video. Se encontraron enlaces a 30000 archivos con extensiones usadas para código fuente, y 600000 archivos con extensiones usadas para programas.

La Web presentó un promedio de 57 páginas por sitio, 17 sitios con 50000 o más páginas estáticas y sólo otros dos sitios sobrepasaron las 4000 páginas web. El tamaño promedio de un sitio web completo, considerando solamente las páginas *HTML*, es de aproximadamente 0,8 MB. Cerca del 55% de los sitios web fueron creados en el 2004. De acuerdo con los resultados obtenidos por el robot de búsqueda la aplicación para servidor web más usado es Apache con un 70% de participación de mercado, y la segunda aplicación más usada es Microsoft IIS (Internet Information Server) con un 20 %. La distribución de sistemas operativos, en la que Unix y GNU/Linux tienen un 65% de participación, se puede inferir que al menos 1/5 de los servidores basados en Windows usaban Apache. Se encontraron más de 700000 enlaces hacia páginas en otros países.

- **Chile 2006 [6].**

En agosto de 2006 se realizó un estudio de la Web de chile para el cual se utilizó el mismo sistema empleado en el 2004. La colección descargada contaba con más de 7 millones de páginas web, más del doble que las descargadas para el estudio del año 2004. La colección utilizó 50 GB de disco, de los cuales 48 GB corresponden al texto de los documentos y 2 GB a meta datos de las páginas.

Se observó un promedio de 43 páginas por sitio. Donde el 10% de los sitios de mayor cantidad de páginas contienen el 90% de los documentos. Por lo que existen muchos sitios que tienen muy

pocas páginas, lo cual puede ser una señal de poco desarrollo de la Web. Además el idioma que más predomina es el español con alrededor del 80% de las páginas seguido del inglés con un 17%, Otros idiomas tienen una presencia muy leve.

En todas las pruebas realizadas usualmente se obtienen entre 75% y 85% transferencias exitosas, disminuyendo 4 puntos porcentuales respecto al último estudio. También disminuyó cerca de 2 puntos la proporción de los enlaces rotos, ahora en un 4,6%. La disminución de los enlaces rotos puede significar que existe mayor conciencia respecto a verificar la validez de los enlaces.

Más de 3,1 millones de las páginas, el 42,5% del total descargadas, eran páginas dinámicas, es decir, páginas generadas en el momento de ser solicitadas, aumentando la proporción en un 4% respecto a la medición del año 2004. Cerca de un 21% de los sitios de Chile no son fáciles de encontrar ya que están hechos con tecnologías no visibles para los motores de búsqueda, como Flash y JavaScript.

Se encontró aproximadamente 1,1 millones de enlaces a documentos en formatos distintos a *HTML*. Los formatos más populares son PDF (Acrobat), *XML* (se consideran archivos *SVG, RSS, RDF, XML*, etc.) y de texto plano TXT. Respecto al año 2004 se aprecia un avance por parte de las tecnologías XML, mientras que los formatos propietarios *DOC, XLS y PPT* han disminuido su participación; aunque sus contrapartes de código abierto, los llamados Open Document Format, basados en XML, casi no tienen presencia. En audio, el formato MP3 casi dobló su participación en la Web chilena respecto al año 2004, y en imágenes GIF es el más popular en la Web con un 83%.

De todas las páginas existentes en la Web chilena, un 25% de ellas fue creada o modificada durante el período 2005-2006 lo que implica un alto grado de crecimiento y dinamismo. A pesar de ello, es necesario considerar que la mayoría de los usuarios no va muy profunda dentro de los sitios web; esto significa que hay miles o millones de páginas que son visitadas muy raras veces. De hecho existe una fracción no despreciable de páginas que no han sido modificadas en los últimos 8 años.

- **Corea [19].**

El primer estudio sobre el espacio Web coreano fue efectuado en diciembre del 2004, en el cual se utilizo el *WIRE* rastreador, para obtener la información necesaria para la investigación. Donde fueron analizadas varias de sus características a partir de una muestra de más de 50000 sitios los que poseen más de 8 millones de páginas, donde el 10% de los mayores sitios web contenían más del 85% de la información descargada, lo que sugirió que la distribución de calidad era muy desigual, ya que sólo algunas páginas poseían cierta relevancia. Los resultados también mostraron un predominio de formatos estándar como XML o PDF, y herramientas de uso común como ASP y ZIP, en su mayoría pertenecientes al software propietario, lo que se explica debido al bajo apoyo de Asia al uso del código abierto.

- **Cuba [7].**

El estudio va dirigido hacia los meses de septiembre del 2002 a agosto del 2003. Siendo Webalizer, el software utilizado para el análisis de los ficheros *logs* de los servidores web que se emplearon para el Portal Cuba.cu desde 1999. Los resultados fueron obtenidos en cuanto a:

Datos generales del funcionamiento del Portal: total de accesos, total de archivos, total de páginas, total de visitas, total de clientes, total de *URL*, total de páginas de entradas y salidas.

Accesos a Recursos de Información: análisis de Páginas de entradas, Temáticas preferidas por los usuarios, análisis de páginas de salidas.

Dentro de los resultados se obtuvo, total de Acceso 33381340, Total de Archivos: 24129194, Total de Páginas: 4723076, Total de Visitas: 1347647, Total Clientes: 804600, Total de *URL* 274161, Total Pág. Entradas 42770, Total Pág. Salidas 44440.

La cantidad de ficheros que se acceden para visualizar o trabajar con *URL* del tipo de Recursos de Información es de 14.9. Resaltan como páginas de entradas página de inicio portal Cuba.cu, discursos del Comandante en Jefe, servicio de noticias del portal. Dentro de los títulos de las páginas con más de 100 enlaces externos se encuentran Portal Cuba, sitio web CITMA, Cincos Cubanos Inocentes. Los temas de preferencia para los clientes son: Discursos del Comandante en Jefe Fidel Castro Ruz, Ciencia, Tecnología, Medio Ambiente en Cuba, José Martí, Constitución de Cuba, Museos en Cuba, Agencias de Viajes, Partido Comunista de Cuba, Ajedrez en Cuba, Cocina Cubana.

- **España [20].**

La colección fue obtenida entre los meses de septiembre y octubre del 2004, utilizando un programa para recolectar páginas web. El cual comienza descargando un conjunto de direcciones iniciales, que en su caso fueron obtenidas a partir de las referencias incluidas en el buscador Buscopio.

Durante el período del estudio, los resultados obtenidos fueron guiados por indicadores cibermétricos que fueron organizados por niveles:

Páginas: Direcciones *URL*, títulos de las páginas, texto de las páginas, idioma, vocabulario, páginas dinámicas, documentos que no están en HTML, enlaces entre páginas web, ordenamiento usando algoritmos de análisis de enlaces.

Sitios: número de páginas, tamaño de las páginas en un sitio web completo, enlaces internos, enlaces entro otros sitios web, sitios web más referenciados, sitios web con más enlaces, suma de las puntuaciones por enlaces, componentes fuertemente conectados y estructura de enlaces entre los sitios web.

Dominios: dirección IP y proveedor de hosting, software utilizado como servidor, número de sitios por dominio, número de páginas por dominio, páginas por cada idioma por dominio, tamaño total de los dominios, títulos de páginas por dominio, enlaces entre dominios, dominios de primer nivel españoles y dominios de primer nivel externos.

El estudio propició como resultado que las direcciones *URL* más usadas en la Web son las que corresponden al protocolo *HTTP*. El 80% de las direcciones *URL* tienen entre 40 y 80 caracteres. Se observa que hay muchas páginas con muy poco texto y unas pocas páginas con un tamaño enorme. El castellano es usado por poco más de la mitad de las páginas, seguido por el inglés y el catalán. La proporción total de páginas escritas en los idiomas oficiales del país es de aproximadamente 62%. Más de 3,5 millones (22%) de las páginas descargadas eran páginas dinámicas. La aplicación más usada para construir páginas dinámicas es PHP4.

Se encontró aproximadamente 200 000 enlaces a ficheros que no eran HTML lo que, si bien es un número grande de documentos, representa sólo un 1% de las páginas totales en la Web. Dentro de las características de los sitios tenemos que el número promedio de páginas por sitios es 52. Además los dos servidores dominantes eran Apache y Microsoft IIS (*Internet Information Server*), con ventaja para Apache. El sistema operativo más usado para servidores era Windows (43%), seguido muy de cerca por sistemas operativos basados en Unix (41%); esto significa que al menos el 15% de los servidores basados en Windows prefieren Apache. En promedio se encontró 2.55 sitios por dominio. Hay un promedio de 133 páginas por dominio. El tamaño promedio de un dominio web completo considerando solamente el texto, es de aproximadamente 373 Kilobytes, sólo un 16% de los títulos son únicos.

- **Hungría [21].**

La primera parte de su investigación se basó en la arquitectura del motor de búsqueda, utilizando finalmente un motor que se diferenciaba en varios aspectos del diseño de la mayoría de los motores; como un ejemplo de aplicación eficiente de rastreo e indexación de las políticas que pueda permitir la búsqueda de las noticias de última hora. Su experimento se dirigió hacia los indicadores como las cantidades de páginas, cantidad de sitios web, tamaño de los archivos y el idioma y tipos de palabras.

Según se estimó se contaba con no mucho más de diez millones de páginas bajo el dominio .hu pertenecientes a aproximadamente 300000 sitios web. Para el estudio se rastreó solo cinco millones de páginas. Donde se determinó que el idioma húngaro se encontraba entre 70-90% de las ocurrencias, mientras inglés 27-34% lo que permitió deducir que las mayoría de las páginas eran multilingües o bilingües. Además fuera de la .hu se encontraron 280000 páginas, sobre todo en húngaro. También se realizó una medición preliminar para la vida de los documentos *HTML*.

- **Perú [22].**

La recolección de la muestra para el estudio se realizó con el software *crawler WIRE*, en el mes de agosto del año 2006. La investigación tomó camino hacia indicadores cibermétricos que fueron agrupados por categoría:

Sitios y Páginas: tamaño de los sitios, páginas por sitios, enlaces entre sitios y contenido de las páginas.

Enlaces y Ranking: distribución de grado, *ranking* de sitios y macro estructura del espacio web.

Se descargaron 1629745 páginas desde 8908 sitios, que corresponden a 7945 dominios de tercer nivel. Existen más del 55% de los sitios con un máximo de 10 páginas, Se halló 5688 sitios (64%) sin enlaces entrantes lo que implica claramente posibles problemas de visibilidad. y 5948 (68%) sin enlaces salientes lo que provoca una baja conectividad.

Por lo que se podía apreciar que el espacio web de Perú se hallaba débilmente interconectado ya que la componente MAIN es pequeña y existe un 53% de sitios en la región islas.

- **Portugal [23].**

Para la caracterización de la Web de Portugal, se recolectaron los datos con el robot de búsqueda *Viúva Negra Crawlers*, insertándolos al Versus donde se guardaban los documentos en archivos y meta-datos.

Los indicadores que fueron utilizados son: el idioma, los servidores web, los dominios, cantidad de direcciones *URL*, tamaño de las *URL*, lenguajes de los documentos, tamaño de los datos producidos y de los textos, dando como resultados que el idioma predominante es el portugués con un 73%, el dominio en que se mueve la Web Portuguesa es el .pt, aunque tienen espacios en la .com y .net. Consta de 4 millones de *URL* para un tamaño de 78 GB y el HTML con un 95%, es el lenguaje predominante.

1.7 Conclusiones del capítulo.

En este capítulo se realizó un estudio que ayuda en gran medida a la realización de la tesis, brindando información que de una forma u otra, ha sido de gran utilidad para el desarrollo de la investigación.

Debido a la abundancia de contenido, por no tener un punto de parada del incremento de sitios en la Web, es que existen páginas que no están reconocidas y más bien pertenecen a las llamadas lagunas de la información. De aquí que surja la Webmetría o ciencia que estudia la Web; que a través de distintas herramientas cibermétricas, como por ejemplo, un robot de búsqueda, realiza la recolección de las páginas de uno o más sitios web, y con esto un análisis cuantitativo y cualitativo a través de indicadores que ayuden a establecer un punto de partida llamado caracterización o estudios de la Web. Estos estudios pueden llegar a ser engorrosos para los que los llevan a cabo, aunque existen países que los han realizado con mucho éxito.

CAPÍTULO II – Definición de herramienta e indicadores cibermétricos para caracterizar la Web de la UCI.

2.1 Indicadores cibermétricos.

Los indicadores, en términos generales, representan una medición agregada y compleja que permite describir o evaluar un fenómeno, su naturaleza, estado y evolución. La aplicación de la métrica y de las medidas cuantitativas a la información electrónica se ha convertido en un área significativa para la investigación.

En el campo de los indicadores, las discusiones conceptuales se quedan un tanto al margen. Casi todos los autores se limitan a su aplicación sin delimitar si pertenecen a una u otra disciplina métrica, y se limitan a apellidarlos en correspondencia con el enfoque seguido en el trabajo -indicadores bibliométricos[3], informétricos[4], cienciométricos[5], patentométricos[6], entre otros-, sin diferenciar mucho en el tipo de indicador sino más bien en la finalidad con la que se emplean [24].

La siguiente es una agrupación de los indicadores según las variables o categorías más generales que permiten evaluar:

Categoría	Indicador
Infraestructura	Número de host, de servidores Web, de usuarios, de dominios, de sitios, de sitios institucionales, etc.
Tamaño	Número de páginas, de objetos, de objetos multimedia, de archivos ejecutables, tamaño de los archivos, distribución por lenguajes, evolución temporal, número de niveles, de enlaces por página, etc.
Calidad	Porcentaje de enlaces válidos, de errores de enlace, apariencia, etc.
Conectividad	Total de enlaces, de enlaces por página, número de enlaces internos, de enlaces externos, etc.
Visibilidad	Número de enlaces recibidos o externos, enlaces nacionales externos, enlaces internacionales externos, etc.
Impacto	Factor de impacto
Popularidad	Número de visitas

[3]http://es.wikipedia.org/wiki/Bibliometría
[4]http://es.wikipedia.org/wiki/Informetría
[5]http://es.wikipedia.org/wiki/Cienciometría
[6]http://es.wikipedia.org/wiki/Patentometría

2.1.1 Indicadores más frecuentes empleados para el análisis métrico.

El desarrollo de las tecnologías de la información y las comunicaciones ofrece nuevos escenarios para la realización de los estudios métricos de la información, sobre todo de aquella que circula por Internet, con las facilidades del *WWW*. Internet ha supuesto una revolución sin precedentes en el mundo de la informática y las comunicaciones. Esta red constituye un universo de recursos de información y un espacio virtual de comunicación entre usuarios.

Es necesario un gran esfuerzo teórico para sentar las bases de la CIBERMETRIA[7] como nueva disciplina de descripción cuantitativa, diferente y complementaria a las ya existentes:

- ✓ Los INDICADORES [24] son los principales resultados de dicho esfuerzo y deben ser definidos y calculados con rigurosidad.
- ✓ Las técnicas de recopilación automática mediante AGENTES [1] demuestran sus grandes posibilidades y capacidad de generar muestras amplias y representativas.
- ✓ Los indicadores básicos de tamaño, popularidad, visibilidad e impacto juegan un papel fundamental.

La siguiente lista recoge los indicadores más frecuentes empleados para el análisis métrico de los recursos digitales en diversos trabajos sobre la temática.

- ❖ Indicadores de tipos institucionales -cantidad de páginas en determinados sectores.
- ❖ Indicadores regionales.
- ❖ Indicadores idiomáticos.
- ❖ Indicadores de tipología de sitios -sitios académicos, comerciales, de sectores públicos o privado.
- ❖ Indicadores de tamaño, en sus dos variantes, tamaño documental (número total de páginas comprendidas en un dominio) o tamaño informático -tamaño en bytes de una sede Web.
- ❖ Indicadores de densidad, también tiene dos variantes, densidad hipertextual -media de enlaces por página- y densidad multimedia -media de objetos multimedia por página.
- ❖ Indicadores de profundidad -número máximo de niveles de una sede.
- ❖ Indicadores de luminosidad -total de enlaces emitidos desde una sede.
- ❖ Indicadores de visibilidad -número total de enlaces externos diferentes recibidos por una sede, existen variantes para calcular la visibilidad nacional con límite a los enlaces recibidos en el mismo país, etc.
- ❖ Indicador de navegabilidad -número total de enlaces internos respecto al total de páginas.
- ❖ Indicadores de validez hipertextual -porcentaje de enlaces válidos respecto al total.
- ❖ Indicadores de cooperación (colegios invisibles).
- ❖ Indicador de diversidad -distribución de las características de los enlaces recibidos por una página.
- ❖ Medidas de popularidad -número y distribución de las visitas recibidas en un plazo determinado.
- ❖ Indicadores de impacto -resultado de dividir el número total de enlaces externos diferentes recibidos por una sede por su tamaño expresado en número de páginas.

[7] http://es.wikipedia.org/wiki/Cibermetría

- Indicadores para el estudio del comportamiento de usuarios en la recuperación de información.

2.2 Definición de indicadores para el estudio de la Web de la UCI.

Los siguientes indicadores están básicamente agrupados en 3 categorías, y son a su vez, los que sirven para realizar los Estudios Webmétricos de la UCI. Cada uno de ellos brinda un dato importante acerca de algún parámetro medible de la Web, que por supuesto, puede ser útil después a la hora de tomar decisiones con respecto a la misma, partiendo de los resultados analizados.

Estos niveles, en los que se agrupan los indicadores son: Colección, Sitio, y Página. No se toma en cuenta un nivel Dominio pues, para este caso, los estudios estarán centrados en un solo Dominio; el .uci.cu y, por tanto, no es necesaria tal agrupación.

A continuación los indicadores seleccionados, así como una explicación de cada uno de ellos:

- Nivel Colección
  - Datos Generales.
    - Este indicador mostrará los datos más generales de la colección de información. Por ejemplo, la cantidad de sitios analizados, la cantidad general de páginas web descargadas durante el estudio, entre otros. Se detallará además acerca del tipo de software utilizado como servidor de aplicaciónes web y los sistemas operativos con que cuentan los distintos servidores de la Web de la UCI.
- Nivel Sitio
  - Cantidad promedio de páginas por sitio.
    - Este indicador ayuda a tener una idea del tamaño de los sitios aunque no es, en todos los casos, un dato verídico; pues en el estudio suelen aparecer algunos sitios con muy pocas páginas, y otros con muchas más páginas que el promedio mencionado. No por esto deja de tener utilidad e importancia.
      - Cantidad promedio de páginas estáticas por sitio.
      - Cantidad promedio de páginas dinámicas por sitio.
  - Tamaño total de la colección de información. (Espacio ocupado en Disco Duro)
    - Este indicador está dado precisamente por el tamaño del texto en bruto descargado en el disco duro durante el estudio. Brinda una media del tamaño de la información disponible en la Web. Aquí se excluyen las etiquetas, imágenes, y ficheros multimedia o de otros formatos, pues solamente se descarga el texto de las páginas web.
  - Tamaño promedio en Megabytes de los sitios.
    - Este indicador está directamente relacionado con el explicado anteriormente, pues precisamente se obtiene a partir del mismo; al dividir el tamaño total de la colección de información entre la cantidad de sitios analizados.
  - Profundidad máxima promedio de los sitios.
    - Este indicador permite conocer de manera aproximada cuan trabajoso o no puede ser para el usuario la navegación por los sitios de la colección analizada. Más

adelante se explica con más detalle este punto.

- Promedio de Grado Interno de los sitios.
  - El grado interno de un sitio está dado por la cantidad de sitios en la colección que hacen referencia a el desde alguna de sus páginas, por lo cual se considera popular un sitio que tenga un grado interno muy elevado, aunque esto pudiera significar además resultado del tipo de sitio que se visita.
- Promedio de Grado Externo de los sitios.
  - El grado externo de un sitio está dado por la cantidad de sitios en la colección a los que hace referencia desde alguna de sus páginas. Un sitio de carácter comercial, tratará de tener un grado externo pequeño, para evitar que los usuarios abandonen el sitio siguiendo enlaces a otros sitios. Conocer esta información ayuda a saber el nivel de comunicación que hay entre los sitios de la colección.
- Distribución de páginas por sitios.
  - Este indicador muestra los sitios con mayor cantidad de páginas de la colección, así como los de menor cantidad. Permite conocer aquellos sitios que sólo tienen una página o un número muy pequeño de las mismas. Los sitios con muchas páginas dinámicas tienden a crecer exponencialmente y suelen tener un número ilimitado de páginas web.

- Nivel Página
  - Cantidad de páginas únicas/duplicadas de la colección.
    - Este indicador ayuda a tener una idea del nivel de replicación del conocimiento que existe en la Web. O sea, cómo un mismo contenido se encuentra disponible en varias direcciones, lo cual es beneficioso para el usuario cuando este contenido replicado es importante, o de consulta frecuente.
  - Cantidad de páginas estáticas/dinámicas de la colección.
    - Este indicador refleja la evolución de la Web. La proporción de páginas dinámicas de la colección ayuda de cierto modo a caracterizar la Web en este sentido; brinda una idea de cuánto ha avanzado desde la Web 1.0 (Estática) hacia la Web 1.5 (Dinámica) o, incluso, pensando de manera futurista hacia la Web 2.0, dirección en la cual se trabaja actualmente en la Universidad.
  - Profundidad de las páginas de la colección.
    - Partiendo de que la profundidad lógica de una página web no es más que la cantidad mínima de veces que el usuario tiene que dar *click* en un vínculo, para llegar a la misma sin abandonar el sitio web y comenzando desde la portada del mismo, este indicador brinda información muy importante y útil. Durante el desarrollo actual de las tecnologías de la información el tráfico de usuarios por las redes ha aumentado considerablemente y, con esto, también han aumentado las necesidades de los mismos y la búsqueda de comodidades en la navegación. Con este indicador se puede establecer una media del trabajo que representa para el usuario llegar a una información determinada en la Web. Las páginas relevantes para el usuario no deben encontrarse a grandes profundidades y, al mismo tiempo, la media de acceso no debe tener valores muy elevados; pues sucede que el usuario se aburre de dar *click* y sencillamente abandona el sitio web. La portada de un sitio web tiene profundidad lógica 1 y las páginas que se encuentran accesibles directamente desde la portada tienen profundidad lógica 2 y así sucesivamente.

También se puede hablar de una profundidad física, en la cual las páginas de la forma http://sitio/pag.html o http://sitio/dir/ están a profundidad 2 y así sucesivamente. La profundidad física mide la organización en archivos y directorios de un sitio web.

- Edad de las páginas de la colección.
  - Este indicador ayuda a conocer la media de creación de páginas web, lo cual con el desarrollo de la Universidad debería ser un número en crecimiento constante. Si este número estuviera decreciendo, significaría que la creación de páginas web lo hace también y, por tanto, si no se realiza una actualización exhaustiva de las páginas existentes, llegaría un punto en que la información brindada por muchas de ellas pudiera ser incierta o desactualizada; lo cual es, desde varios puntos de vista, poco deseado.
- Extensiones encontradas en el estudio.
  - Con este indicador se pueden conocer las extensiones que prevalecen en la Web, según cada una de las siguientes clasificaciones y de manera general. Además ayuda a identificar la existencia de extensiones desconocidas y el nivel en el cual están presentes las mismas. Se pueden establecer medias del uso de extensiones de Software Libre por estos datos, que dan una idea del nivel de impacto y utilización del Software Libre en la Universidad.
    - Extensiones de audio, video e imagen.
    - Extensiones de interfaz de entrada común (*CGI, por sus siglas en inlgés*) y código fuente.
    - Extensiones de software.
    - Extensiones de documentos que no son *HTML* ni *TXT*.
    - Extensiones de ficheros comprimidos.
    - Extensiones extras.
- Lenguaje de las páginas de la colección.
  - Este indicador ayuda a identificar los distintos idiomas utilizados en la Web analizada, así como su nivel de presencia en la misma. Siempre teniendo en cuenta que para poder definir el lenguaje de una página web debe tener al menos 50 palabras y, de ellas, al menos unas 10 en el lenguaje en cuestión; de lo contrario, no se puede identificar un lenguaje para la misma.

2.3 Herramientas tecnológicas empleadas para desarrollar estudios cibermétricos.

Hay muchas razones para explorar de forma automática la red. La más inmediata deriva de su tamaño, que hace implanteable una exploración manual, salvo para fracciones muy pequeñas. La estructura de grafo hace factible la utilización de programas de navegación automática; tales programas se conocen con diversos nombres: *agentes o robots web, spiders, wanderers* o, incluso, gusanos (*worms*) [1].

Algunas de las razones más frecuentes para el uso de estas herramientas que recorren automáticamente el ciberespacio son:

1. Análisis estadístico de la red. Desde la obtención de datos simples (tamaño de sedes web, por ejemplo) hasta trabajos mucho más complejos que requieren de herramientas automáticas que recolecten, recuenten y calculen coeficientes.
2. Labores de mantenimiento: detección de enlaces rotos o erróneos, verificación de sintaxis *HTML*, etc.
3. Tareas de copiado o *mirroring* con diversos propósitos.
4. Recuperación de Información, una de las tareas más comunes.

Entre las principales herramientas empleadas se encuentran:

- Agentes Mapeadores (Gestores de sitios web):
    - Astra Site Manager 2.0 (www.merc-int.com)
    - COAST Web Master 7.0 (www.coast.com/)
    - Content Analyzer 3.0 (www.microsoft.com/siteserver)
    - Funnel Web Profiler 2.0 (www.quest.com/)
    - LinkBot Pro 6 (www.watchfire.com)
    - LinkViewer 3.0 (www.gradetools.com)
    - Microsoft Site Analyst 2.0.
    - Site Mapper 2.0 (www.msw.com.au/mapper)
    - SiteXpert 9.0 (www.xtreeme.com/sitexpert)
    - WebAnalyzer 2.01 (wsa-web-site-analyzer.softonic.com)
    - WebKing 4.1 (www.parasoft.com/)
    - WebTrends Prof Suite 7.0 (www.webtrends.com)
- Verificadores de Enlaces: (Software)
    - Alert Link Runner 6.0 (www.alertbookmarks.com/lr)
    - CSE HTMLValidator 9.0 (www.htmlvalidator.com)
    - CyberSpyder 3.4.0 (www.cyberspyder.com)
    - LinkBot Pro 6 (www.watchfire.com)
    - LinkMan Prof 7.6 (www.outertech.com)
    - LinkScan 12.0 (www.elsop.com)
    - LinXCop 2.6 (www.filehouse.com/linxcop
    - Web Link Validator 5.0 (www.relsoftware.com/wlv/)

- Verificadores de Enlaces: (Online)
    - W3C Link Checker (validator.w3.org/checklink/)
    - Volcadores de sitios web.
    - AaronWebVacuum 2.8 (www.surfwarelabs.com
    - BlackWidow 5.0 (www.softbytelabs.com)
    - HTTrack Website Copier 3.43 (www.httrack.com)
    - inSITE 1.0 (www.rocketdownload.com/Details/Inte/insite.htm)
    - JOC WebSpider 5.5.2 (www.jocsoft.com)
    - Offline Explorer Pro 5.4 (www.metaproducts.com)
    - PageNest Free Offline Browser 3.17 (pagenest.com)
    - SuperBot 2.60 (www.sparkleware.com/superbot
    - Teleport Pro 1.59 (www.tenmax.com/)
    - Website Extractor 9.85 (www.asona.org/)
    - WebCopier Pro 5.0 (www.maximumsoft.com/)
    - WebReaper 10 (www.webreaper.net)
    - WebWhacker 5.0 (www.bluesquirrel.com
    - WebZip 7.1 (www.spidersoft.com/)
    - Wysigot 6.0 (www.ecatch.com)

2.4 Selección de herramienta a utilizar en el estudio de la Web de la UCI.

Existen, como se puede apreciar en el epígrafe anterior, muchas herramientas para el estudio del ciberespacio; pero no todas están realmente disponibles para uso abierto. Ciertamente existen, por ejemplo, varios servicios de búsqueda en Internet, cada uno de los cuales tiene su propio *spider* o robot; pero debemos contentarnos con los datos que estos nos facilitan, lo cual muchas veces no resuelve nuestros problemas. De igual manera, el valor económico de muchos de estos servicios hace que, en ocasiones, la literatura sea bastante evasiva acerca de detalles concretos sobre el funcionamiento de muchos de ellos. Por tanto, no son muchos los spiders disponibles para uso abierto, es decir, con licencia GPL o parecida.

No obstante, algunos de los que se pueden tener disponibles son:

- **WebBot**. Disponible en http://www.w3.org/Robot/, se trata de un proyecto del World Wide Web Consortium (W3C).
- **Harvest-NG**. Disponible en http://webharvest.sourceforge.net/ng/, se trata de un conjunto de utilidades para construir webcrawlers y está escrito en lenguaje perl.
- **Webvac Spider**. Disponible en http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/webbase-pages.html, es un proyecto de la Universidad de Stanford.

- **SocSciBot 4 y SocSciBotTools**. Disponibles en http://socscibot.wlv.ac.uk/, es una opción interesante con utilidades adicionales.
- **WIRE Crawler**. Disponible en http://www.cwr.cl/projects/WIRE/, es el desarrollo del Centro de Investigación del Web (CIW) dirigido por el Dr. Ricardo Baeza-Yates.
- **SacarinoBot y EloisaBot Tools**.

2.4.1 Funcionamiento básico de un Spider [1].

El procedimiento básico de un robot consiste en suministrar una *URL* inicial o un conjunto de ellas, obtener la página web correspondiente y a continuación extraer todos los enlaces existentes en dicha página. Con los enlaces obtenidos es necesario realizar una serie de operaciones previas de normalización y luego comprobar las *URL* que se habían seguido previamente y, en caso de no haberlas recorrido, introducirlas en una cola de direcciones *URL* a seguir. Por último, se almacena la información bien en bases de datos o en estructuras de ficheros con codificación ASCII; se obtiene la *URL* del siguiente enlace a seguir y comienza de nuevo el proceso. (Ver Fig. 1)

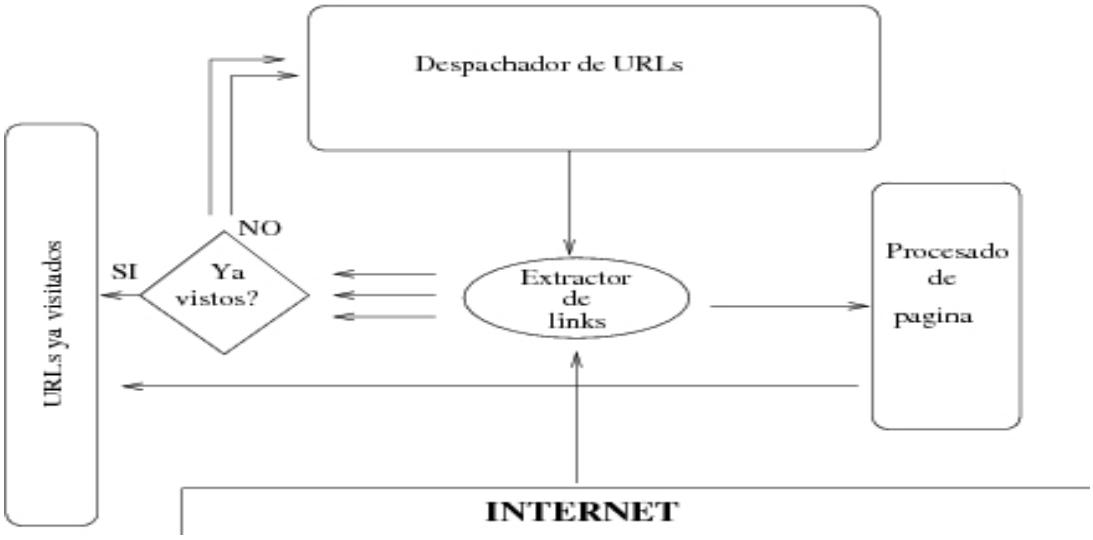


Figura 1: Esquema básico de un robot.

Uno de los aspectos más interesantes, en dicho proceso, es la gestión de la lista que almacena los enlaces que se van encontrando. El aspecto clave es el orden en que se exploran las direcciones que se guardan en dicha lista, a la que constantemente se están añadiendo nuevas direcciones.

Suelen distinguirse, en función de esto, tres formas básicas de exploración:

1. Recorrido en anchura (*breadth-first*).
2. Recorrido en profundidad (*depth-first*).
3. El mejor posible (*best-first*).

Las dos primeras son bien conocidas y dependen de que dicha lista se organice como una pila o como una cola. En el primer caso, el entorno inmediato a una página es explorado primero; resulta útil cuando se buscan exploraciones exhaustivas de un dominio, aún cuando se limiten a una determinada profundidad solamente. Si la lista se organiza como una cola, entonces la exploración se aleja inmediatamente del punto de partida y, hasta que cada posible camino no ha sido recorrido, no se vuelve a éste; lo cual resulta útil, pues ayuda a descubrir información poco obvia o difícil de obtener mediante navegación manual. (Ver Fig. 2)

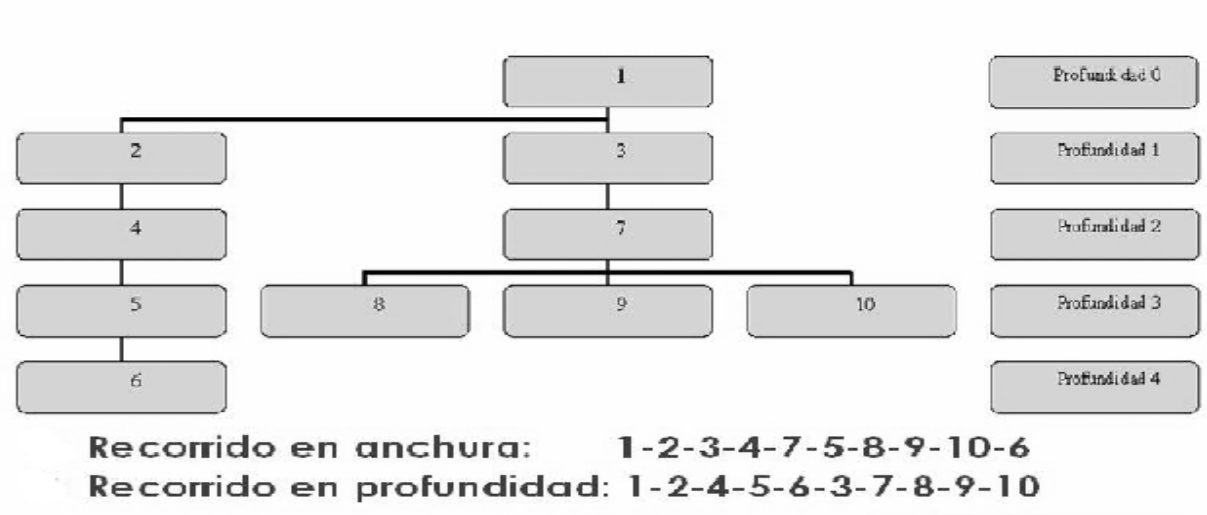


Figura 2: Resultados según el tipo de recorrido.

El tercero significa ordenar la lista de direcciones a explorar por el mejor candidato y aplicar el criterio que se estime más adecuado, en función del uso que se quiera hacer del robot. Un ejemplo simple sería priorizar o explorar antes aquellos elementos de la lista más citados, es decir, los enlaces que se repiten más veces, en la suposición de que deben ser los más populares o importantes.

Naturalmente , es posible aplicar otros coeficientes más sofisticados. Por ejemplo, cuando lo que se busca es que el robot obtenga páginas o direcciones sobre un tópico determinado, es priorizar aquellos enlaces que presumiblemente lleven a nodos más relacionados con ese tópico.

El comportamiento de un robot de búsqueda es el resultado de una combinación de cuatro políticas:

1) ***Política de selección***: Indica las páginas a descargar.
2) ***Política de revisión***: Revisa si ha existido algún tipo de cambio en el estado de las páginas.
3) ***Política de cortesía***: Establece la manera de evitar la sobrecarga en los sitios web.
4) ***Política de paralelismo***: Establece la forma de coordinar los rastreadores web que están distribuidos en ese momento.

Política de Selección.

Dado el tamaño actual de la Web, incluso los grandes robots de búsqueda sólo cubren una parte de la disposición del contenido de la misma, descargando sólo una fracción de las páginas web, es muy conveniente que esa fracción descargada contenga las páginas más relevantes y no sólo una muestra aleatoria de la Web.

Para ello se requiere un indicador de importancia que otorgue prioridad a las páginas web. La importancia de una página es una función de su calidad intrínseca, su popularidad en términos de enlaces o visitas, edad o tamaño, e incluso de su *URL*. Diseñar una buena política de selección tiene una dificultad añadida, ya que debe trabajar con información parcial, como el conjunto completo de páginas Web que no se conoce durante el rastreo.

Política de revisión.

La Web tiene un carácter dinámico y rastrear una fracción de las misma puede llevar mucho tiempo, por lo general medido en semanas o meses si es muy grande dicha fracción. En el momento en que un rastreador web ha concluido, muchos acontecimientos podrían haber sucedido. Estos eventos se caracterizan como creaciones o actualizaciones en la Web.

Política de cortesía.

Es una norma que le permite a los administradores indicar qué partes de sus servidores web no deben ser visitados por los robots, aunque realmente no garantiza que sea respetada por los mismos de manera alguna; lo anterior depende completamente de la programación de los mismos. La presente política no incluye una sugerencia para el intervalo de visitas a un mismo servidor, aunque este intervalo es la forma más eficaz de evitar la sobrecarga del mismo.

Política de paralelismo.

Un rastreador paralelo ejecuta múltiples procesos simultaneamente. Los objetivos son aumentar al máximo la descarga, reducir al mínimo los gastos generales de paralelización y evitar una repetición de las descargas de la misma página.

2.4.2 Alternativas de herramientas libres.

- **WebBot**. Disponible en http://www.w3.org/Robot/, se trata de un proyecto del *World Wide Web Consortium (W3C)*. Fue desarrollado a finales de 1990, en principio, para realizar diversas predicciones para asuntos del mercado.
  Es un robot buscador muy rápido trabajando en una web, que soporta expresiones regulares y registro de logs *SQL*. Está basado en la librería libwww *HTTP*/1.1[8] y se puede utilizar, entre otras cosas, para comprobar links, validación de código *HTML* en páginas, descarga de imágenes, creación de mapa de un sitio web, modificación de fechas sobre la base de la última modificación en el campo de cabecera *HTTP* y distribución de contenidos y tipos de caracteres encontrados en el recorrido por el content-type[9] de los documentos. Realiza una búsqueda tradicional, basada en la tala de archivos comunes utilizando los formatos del archivo de registro y las comprobaciones de los hipervínculos, así como la de las imágenes robustas. Realiza un uso limitado de las peticiones GET y HEAD pues solo descarga lo estrictamente necesario. Puede ser usado para recorrer numerosos enlaces, mas debe utilizarse con cuidado, pues no está diseñado para recorrer el Internet en general.

- **Harvest-NG**. Disponible en http://webharvest.sourceforge.net/ng/, es una colección de scripts de Perl y módulos que proporcionan una potente red de rastreo. Fue desarrollado en este lenguaje aprovechando muchas de las actuales herramientas del mismo y tiene por objetivo proporcionar un código abierto, compatible con las normas y la herramienta para recopilar el contenido de una amplia variedad de fuentes de información, que se resume en un conjunto de descripciones de recursos y el almacenamiento de estos en una base de datos de fácil acceso, que posee servicios de búsqueda a partir de los cuales se puede construir la información estadística recopilada.
  Harvest-NG soporta una gran variedad de formatos de contenido, principalmente a través de la utilización de convertidores externos. El código básico está diseñado para trabajar con *HTML* y texto plano pero, al añadir convertidores y traductores, varios tipos de contenido pueden ser soportados. Está diseñado para ser capaz de interactuar con muchos convertidores de libre adquisición, como mswordview, pdftotext y pstotext, aumentando la gama de tipos de contenido soportados.
  Harvest-NG almacena todos los recursos de las descripciones en una base de datos, junto con otra información sobre el contenido. Esta base de datos está gestionada internamente,

[8] http://www.w3.org/Library/

[9] http://es.wikipedia.org/wiki/Multipurpose_Internet_Mail_Extensions

sin necesidad de sistemas externos. La interfaz de la base de datos es clara y bien documentada, por lo que además de utilizar una serie de herramientas incluidas con el programa, puede ser usada también para crear utilidades que puedan ser ejecutadas sobre los datos recogidos.

- **Webvac Spider**. Es un proyecto de la Universidad de Stanford, disponible en http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/webbase-pages.html.
  Este robot rastrea generalmente hasta una profundidad de 7 a 12 niveles tanto para páginas estáticas como dinámicas y obtiene un máximo de 10KB de páginas por sitio. Sólo sigue los vínculos del dominio, por lo que los rastreos los hace más estables sobre la lista de los sitios. Se demora de uno a diez segundos entre las páginas.
  Actualmente la Universidad de Stanford cuenta con un repositorio de más de 117 Terabytes de información, a partir de los distintos recorridos realizados de diversas páginas web destinadas a la investigación, en temas como el análisis gráfico web e indexación de páginas.
  Generalmente rastrea la misma lista de sitios cada vez que hace un recorrido. Presenta una colección de los enlaces de cada uno de los rastreos, así como la información recolectada. El texto general recolectado tiene alrededor de 0.5 Terabytes comprimido y unos 1.5 Terabytes sin comprimir.

- **SocSciBot 4 y SocSciBotTools**. Disponible en http://socscibot.wlv.ac.uk/, es una opción interesante con utilidades adicionales. Es un rastreador diseñado con fines de investigación. Junto con el están los programas de apoyo de las Herramientas SocSciBot. Se puede utilizar para llevar a cabo análisis de enlaces en un sitio o en los sitios de recolección, o para ejecutar un motor de búsqueda en una colección de sitios.
  El programa se ejecuta en Windows 95, rastrea los sitios con un máximo de 15000 páginas y no presenta restricciones en la velocidad. Los que utilizan este robot no tienen garantía de que el mismo funcionara como debe ser, ni de que los resultados sean los esperados sobre la recolección de datos. Necesita un ancho de banda relativamente grande para que su funcionamiento sea aceptable. No trabaja en servidores que presenten sobrecarga. Para su utilización hay que aceptar que sea conectado a distancia, es para que los propietarios se puedan asegurar de que no está siendo usado en un modo poco ético.
  Asimismo, sus creadores no aceptan responsabilidad por los daños derivados de su utilización o por la pérdida de datos o de otros problemas causados por las operaciones de los programas descargados.

- **UbiCrawler**. Disponible en http://law.dsi.unimi.it/ubicrawler/
  Es un rastreador distribuido y escrito en Java, que no tiene un proceso centralizado. Se compone de un número de agentes y la función de asignación se calcula utilizando de forma coherente los nombres de host. Hay superposición de ceros, lo que significa que la página no se rastrea en dos ocasiones, a menos que exista un rastreo agente de accidente lo que provocará que el otro agente deba volver a rastrear las páginas. El rastreador es diseñado para lograr alta escalabilidad y ser tolerante a fallos, aunque no se distribuye públicamente, sino que puede ser utilizado para la investigación o fines comerciales, siempre y cuando se obtenga el permiso de sus autores para su utilización.

- **SacarinoBot y EloisaBot Tools**.
  Son proyectos en desarrollo, del Grupo de Recuperación Avanzada de Información, de la Universidad de Salamanca. na muestra de su trabajo puede ser encontrada en http://www.fesabid.org/madrid2005/descargas/presentaciones/actividades/alonso_jl.pps

- **WIRE Crawler**. Disponible en http://www.cwr.cl/projects/WIRE/, este proyecto chileno *WIRE* es un esfuerzo iniciado por el Centro de Investigación Web (Center for Web Research) dirigido por el Dr. Ricardo Baeza-Yates, para crear una aplicación que permita la recuperación de información; diseñada para ser utilizada en la Web.
  Actualmente el software *WIRE* incluye:
  1. Un formato simple para almacenar una colección de documentos web.
  2. Un rastreador web.
  3. Herramientas para la extracción de las estadísticas de la colección.
  4. Herramientas para la generación de informes acerca de la colección.

  Las principales características del software *WIRE* son las siguientes:
  Escalabilidad: diseñado para trabajar con grandes volúmenes de documentos, ha sido probado con varios millones de documentos.
  Prestaciones: programado en C/C++ para un alto rendimiento.
  Configurable: todos los parámetros para el rastreo y la indexación se pueden configurar a través de un archivo XML.
  Análisis: incluye varias herramientas para analizar, extraer estadísticas y la generación de informes sobre sub-conjuntos de la Web, por ejemplo: la web de un país o de una gran intranet.
  De código abierto: el código está libremente disponible.

  Además el sistema está diseñado para centrarse en la evaluación de la calidad de la página, utilizando diferentes estrategias de rastreo y la generación de datos web para la caracterización de los estudios.

  El robot WIRE se compone de diversos programas o módulos que lo ayudan con el funcionamiento, normalmente el mismo funciona de forma reiterativa, los programas son ***wire-bot-reset, wire-bot-seeder, wire-bot-manager, wire-bot-harvester, wire-bot-gatherer y wire-bot-run.***

  WIRE-BOT-RESET
  Este módulo se utiliza para resetear el repositorio donde se almacenará la información del recorrido del robot web; o sea borrar, limpiar toda la información del mismo y crear los directorios, carpetas o estructuras de datos necesarios para un nuevo recorrido. El tiempo utilizado para esto es relativo y depende directamente de los valores de las variables maxsite y maxdoc en el fichero de configuración del spider.

  WIRE-BOT-SEEDER
  Este módulo recibe las direcciones *URL* iniciales para el recorrido y añade al repositorio los documentos necesarios para las mismas. El mismo se utiliza tanto para dar al rastreador un conjunto inicial (lista de partida) de direcciones *URL*, como para analizar las direcciones *URL* que se extraen de los programas recolectores de las páginas descargadas.

  WIRE-BOT-MANAGER
  Este módulo organiza y muestra los documentos de la colección mediante sus resultados, y crea lotes de documentos para el módulo de recolección (wire-bot-harvester). Los resultados se otorgan por una combinación de factores que se describen en el archivo de configuración. Proporciona una noción de los documentos descargados hasta el momento y de los que restan dentro de la lista de direcciones por analizar.

WIRE-BOT-HARVESTER
Este módulo descarga los documentos de la Web. El programa trabaja en su propio directorio con sus propias estructuras de datos y se puede detener en cualquier momento, utilizando el módulo wire-bot-manager que comprueba el estado del lote de documentos descargados y lo borra en caso de estar incompleto o incorrecto.

WIRE-BOT-GATHERER
Este módulo analiza los documentos descargados de la Web durante el lote actual y extrae las nuevas direcciones *URL*. El mismo toma las páginas descargadas por el módulo recolector (wire-bot-harvester) de su directorio y las combina en la colección principal.

WIRE-BOT-RUN
Este módulo ejecuta varios ciclos rastreadores de la forma "seeder-manager-harvester-gatherer".

2.4.3 Herramienta seleccionada para el estudio.

Luego de haber terminado el estudio de las distintas herramientas propuestas, analizando por separado las características principales de las mismas para la recolección de datos y teniendo en cuenta factores como los objetivos propuestos por sus creadores para el funcionamiento, la forma y sistematicidad en que han sido utilizadas para caracterizaciones anteriores y los resultados obtenidos en dichos procesos; de todas ellas la más adecuada, por adaptarse a las necesidades del estudio de la Web de la UCI, es el *WIRE* Crawler.

Esta herramienta permite realizar un estudio cuantitativo de la Web, generando estadísticas importantes y una serie de reportes que son el pilar para confeccionar un estudio o consideraciones de carácter cualitativo sobre la Web analizada. Ha sido ampliamente utilizada para el estudio no sólo de sedes web, sino en estudios de carácter nacional; incluyendo otros países además de Chile, que fue donde se creó la aplicación. Esta es una característica de mucho peso en la selección, aunque se pueden mencionar las siguientes: está diseñado para trabajar con grandes volúmenes de información; está programado en C++ para un alto rendimiento y es altamente configurable pues todos los parámetros para el rastreo, la indexación, el análisis de los datos y la creación de los reportes estadísticos, se pueden configurar a través de un archivo *XML*.

Por otro lado, Incluye varias herramientas para analizar, extraer estadísticas y generar una serie de informes sobre subconjuntos de la Web. Con el *WIRE* se pueden descargar y trabajar sobre cientos de miles de documentos, que con las otras herramientas no se lograría de manera eficiente. La Web de la Universidad de las Ciencias Informáticas es de un tamaño considerable, aunque es más pequeña que la Web de un país, por ejemplo, como Chile.

La aplicación genera 6 informes sobre la colección:

1. Informe General sobre la colección descargada.
2. Informe sobre las extensiones encontradas en el recorrido.
3. Informe sobre los sitios web analizados en el recorrido.
4. Informe sobre los enlaces de los sitios web analizados en el recorrido.
5. Informe sobre los idiomas o lenguas encontrados en el recorrido.
6. Informe sobre los distintos ciclos realizados por el Módulo de Recolección de Datos.

No se selecciona alguna de las otras herramientas debido a las inconveniencias que presentan para realizar el estudio de la Web de la UCI. Por ejemplo, el ***SocSciBot 4*** se ejecuta en Windows 95, sistema operativo de carácter privativo, y al estar impulsando el uso del Software Libre en la Facultad 10 y en la Universidad, se convierte en una opción que no se debe utilizar; además de que no existe garantía de que funcione correctamente, ni de que los resultados obtenidos de la Web sean los esperados. El ***Webvac Spider*** no se escoge, pues sus posibilidades a la hora de configurar están limitadas, al no permitir cambios necesarios en su configuración que permitan adecuarlo a las necesidades que se requieran para el estudio de cada web. El ***UbiCrawler*** se desecha debido, principalmente, a que para su uso es necesaria la autorización de sus creadores. El ***WebBot*** no se escogió, pues no hace un buen uso de las peticiones GET y HEAD las cuales arrojan resultados necesarios para la investigación de una Web, que son más aprovechadas por las demás alternativas. ***SacarinoBot y EloisaBot Tools*** se encuentran aún en desarrollo y existen posibilidades de que se conviertan en softwares privativos, una vez concluidos y en funcionamiento. Por último el ***Harvest-NG*** no se selecciona debido a lo trabajoso que resulta su configuración para una Web determinada.

2.5 Conclusiones del capítulo.

En el presente capítulo se realiza un análisis de los distintos indicadores webmétricos utilizados para realizar caracterizaciones de la Web. Se seleccionan los indicadores para realizar el estudio de la Web de la UCI, así como la herramienta cibermétrica que se adapta más a las necesidades del mismo. Partiendo del análisis de las distintas alternativas libres para el estudio del ciberespacio, se valoran las potencialidades ofrecidas por cada una y se selecciona la mejor para satisfacer las necesidades y requerimientos para realizar un estudio de la Web Universitaria.

Posteriormente se realiza un estudio más completo de dicha herramienta, sobre cuestiones de instalación, configuración y uso de la misma; que sirve de apoyo para otras personas que decidan utilizar el *WIRE* en trabajos futuros.

CAPÍTULO III – Características de la Web de la Universidad de las Ciencias Informáticas.

En el presente capítulo se muestran los principales resultados de los Estudios Webmétricos realizados en la Universidad de las Ciencias Informáticas (UCI), que permitirán analizar datos exhaustivos de los sitios web de dicho centro:

- Primer Estudio Webmétrico, diciembre de 2008 [25].
- Segundo Estudio Webmétrico, febrero de 2009 [26].
- Tercer Estudio Webmétrico, abril de 2009 [27].

Se ofrecen además varios datos comparativos entre dichos estudios webmétricos, lo cual constituye la base fundamental para medir la evolución de la Web durante los últimos meses, tomando como base de partida los estudios webmétricos realizados con anterioridad. El último estudio presentado fue realizado entre el 1 de abril de 2009 y el 8 de abril de 2009 que, al igual que los dos anteriores, fue desarrollado por el Proyecto Generador de Estudios Webmétricos (GEWEB) del Grupo de Proyectos de Cibermetría Aplicada (CIBA), perteneciente al Polo Productivo de Soluciones Informáticas para Internet (SINI). Se utilizó en todos los casos como Spider, el sistema WIRE, desarrollado en el Centro de Investigación de la Web (CIW) de Chile.

A continuación se muestra una tabla resumen de los recorridos realizados en la UCI:

Variable\Indicador	Primer Estudio Webmétrico	Segundo Estudio Webmétrico	Tercer Estudio Webmétrico
Total de Páginas Analizadas	127718	458457	461831
Texto Total de la Colección	3641.76 MB	8834.68 MB	8307.75 MB
Texto Promedio por Página	0.029 MB	0.019 MB	0.018 MB
Total de sitios web	108	164	159
Páginas Promedio por Sitio	1356	3336	4414
Texto Promedio por Sitio	33.72 MB	53.87 MB	52.25 MB

Tabla 1: Resumen de Tabla Comparativa.

3.1 Nivel Colección. Datos Generales.

La Web de la UCI está compuesta por aproximadamente 160 sitios que contienen más de 700000 páginas, aunque sólo se descargaron poco más de 461800.

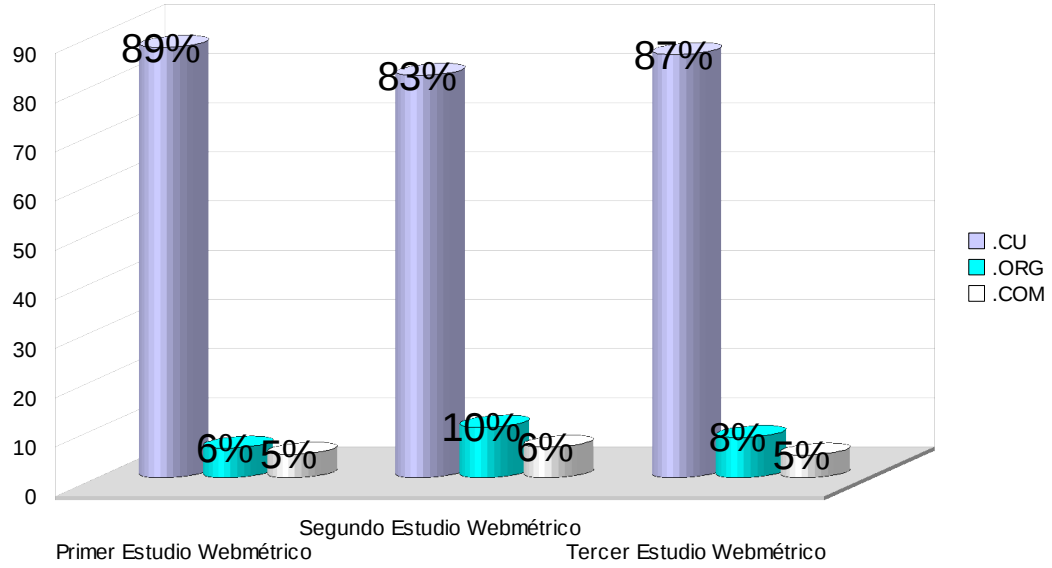
La Tabla 2 muestra algunas distribuciones interesantes de las páginas web descargadas. Estos datos se explican más adelante, en el nivel correspondiente a las páginas.

Total de Páginas Descargadas	**461 831**	
Páginas Únicas	443 664	96.07%
Páginas Duplicadas	18 167	3.93%
Páginas Estáticas	201 668	43.67%
Páginas Dinámicas	260 163	56.33%

Tabla 2: Resumen acerca de las páginas descargadas.

3.1.1 Enlaces a dominios externos.

A partir de los estudios realizados, se localizaron un total de 125 dominios externos al .uci.cu, entre los que predominan (por este orden): .cu, .org, y .com con 4709462, 446744 y 251308 enlaces en nuestra web respectivamente. Esto representa en ese mismo orden, un 85.98, 8.16 y 4.59 porciento del total de enlaces a dominios externos encontrados en la colección, incluyendo tanto enlaces a páginas web como a ficheros multimediales u otras extensiones. (Ver Dibujo 1)



Dibujo 1: Dominios Externos más Referenciados en la UCI.

Estos altos números de enlaces a direcciones fuera de la Web de la UCI tienen un impacto positivo en los usuarios, pues significa que los mismos tienen acceso a través de las redes del dominio .uci.cu a múltiple información de sitios tanto nacionales como internacionales; lo cual tiene una incidencia directa en la docencia y los procesos investigativos/productivos de la universidad. La Tabla 3 muestra los 10 Dominios Externos más referenciados en la UCI.

DOMINIO	Cantidad de Enlaces	Porciento
CU - CUBA	4 709 462	85.98
ORG	446 744	8.16
COM	251 308	4.59
NET	42 943	0.78
GOV - Gobierno	11 305	0.21
ES - España	3 875	0.07
UK - Ucrania	1 618	0.03
DE - Germany	1 373	0.03
RU - Federación Rusa	1 151	0.02
EDU - Educación	1 135	0.02

Tabla 3: Dominios Externos más referenciados en la UCI.

3.1.2 Software utilizado como Servidor Web. Sistemas Operativos.

Luego de terminar el estudio, se realiza una búsqueda DNS de la dirección IP de cada uno de los sitios identificados; para ello se realiza una petición HEAD con parámetros específicos de respuesta, donde además se obtienen los datos acerca del software utilizado como servidor web, las distintas extensiones instaladas e incluso, en ocasiones, el sistema operativo. Una respuesta típica tiene la siguiente forma:

HTTP/1.1 200 OK

Date: Tue, 14 Apr 2009 00:24:22 GMT

Server: Apache/2.2.3 (Debian)

Content-Type: text/html; charset=utf-8

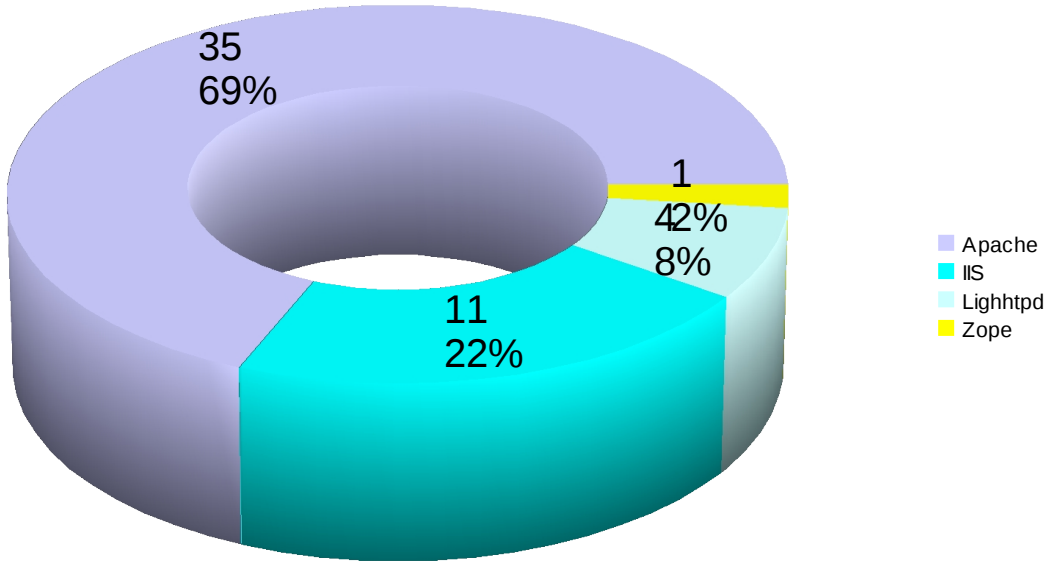
Last-Modified: Tue, 14 Apr 2009 00:24:22 GMT

Client-Date: Tue, 14 Apr 2009 00:24:49 GMT

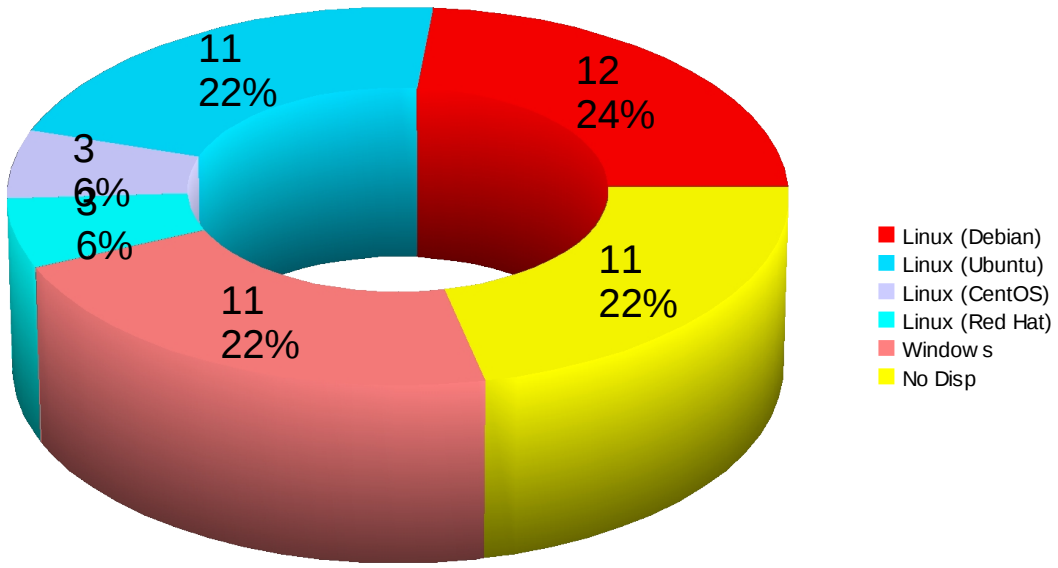
Client-Peer: 10.0.0.13:80

X-Powered-By: PHP/5.2.5

En algunos casos –como en el ejemplo– la información es bastante completa, incluyendo el nombre del servidor (Apache), la versión del software instalada (2.2.3), el sistema operativo utilizado (Debian) y las extensiones instaladas (PHP/5.2.5); no obstante, en otros casos, no siempre la respuesta contiene todos los datos necesarios. En los Dibujos 2 y 3 se puede apreciar la distribución del tipo de servidor web y de sistema operativo por dirección IP.



Dibujo 2: Distribución del Tipo de Servidor Web por dirección IP.



Dibujo 3: Distribución del Tipo de Sistema Operativo por dirección IP.

El Servidor Web dominante es Apache, usado en más del 68% de las direcciones IP encontradas, esto triplica el uso de Microsoft IIS (Internet Information Server), que sólo es utilizado en un 21% de las mismas. Se puede decir con certeza que Apache tiene el mayor porciento de participación en la Web de la UCI. Unas 4 direcciones de IP, lo cual representa el 8%, todas pertenecientes a la Facultad Regional de Granma, utilizan el Servidor Web Lighttpd. En la Web mundial la proporción de servidores es 69% para Apache y 21% para IIS; lo cual se corresponde con los resultados obtenidos. También se encontró el servidor de aplicaciones web Zope.

En el caso de los sistemas operativos utilizados, de igual manera los resultados son bastante evidentes. De un total de 51 direcciones de IP conocidas, 29 utilizan GNU/Linux (12 utilizan Debian, 11 Ubuntu, 3 Red Hat y 3 CentOS), lo cual corresponde a más del 57%. Sólo el 21.5% de ellas utilizan Windows, equivalente a 11 direcciones IP. Igual número de direcciones IP, no devuelven información acerca del sistema operativo que utiliza. Si los casos no determinados se distribuyeran proporcionalmente de acuerdo a los casos conocidos, ya no existiría para Windows la brecha con los sistemas operativos de código abierto en nuestra web. Los resultados evidencian claramente el auge del Software Libre en la Universidad de las Ciencias Informáticas.

3.1.3 sitios web por dirección IP.

Es muy común que en una misma dirección IP se encuentren múltiples sitios web, soportados con uno o más servidores de aplicaciones de este tipo. En el Anexo 4 se muestra la cantidad de sitios encontrados por cada dirección de IP identificada durante los estudios.

De manera general se puede decir que existen 199 sitios web con dirección IP conocida y que, a su vez, se identificaron de estas últimas unas 51 direcciones IP; o sea, que hay un promedio aproximado de 3.9 sitios web por cada dirección IP conocida. La mayor cantidad de sitios web en una misma dirección IP es 52. Por otro lado, se encontraron 37 direcciones IP que sólo cuentan con un sitio web.

3.2 Nivel Sitio. Datos Generales.

En esta sección se presentan los resultados más generales del nivel mencionado; tales como, la cantidad promedio de páginas web por sitio, así como datos referentes a la profundidad y peso en términos de contenido de los mismos.

Total de Sitios OK	**159**
Promedio de Enlaces Internos	48 950
Promedio de Páginas Web por Sitio	4414
Promedio de Páginas Dinámicas	2514
Promedio de Páginas Estáticas	1900
Tamaño promedio en MB	52.25
Promedio de Profundidad Máxima	5,67
Promedio de Grado Interno	2,66
Promedio de Grado Externo	2,66

Tabla 4: Resumen acerca de los sitios analizados.

3.2.1 Cantidad promedio de páginas por sitio.

Se podría inferir que un sitio promedio de la colección tiene aproximadamente 4414 páginas web, contenidas en más de 50 MB, con 2.66 referencias desde otros sitios de la colección. Con un promedio de páginas estáticas y dinámicas por sitio de 1900 y 2514 respectivamente, para un 43 y un 57 porciento del total de páginas. Su profundidad promedio es de 5.67. En la Tabla 4 se pueden apreciar estos y otros datos relacionados.

Según estos datos se puede apreciar claramente una mayoría de páginas dinámicas respecto a las páginas estáticas en la Web de la UCI. Lo cual refleja, a su vez, un alto nivel de desarrollo; pues brinda una idea de cuánto se ha avanzado desde la Web 1.0 (Estática) hacia la Web 1.5

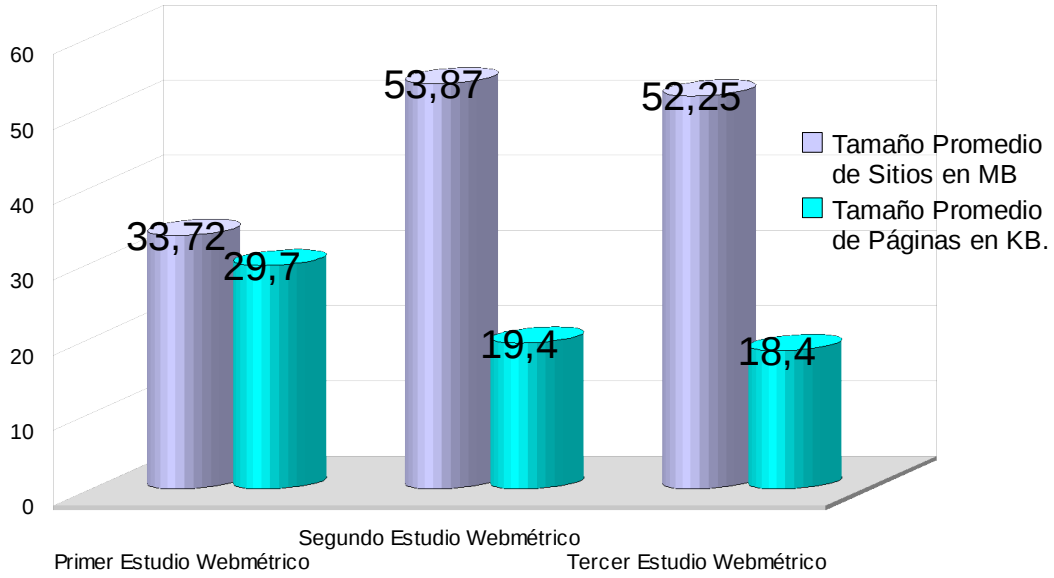
(Dinámica) o, incluso, pensando de manera futurista hacia la Web 2.0, dirección en la cual se trabaja actualmente en la Universidad.

3.2.2 Tamaño total de la colección de información analizada.

Sólo se descargó el texto plano de las páginas web analizadas, lo que representa más de 8 GB de información. Al agregar a esta cifra los metadatos descargados, los registros de enlaces analizados y descargados de la Web y los ficheros generados para el análisis de la colección de información y los respectivos reportes; la colección alcanza un tamaño total de más de 16 GB. Esta cifra es mayor, en todos los sentidos, a las colecciones de los dos primeros estudios; cuyos valores alcanzan unos 5 y 7 GB respectivamente, debido principalmente a los valores utilizados en las variables del fichero de configuración del Spider.

3.2.3 Tamaño promedio de los sitios en MB.

Como se presentó anteriormente, un sitio promedio de la UCI está contenido aproximadamente en 52.25 MB (Ver Tabla 1); esta es una cifra que, aunque disminuyó respecto al 53.87 MB del segundo estudio, la variación es pequeña. Como es lógico, al aumentar además la cantidad de páginas descargadas y mantenerse tan parecido el tamaño promedio de los sitios, el tamaño promedio de las páginas web disminuyó considerablemente llegando hasta 0.018 MB, lo cual es equivalente a aproximadamente 18.4 KB. O sea, que se puede afirmar que una página promedio de la UCI tiene poco más de 18 KB. (Ver Dibujo 4)



Dibujo 4: Tamaño Promedio de sitios y páginas web.

3.2.4 Profundidad máxima promedio de los sitios.

A partir de este valor se puede conocer para el usuario, el trabajo exigido para navegar por los sitios de la colección analizada. En la Tabla 4 se puede observar que el promedio de profundidad máxima de la colección analizada es 5.67. Este valor es muy superior respecto a los obtenidos anteriormente, que fueron de 4.31 para el primer estudio y 4.38 para el segundo; lo cual se corresponde con el aumento del nivel de profundidad utilizado por el Spider para recorrer la Web, que pasó desde 15 en el primer estudio, hasta 25 en el segundo y 50 en el tercero.

3.2.5 Promedio de Grado Interno.

Los sitios web más referenciados se muestran en la Tabla 5. Se cuentan todos los sitios distintos de la colección que apuntan a un sitio específico. Este indicador está generalmente asociado a la popularidad de un sitio y se denomina grado interno. De manera general se podría inferir que un sitio promedio es referenciado aproximadamente desde 2.66 sitios web pertenecientes a la colección analizada. (Ver Tabla 4)

Dirección del Sitio	Grado Interno (Enlaces Entrantes)
intranet.uci.cu	35
softwarelibre.uci.cu	24
teleformacion.uci.cu	22
primavera.uci.cu	18
inter-nos.uci.cu	15
web21.uci.cu	10
softwarelibre.hab.uci.cu	10
investigaciones.uci.cu	9
guiatelefonica.uci.cu	9
feu.uci.cu	8

Tabla 5: Sitios con mayor grado interno de la colección.

3.2.6 Promedio de Grado Externo.

También existe lo que se denomina grado externo de un sitio, que no es más que la cantidad de sitios a los que hace referencia un sitio específico. Los sitios web de mayor grado externo se pueden apreciar en la Tabla 6 y son aquellos desde los cuáles se puede acceder directamente a la mayor cantidad de sitios web de la colección. Conocer esta información puede ayudar a conocer el nivel de comunicación que existe entre los sitios web de la colección y, de cierto modo, incluso se puede analizar en ocasiones el tipo de sitio web al que se accede. Por ejemplo, un sitio de carácter comercial tratará de tener un grado externo pequeño, para evitar que los usuarios abandonen el sitio siguiendo enlaces a otros.

Dirección del Sitio	Grado Externo (Enlaces Salientes)
intranet.uci.cu	45
foro.uci.cu	31
ucipedia.uci.cu	27
web21.uci.cu	22
primavera.uci.cu	21
feu.uci.cu	20
softwarelibre.uci.cu	13
java.uci.cu	12
softwarelibre.hab.uci.cu	11
intranet.hab.uci.cu	11

Tabla 6: Sitios con mayor grado externo de la colección.

Un 31.45% de los sitios están conectados entre sí a través de enlaces. Por otro lado, el 67.3% de los sitios está completamente desconectado en términos de enlaces (Islas de Información). De los 159 sitios analizados, 53 no son referenciados por ningún otro sitio de la colección, así mismo 77 no referencian a ningún otro sitio de la colección.

3.2.7 Distribución de páginas web por sitio.

En esta sección se muestran los sitios de la colección que presentan las mayores cantidades de páginas web descargadas, entre los que se encuentran el Portal del Proyecto ERP-CUBA, el Sitio de Gestión de Proyectos de la Facultad 10, el Foro de las Comunidades de Desarrollo de la UCI y los distintos sitios y repositorios de Software Libre, entre otros. Presumiblemente, se puede inferir que estos sitios son además aquellos que lógicamente poseen la mayor parte del contenido analizado en el estudio. (Ver Tabla 7)

Además, suele suceder que los sitios con mayor cantidad de páginas web también están formados, en su mayoría, por páginas dinámicas de contenido y tienden a crecer exponencialmente a medida que pasa el tiempo. Normalmente corresponden a sitios que tienen instalados uno o más Sistemas de Gestión de Contenidos (CMS, por sus siglas en inglés) para brindar servicios de blogs, foros o galería de imágenes. Los CMS actuales permiten utilizar *URL Rewriting* para recuperar las páginas y una serie distinta de parámetros puede llevar al mismo documento. Además agregan distintos enlaces a distintas partes internas del documento (como los comentarios a una entrada en un blog o las distintas opiniones en un foro), los que crean recursión en las páginas. Estos sistemas no tienen diseños estéticos (por ejemplo, una vista de un documento entregando un identificador muestra enlaces a otras páginas del sitio que verlo entregando la fecha del mensaje como parámetro no se muestran) por lo que es difícil detectar los documentos duplicados.

Dirección del Sitio	Cantidad de Documentos
portal-erp.prod.uci.cu	120 874
gforge.f10.uci.cu	75 854
foro.uci.cu	75 000
gentoo.prod.uci.cu	56 659
ubuntu.prod.uci.cu	48 330
mirror.prod.uci.cu	39 935
softwarelibre.uci.cu	38 385
gpi.uci.cu	35 716
cpav.uci.cu	35 122
isos-linux.prod.uci.cu	22 778
ucipedia.uci.cu	18 465
informatizacion.uci.cu	17 278
softwarelibre.hab.uci.cu	15 082
documentacion.prod.hab.uci.cu	14 887
primavera.uci.cu	14 855
php.uci.cu	9 624
ujc.uci.cu	5 925
seriecientifica.uci.cu	5 904
invsoftedu.uci.cu	5 891
opensuse.prod.uci.cu	4 829

Tabla 7: Sitios con mayor cantidad de documentos de la colección.

Por otro lado, alrededor de 84 sitios web tienen sólo una página web. Esto puede significar un consumo innecesario de recursos en un sitio que prácticamente no brinda funcionalidades al usuario; puesto que se está utilizando un ordenador para publicar un sitio sin contenido que mostrar a los usuarios, lo cual por otro lado tiene un gran impacto negativo en el visitante del sitio web en cuestión. Aunque no siempre es así, en ocasiones puede estar relacionado con otros hechos que impiden que el *Spider* los analice.

En algunos casos puede tener otras razones como:

- ➔ La navegación de la página está basada en JavaScript, por lo que es necesario interpretar esta técnología para poder navegar.
- ➔ La página es sólo una redirección a otro sitio, tanto usando la etiqueta "*Refresh*" como teniendo únicamente un enlace comunicando al usuario la dirección del otro sitio.
- ➔ La página, en efecto, es la única página del sitio.

- ➔ La página requiere un plug-in de Flash para poder ser visualizada. Es una tendencia entre sitios web el tener una introducción animada al sitio, sin usar verdaderamente Flash para mostrar el contenido u organizar la página. De este modo muchos de estos sitios, a pesar de ser “normales”, no logran ser indizados por los buscadores por no incluir un enlace del tipo “Omitir Introducción”.
- ➔ La página contiene solamente enlaces externos.
- ➔ La página efectivamente tiene enlaces internos, pero estos están incorrectamente formados y el recolector no pudo interpretarlos.
- ➔ La página utiliza Applets Java para la navegación.

3.3 Nivel Página. Datos Generales.

En esta sección se muestran una serie de datos acerca de las páginas descargadas durante el estudio. Sin tener en cuenta su agrupación en los distintos sitios analizados, sino más bien desde un punto de vista independiente de las mismas. Se muestran distribuciones de profundidad, edad, idiomas, etc.

3.3.1 Cantidad de páginas únicas/duplicadas de la colección.

Del total de páginas web descargadas, un 96,07% son páginas únicas. Lo que equivale a más de 443600 páginas web, de 461831 que se descargaron. Generalmente el índice de las páginas duplicadas suele ser mayor, como sucedió con el primer estudio y su 7.6% de páginas duplicadas, o el segundo con un 5.1%. Esta vez, se redujo hasta un 3.93%. (Ver Tabla 2)

Estos datos pueden ser interpretados de distintas maneras. Cuando la información duplicada se refiere a datos de gran importancia, que requieren ser mostrados de manera íntegra en distintas ubicaciones debido al alto valor que tiene la misma para los usuarios, se puede tomar como positivo. En este caso refleja un grado aceptable de replicación de la información importante en la Web. Por otra parte, si la influencia de una información determinada tiene una incidencia negativa en los usuarios, la replicación de la misma en la Web no puede ser de ninguna manera positiva.

3.3.2 Cantidad de páginas dinámicas/estáticas de la colección.

Se obtuvo que 260163 páginas web son dinámicas, lo que respresenta un 56.33% del total de la colección de información. El resto, correspondiente a 201668, son lógicamente páginas web estáticas. Estos datos están en correspondencia directa a la distribución promedio de páginas web dinámicas y estáticas por sitio. (Ver Tabla 2)

Es positivo el predominio de las páginas web dinámicas, pues con las mismas es mucho mayor la interacción del usuario y el sitio web; permitiendo que exista un incremento en la aceptación por parte de los usuarios hacia las mismas, al incluir novedosas funcionalidades para las que son necesarias la utilización de otros lenguajes de programación, como la programación especial en PHP y ASP. Así como la utilización de JavaScript, AJAX y otras tecnologías.

Este aumento, en los últimos años, viene dado principalmente por la utilización de distintos CMS en la creación de los sitios web, debido al incremento de los conocimientos por parte de los desarrolladores de sitios web de la UCI y, por ende, a la utilización por los mismos de herramientas y técnicas avanzadas en este sentido. Algunos de los principales *CMS* utilizados son, por ejemplo: PhpFusion, PhpNuke, Drupal, E107 y Joomla. También es muy utilizado Zope Plone en la creación de sitios web.

Se asume que el número de páginas dinámicas seguirá creciendo en el futuro, debido a la tendencia actual de tener sitios cuyos contenidos se puedan administrar en línea y que sean independiente del diseño y de la estructura de los documentos, al ser más fácil y práctico tener el contenido del sitio en una base de datos que en archivos *HTML*, donde se hace mucho más difícil modificarlos para ingresar o modificar alguna información. Además, es válido mencionar que existen muchas páginas estáticas, con extensiones *HTML*, que son generadas por procesos en lote en los servidores, que se ejecutan constante y automáticamente.

3.3.3 Profundidad de las páginas de la colección.

Partiendo de que la profundidad lógica de una página web no es más que la cantidad mínima de veces que el usuario tiene que dar *click* en un vínculo, para llegar a la misma sin abandonar el sitio web y comenzando desde la portada del mismo, este indicador brinda información muy importante y útil. Durante el desarrollo actual de las tecnologías de la información el tráfico de usuarios por las redes ha aumentado considerablemente y, con esto, también han aumentado las necesidades de los mismos y la búsqueda de comodidades en la navegación. Con este indicador se puede establecer una media del trabajo que representa para el usuario llegar a una información determinada en la Web. Las páginas relevantes para el usuario no deben encontrarse a grandes profundidades y, al mismo tiempo, la media de acceso no debe tener valores muy elevados; pues sucede que el usuario se aburre de dar *click* y sencillamente abandona el sitio web.

Unas 450489 páginas web se encuentran entre las profundidades 1 y 10, representando un 97.5% de la colección analizada (461831 páginas web). El resto, equivalente a unas 11342 páginas web, está entre las profundidades 11 y 35 y, aproximadamente, una mitad de esta cifra se ubica por encima de dicha profundidad. A medida que aumenta la profundidad se puede

observar que va disminuyendo la cantidad de páginas web, aunque esta variación es poco significativa tanto en el porciento respecto al total de páginas analizados como en la propia cantidad de páginas web.

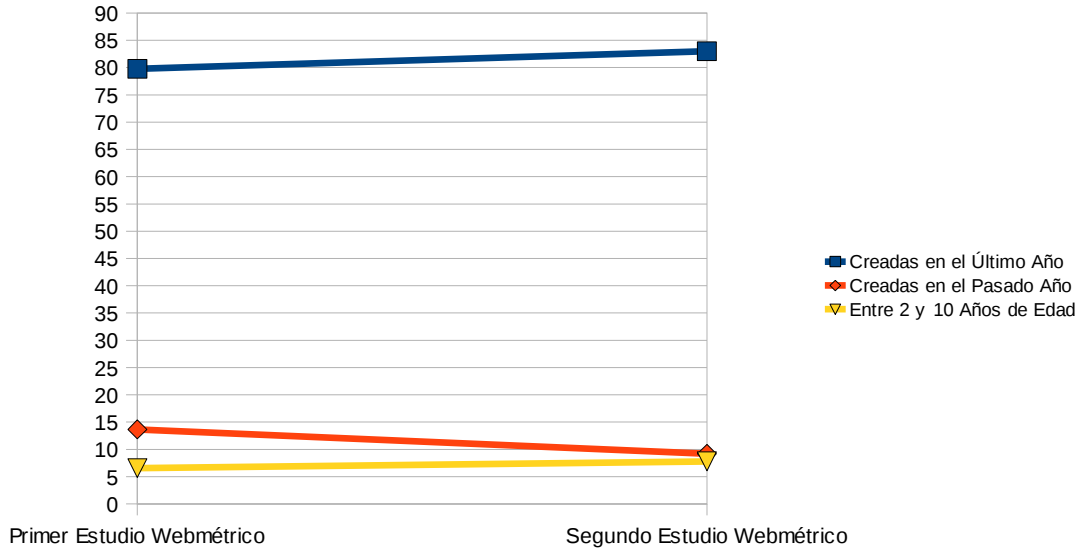
Por otra parte, considerando que la profundidad lógica de una página está dada básicamente por la "cantidad de clicks que el usuario debe dar en un sitio para llegar a ella", es preocupante saber que existen más de 5600 páginas web por encima de la profundidad 35; aunque en muchos de los casos se debe a la presencia de algunas *URL* con múltiples elementos repetidos, lo cual además viola ciertos estándares para el largo de las mismas [28]. Ejemplo de esto son las siguientes *URL* mostradas (recortadas por motivos de legibilidad, y que varían de longitud durante el recorrido):

- http://softwarelibre.uci.cu/author/resolveUid/resolveUid/[...]/resolveUid/resolveUid/...
- http://primavera.uci.cu/news/link/load/link/load/[...]/link/load/link/load//news/...
- http://ubuntu.prod.uci.cu/Ubuntu/releases/releases/[...]/releases/releases/...

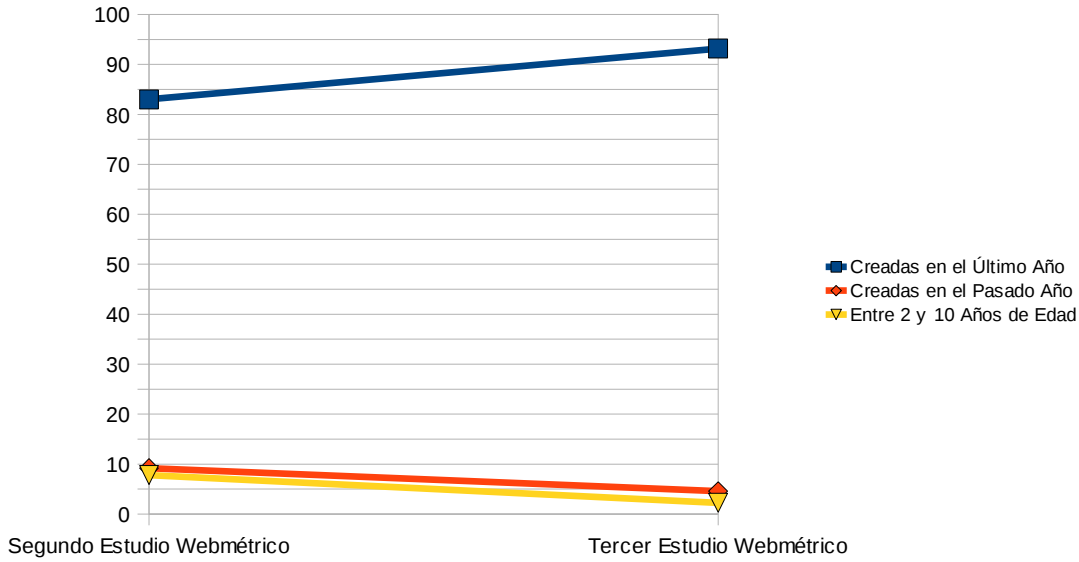
3.3.4 Edad de las páginas de la colección.

La Web de la UCI ha ido creciendo aceleradamente en los últimos años. Prueba de ello es que un 93.17% de las páginas web analizadas fueron creadas o modificadas en el último año, así mismo un 4.58% en el año anterior; sólo un 2.25% de las páginas web no han sido actualizadas en los últimos dos años. Estos datos evidencian un alto grado de crecimiento y dinamismo en la colección analizada. En los Dibujos 5 y 6 se muestran datos comparativos respecto al primer y segundo estudio de la Web de la UCI.

Para determinar la edad de las páginas, el spider observa la última fecha de modificación (*Last-Modified Date*) entregada por el servidor en la petición HEAD realizada. Se pueden dar casos de fechas incorrectas, debido a que el servidor no tiene sus relojes sincronizados con la hora y fecha actual del país o que simplemente no han sido configurados para ello. Incluso, un servidor web pudiese delvolver una fecha del futuro o alguna tan antigua como la misma creación de la Web en el mundo.



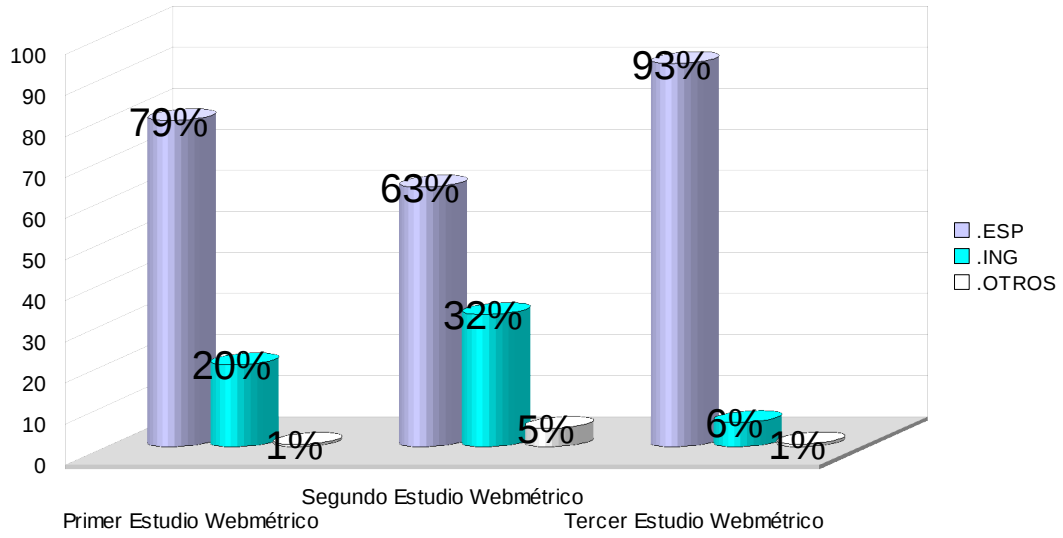
Dibujo 5. Primer y Segundo Estudio Webmétrico.



Dibujo 6. Segundo y Tercer Estudio Webmétrico.

3.3.5 Idioma de las páginas de la colección.

En la Web de la UCI predominan los idiomas español e inglés, presentes en aproximadamente un 93.33% y 5.96% respectivamente. También se encuentran pequeñas muestras de danés, francés e italiano. El aumento del porciento de páginas en español y la consecuente disminución del porciento de páginas en inglés se debe al mayor número de páginas descargadas para sitios con gran cantidad de contenido en el idioma español, tales como el Portal del Proyecto ERP-CUBA, entre otros. (Ver Dibujo 7)



Dibujo 7: Distribución aproximada de idiomas en la Web de la UCI.

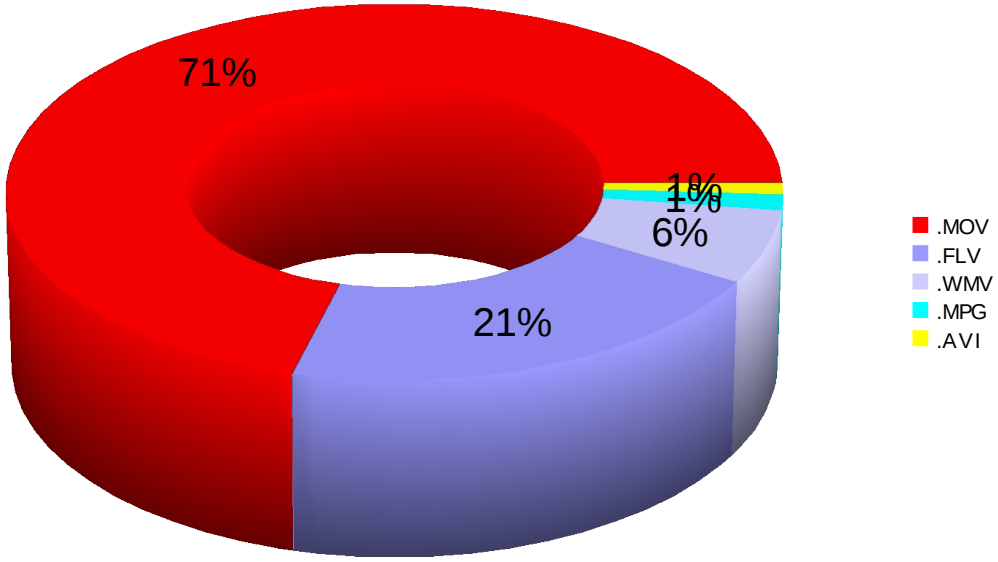
3.3.6 Extensiones encontradas durante el estudio.

Con este indicador se puede conocer las extensiones que predominan en la Web, según cada una de las siguientes clasificaciones y de manera general. Además ayuda a identificar la existencia de extensiones desconocidas y el nivel en que están presentes las mismas. Se pueden establecer medias del uso de extensiones de Software Libre por estos datos, que dan una idea del nivel de impacto y utilización del Software Libre en la Universidad.

3.3.6.1 Extensiones de Audio, Video e Imagen.

En la Web de la UCI predominan las extensiones de video .MOV con 3676 ficheros para un 70.9% del total de extensiones de video encontradas, que indica un 20% más de presencia en la Web que en el segundo estudio. Se encontró además 1063 ficheros .FLV (20.5%) y 322 ficheros .WMV (6.21%). Otras extensiones como .MPG y .AVI sólo tienen poco más de 50 ficheros. (Ver Dibujo 8)

La extensión .MOV se refiere al formato común para las películas Quick Time, la plataforma nativa de Macintosh para películas. Es una extensión de ficheros y un excelente formato de video desarrollado por Apple Computer para videos o animaciones comprimidas. En la actualidad es un formato de video muy utilizado, principalmente, en la realización de las presentaciones (trailers) de películas.



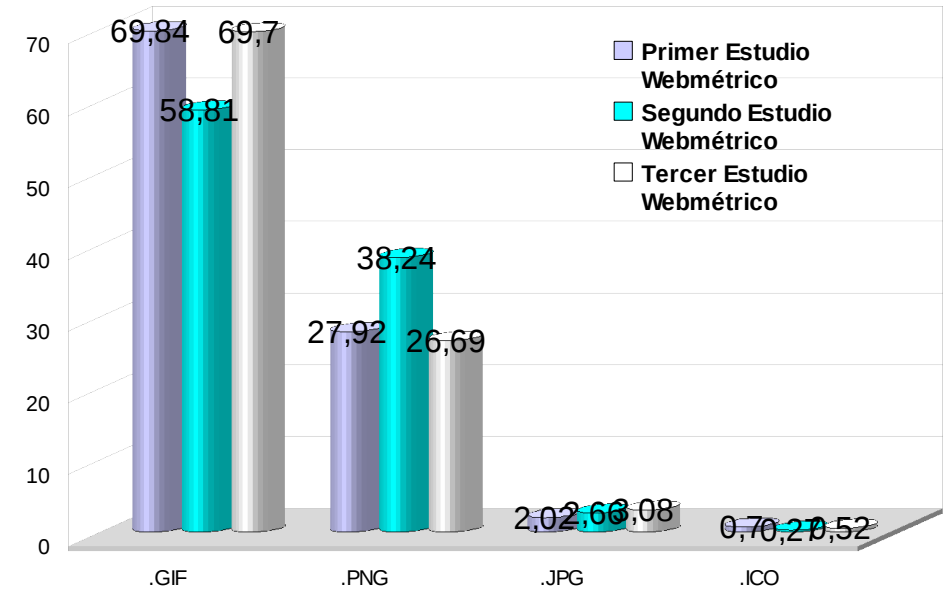
Dibujo 8: Distribución de extensiones de video (% de presencia en la Web).

La extensión de audio que predomina es .MP3 (68.52% de presencia en la Web), aunque en el segundo estudio con 10 ficheros menos tenía un 11% más de presencia en la Web. Esto indica que otras extensiones de audio han ido aumentando su cantidad últimamente. Se aprecia la aparición de extensiones .DJVU con 18 ficheros y el aumento de las extensiones .OGG hasta 15 (ficheros contenedores que pueden tener tanto audio como video). Las extensiones .RAM (11 ficheros), .MID (2 ficheros) y .WMA (1 fichero) se mantienen invariables, aunque sí disminuye su porciento de presencia en la Web respecto al total de extensiones de audio encontradas.

El predominio de la extensión .MP3 se debe probablemente al auge de los reproductores portátiles, este es el estándar utilizado para oír un archivo directamente en una página web, sin necesidad de descargarlo antes al ordenador y poseer una compresión de audio de alta calidad gracias a la posibilidad de ajustar la calidad de la compresión y, por tanto, el tamaño final del archivo que podía llegar a ocupar 12 e incluso 15 veces menos que el archivo original sin comprimir.

Desde el primer estudio webmétrico en la UCI, las extensiones de imágenes que predominan siempre han sido las mismas, aunque como es lógico han ido aumentando considerablemente la cantidad de ficheros encontrados por cada una de ellas. De manera estable la extensión de imagen .GIF siempre ha mantenido el mayor nivel de presencia en la Web; esta vez con más de 5000000 de ficheros para casi un 70% del total encontrado, aumentando en un 11% respecto al segundo estudio realizado. Le siguen en este mismo orden las extensiones .PNG, .JPG, y .ICO, con más de 1929000, 222700 y 37600 ficheros respectivamente por cada una, en el presente

estudio. Estas extensiones han mantenido un porciento de presencia en la Web bastante parecido en todos los estudios y en el mismo orden presentado para esta ocasión. (Ver Dibujo 9)



Dibujo 9: Distribución de extensiones de imagen (% de presencia en la Web).

El alto porciento de la extensión .GIF se debe a que la misma es muy usada a la hora de diseñar una página, debido a que presenta compresión sin pérdida y que pueden incluir animación. Esta extensión es independiente de plataforma, lo que posibilita su utilización en cualquier ordenador independientemente del Sistema Operativo utilizado, siempre y cuando la misma posea un visor de imágenes. Es un formato de imágenes diseñado para minimizar el tiempo de transferencia de archivos sobre la red. Por otro lado, la presencia de la extensión PNG se debe a que es el formato de archivo nativo de Macromedia Fireworks. Los archivos PNG conservan la información original de capa, vector, color y efectos (como por ejemplo las sombras) y todos los elementos pueden editarse siempre que se desee.

3.3.6.2 Extensiones de interfaz de entrada común y código fuente.

La extensión de interfaz de entrada común (CGI) que predomina en la Web es .PHP con un 99,61% de presencia en la misma. Se encontró, para el estudio actual, más de 7200000 ficheros con esta extensión en la Web de la UCI. Otras extensiones de este tipo como .DLL, .PL y .ASP solo ocupan un 0.13%, 0.12% y 0.11% respectivamente, aunque con más de 8200 ficheros en todos los casos, llegando incluso a 9500 ficheros para la primera extensión mencionada. Los distintos estudios realizados han mostrado siempre la insuperable cantidad de ficheros .PHP en la Web de la UCI, con respecto a las otras extensiones identificadas de este tipo. PHP es un

lenguaje interpretado de alto nivel embebido en páginas HTML y ejecutado en el servidor. Sus altas potencialidades justifican la utilización que se le da actualmente en la universidad.

El lenguaje en cuestión, al nivel más básico, puede hacer cualquier cosa que se pueda hacer con un script CGI, como procesar la información de formularios, generar páginas con contenidos dinámicos, o mandar y recibir *cookies*. Quizás la característica más potente y destacable de PHP es su soporte para una gran cantidad de bases de datos. Escribir un interfaz vía web para una base de datos es una tarea simple con PHP.

Ejemplo, es que las siguientes bases de datos son soportadas actualmente por PHP:

Adabas D, Ingres, Oracle (OCI7 y OCI8), dBase, InterBase, PostgreSQL, Empress, FrontBase, Solid, FilePro, mSQL, MySQL, Sybase, IBM DB2, Velocis, Informix, ODBC y Unix dbm.

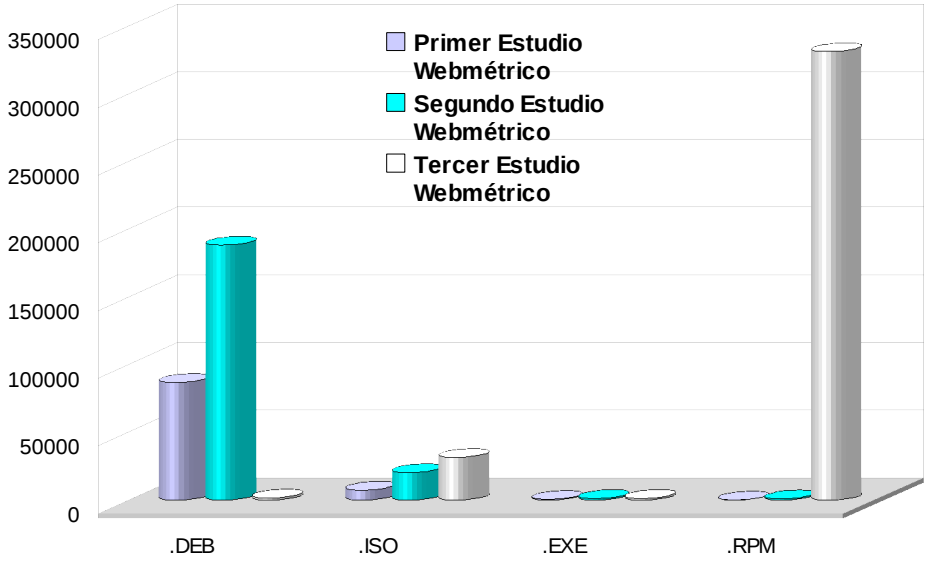
Por otro lado, PHP también soporta el uso de otros servicios que usen protocolos como *IMAP, SNMP, NNTP, POP3, HTTP* y derivados. También se pueden abrir sockets de red directos (raw sockets) e interactuar con otros protocolos. Soporta la programación orientada a objetos (*POO*).

En la Web se encuentran muestras de extensiones de código fuente en distintos lenguajes, tales como: .JAVA con 312 ficheros y .BAT con 77 ficheros. Por otro lado, se hallaron además 37 ficheros .SH (170 menos que en el segundo estudio), 23 ficheros .C (292 menos que en el segundo estudio) y 22 ficheros .H (475 menos que en el segundo estudio). La menor presencia de estas extensiones en la Web es comprensible, debido a que en la actualidad la mayoría de los usuarios y desarrolladores distribuyen las aplicaciones compiladas y su código fuente en ficheros compactados.

3.3.6.3 Extensiones de software.

Respecto a las extensiones de software, se aprecian cambios significativos respecto al pasado estudio. La extensión de software que predomina en este caso es .RPM con más de 331400 ficheros; en los dos estudios anteriores, dicha cifra, nunca llegó a 1000. Sin embargo, ocurrió todo lo contrario con la extensión .DEB que, aún presentando más de 188400 ficheros en el pasado estudio, esta vez solo llegó a poco más de 1680.

Estos cambios pueden tener su causa en la disminución de las páginas en los sitios y repositorios de Software Libre. La extensión .ISO, por su parte, aumentó en más de 10000 su cifra para llegar a 31620 ficheros. Otras extensiones presentan variaciones menos significativas. Estos valores suelen ser en realidad mayores, pues muchos softwares desarrollados son distribuidos en ficheros compactados y de tal forma los identifica el Spider (Ej. .ZIP, .RAR, .GZ). (Ver Dibujo 10)



Dibujo 10: Distribución de extensiones de software (% de presencia en la Web).

Los ficheros con extensión .DEB son paquetes de aplicaciones ya preparados para instalarse de una forma sencilla en el Sistema Operativo GNU/Linux. Al ejecutarlos, automáticamente se encargan de instalar el paquete correspondiente y buscar las dependencias de otros paquetes que pudiera necesitar para su correcta instalación. Se pueden instalar tanto con aplicaciones de modo gráfico como utilizando algunas líneas de comando. Por otro lado, el formato RPM es el más utilizado en la actualidad. Tiene como ventaja principal el encargarse de verificar las posibles dependencias o requisitos para la instalación o actualización de un paquete en particular, así como también el verificar si el paquete que se procederá a desinstalar es requerido por otros paquetes presentes en el sistema. Los paquetes de este tipo son programas previamente compilados, almacenados y listos para ser instalados en el sistema. Debido a estas razones es muy común encontrar, en la Web de la UCI, un alto porciento de estos tipos de extensiones de software.

3.3.6.4 Extensiones que no son *HTML* ni *TXT*.

Predominan, entre los documentos que no son *HTML,* las extensiones .PDF con más de 25500 ficheros en la Web. La superioridad de la extensión .PDF, sobre los restantes tipos de documentos, se debe a que el mismo permite realizar cualquier combinación de texto, gráficos, imágenes e incluso música. Al igual que la extensión .GIF es multiplataforma, lo que posibilita su utilización en cualquier ordenador independientemente del sistema operativo utilizado, siempre y cuando la misma posea un visor de documentos PDF. De este modo, permite publicar cualquier información sin el temor de que la misma no pueda ser consultada o simplemente se modifique el

aspecto y la estructura del documento. Además de permitir proteger el contenido mediante un método de cifrado e incluso firmarlo digitalmente.

3.3.6.5 Extensiones de ficheros comprimidos.

Un archivo comprimido es como una caja en la que puedes guardar cualquier tipo de información (imágenes, documentos, música, entre otros), esta información se codifica y se comprime para ocupar un menor espacio, por lo que resulta mucho más fácil transportar los archivos, enviarlos a través de internet o, incluso, almacenarlos. Con el mismo ejemplo de la caja, se puede ver lo que hay en el interior del archivo comprimido (la caja) pero, para usarlo, hay que extraerlo.

Un 88.27% de los archivos compactados en nuestra web tienen extensión .GZ y TAR.GZ; esta superioridad respecto a otras extensiones de ficheros compactados se ha mantenido desde el primer estudio realizado. Aunque es mayor al 74.92% obtenido en el primero, disminuyó en un pequeño porciento respecto al 90.4% del segundo.

Otras extensiones presentes en la Web de la UCI son: .ZIP (6876 ficheros), .BZ2 (3025 ficheros), .RAR (1104 ficheros) y .TAR (1097 ficheros); con un 6.64%, 2.92%, 1.07% y 1.06% de presencia respectivamente.

La extensión .ZIP es un tipo de formato muy usado para comprimir imágenes, videos, pdf, entre otros. Este formato de compresión es el más usado en *Windows,* junto con .RAR. Este último formato de compresión mencionado, en GNU/Linux, apenas se utiliza.

Los archivos .BZ2 son archivos comprimidos con una calidad excelente, aunque también depende del tamaño del archivo comprimido; es importante añadir que al comprimir un archivo el original se pierde, desaparece.

La extensión .GZ sólo comprime ficheros, ya que junto con .TAR y .BZ2 son las extensiones primarias de compresión de ficheros. Esto se resuelve totalmente con .TAR.GZ, el cual es el formato más utilizado; comprime todo tipo de ficheros con una buena calidad, este formato ya es secundario, acoge la extensión .TAR (para carpetas) y .GZ (para ficheros), lo que la hace una potente extensión de compresión.

3.3.6.6 Extensiones extras.

Se encontraron otras extensiones como .CSS con 252751 documentos, lo cual evidencia el uso popularizado en la universidad de esta variante para la configuración estándar de las páginas web. También se encontraron 32383 documentos con extensión .SWF (20000 más que en el estudio anterior), referentes a ficheros de flash. Su uso, con el desarrollo de las tecnologías, ha ido aumentando considerablemente en los últimos años; pues permite crear archivos pequeños,

que funcionan en cualquier plataforma, y que a su vez permiten la interactividad y la trasmisión de gráficos sobre un ancho de banda reducido. De igual manera, se justifica la aparición de 1650 documentos con extensión .js de Java Script este valor, por otro lado, disminuyó en 2000 documentos. También se encontraron 3972 ficheros .GPG (Llaves OpenPGP) y 2818 ficheros .PSD (Proyectos o imágenes de Adobe Photoshop).

3.3.6.7 Extensiones desconocidas.

El Anexo 5 muestra una tabla de las extensiones desconocidas más encontradas durante el estudio. Estas no son más que las extensiones que no fueron previamente categorizadas, durante la configuración del modulo de descarga del *spider*. Es muy probable que varias de las extensiones mostradas sean, en efecto, hasta cierto punto conocidas; solo que hasta el momento no habían sido encontradas en nuestra web ni se tenían identificadas. Con estos resultados se puede realizar una mejor clasificación para estudios posteriores.

3.3.6.8 Extensiones más encontradas en la Web de la UCI.

La Tabla 8 muestra las 10 extensiones más presentes en la Web de manera general.

Variable/Indicador	PHP	GIF	PNG	RPM	CSS	JPG	GZ	HTML	ICO	SWF
Cantidad de Ficheros	7201142	5038196	1929007	331408	252771	221557	91431	59026	37666	32374
Por Ciento de Presencia	46,4	32,46	12,43	2,14	1,63	1,43	0,59	0,38	0,24	0,21

Tabla 8: Extensiones más presentes en la Web de la UCI de manera general.

3.4 Códigos de estado de las páginas web descargadas.

El recolector de páginas funciona extrayendo direcciones de las páginas que han sido descargadas y es frecuente que entre estas direcciones aparezcan páginas que ya no existen o que simplemente se escribieron incorrectamente. Cada vez que el recolector contacta con un servidor Web éste retorna un código de estado que indica si la página existe o no, o si existe un motivo por el cual no se puede entregar el documento solicitado. La Tabla 9 muestra la distribución de las páginas de acuerdo a estos códigos de estado.

Los resultados mostrados demuestran que las páginas descargadas presentan un alto porciento de enlaces válidos; lo cual es muy positivo, pues evidencia un buen acceso y navegación de los usuarios por la Web de la UCI.

Estado HTTP	Código Estado HTTP	Documentos	Por ciento
OK	200	365 183	79,07%
PARTIAL CONTENT	206	452	0,10%
MOVED	301	64 953	14,06%
FOUND	302	24 141	5,23%
SEE OTHER	303	1 711	0,37
NOT FOUND	404	2 422	0,52%
ERROR PROTOCOL	94	8	0
ERROR TIMEOUT	95	977	0,21
ERROR DISCONNECTED	96	114	0,02
ERROR CONNECT	97	139	0,03
ERROR DNS	98	173	0,04
INTERNAL ERROR	500	726	0,16
NO CONTENT	204	7	0
BAD REQUEST	400	122	0,03
UNAUTHORIZED	401	66	0,01
FORBIDDEN	403	621	0,13
NOT ACCEPTABLE	406	16	0

Tabla 9: Distribución de las páginas web por códigos de estado.

De igual manera se puede observar un bajo porciento de páginas web con necesidad de autenticación por parte de los usuarios, lo cual hace menos trabajoso y más asequible el acceso a los sitios web de la Universidad. Así mismo, los errores por parte de los servidores web son prácticamente insignificantes ante el porciento total de páginas.

Por último aclarar que, aunque existe un bajo porciento de enlaces rotos (lo cual puede estar dado por la adquisición de una mayor conciencia en el momento de elaborar y dar mantenimiento a los sitios), no se debe descuidar este tema debido al impacto negativo que tiene sobre los usuarios. Actualmente existen más de 3800 enlaces rotos en la Web de la UCI y una gran cantidad de códigos de estado, aunque se pueden agrupar de la siguiente manera: (Ver Dibujo 11)

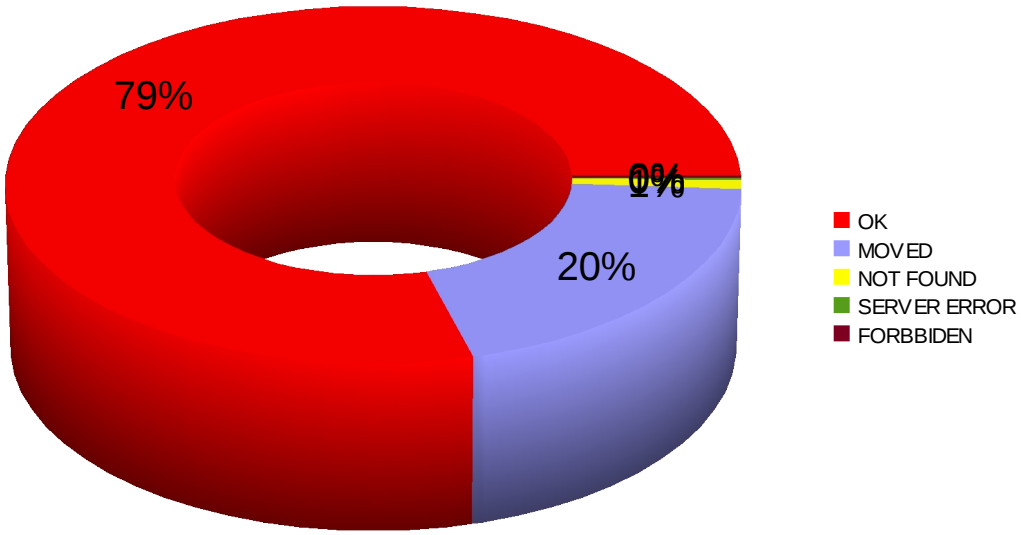
OK: Incluye todos los requerimientos exitosos: OK (200) y PARTIAL CONTENT (206).

NOT FOUND: El servidor no encuentra el documento pedido: NOT FOUND (404), ERROR PROTOCOL (94), ERROR TIMEOUT (95), ERROR DISCONNECTED (96), ERROR CONNECT (97), Y ERROR DNS (98).

MOVED: Incluye todos los requerimientos en los cuales el servidor redirige al recolector a otra página: MOVED (301), FOUND (302), SEE OTHER (303) y TEMPORARY REDIRECT (307).

SERVER ERROR: Incluye todas las fallas en el lado del servidor: INTERNAL SERVER ERROR (500), BAD GATEWAY (502), UNAVAILABLE (503), BAD REQUEST (400) y NO CONTENT (204).

FORBIDDEN: Incluye todos los requerimientos que no son permitidos, principalmente por tratarse de páginas protegidas con clave: UNAUTHORIZED (401), FORBIDDEN (403) y NOT ACEPTABLE (406).



Dibujo 11: Distribución de las páginas web por códigos de estado.

3.5 Estudio de las SCC de la Web de la UCI.

En un grafo, se dice que una parte de el es una componente conexa si es posible ir desde cualquier nodo de esa parte a cualquier otro nodo dentro de la misma parte. Se dice que una componente del grafo es una componente fuertemente conexa si esto es posible respetando la dirección de los enlaces. Dentro de una parte fuertemente conexa es posible ir desde cualquier sitio a cualquier otro sitio siguiendo enlaces. No toda la Web de la UCI es fuertemente conexa. Se considera en las componentes de tamaño 1 solamente los sitios que tienen al menos un enlace entrante o un enlace saliente.

DATOS INICIALES:

Para el estudio se tiene un número total de 240 sitios conocidos, de los cuales sólo 159 se mostraron con al menos una página en estado OK. El tamaño de la SCC más grande de la colección es de 50, con SCC-id 139; y existen además 107 SCC con sólo un sitio web.

DEFINICIONES:

MAIN, los sitios en la componente fuertemente conexa.

OUT, los sitios que son alcanzables desde ***MAIN***, pero que no tienen enlaces hacia ***MAIN***.

IN, los sitios que pueden alcanzar a ***MAIN***, pero que no tienen enlaces desde ***MAIN***.

ISLAS, sitios que no son accesibles ni hacia ni desde ***MAIN***.

TENTACULOS, sitios que sólo se conectan con ***IN*** (***TIN***) u ***OUT*** (***TOUT***), pero en el sentido inverso de los enlaces.

TUNEL, una componente que une las componentes ***IN*** y ***OUT*** sin pasar por ***MAIN***.

La notación de la parte ***MAIN*** se extiende con las siguientes componentes:

MAIN-MAIN, sitios que pueden ser alcanzados directamente desde la componente ***IN*** o que pueden alcanzar directamente la componente ***OUT***.

MAIN-IN, sitios que pueden ser alcanzados directamente desde ***IN*** pero no están en ***MAIN-MAIN***.

MAIN-OUT, sitios que pueden alcanzar directamente a ***OUT*** pero no están en ***MAIN-MAIN***.

MAIN-NORM, sitios que no pertenecen a las subcomponentes definidas anteriormente.

COMPONENTE	CANTIDAD DE SITIOS	POR CIENTO
MAIN NORM	18	11,32%
MAIN MAIN	10	6,29%
MAIN IN	2	1,26%
MAIN OUT	20	12,58%
IN	11	6,92%
OUT	52	32,70%
TIN	2	1,26%
TOUT	14	8,81%
ISLAS	30	18,87%
TUNEL	0	0

Tabla 10: Distribución de los Sitios en las Componentes Web.

Partiendo de estas definiciones, se puede conformar una vista macroscópica de la Web analizada. (Ver Fig 3)

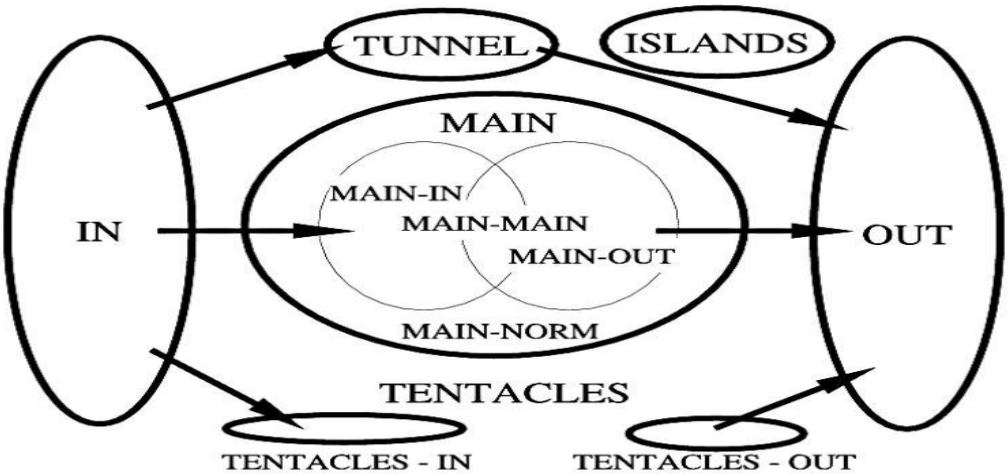


Figura 3. Vista macroscópica de la Web.

3.6 Tecnologías y herramientas utilizadas en los estudios webmétricos realizados.

Sistema Operativo: Debian GNU/Linux 4.0 (Etch).

Spider para la Recolección de Datos: WIRE Software V.0.21

3.7 Conclusiones del capítulo.

El presente estudio ha mostrado los principales resultados obtenidos en los Estudios Webmétricos realizados en la UCI. Los indicadores se dividieron en distintas clasificaciones para un mejor entendimiento; de igual forma se utilizaron imágenes y tablas de contenido, como método de apoyo para la información brindada. Se realiza un estudio de las componentes macroscópicas de la Web, que permite realizar una vista aproximada de su estructura. Se crean bases, no solo para futuros estudios en la Web de la UCI, sino además para otras áreas del país; dígase otros centros de estudio o trabajo, o incluso provincias. El desarrollo creciente de las tecnologías en Cuba implica la necesidad de extender estos estudios en todo el país [26]. Los resultados muestran la heterogeneidad de la Web de la UCI, algo positivo desde el punto de vista de su diversidad; sin embargo, por otro lado, negativo debido a la carencia de calidad por la presencia de numerosos sitios aislados, con poco contenido y pocas referencias.

Conclusiones Generales

En el presente trabajo se muestran los principales resultados de los tres Estudios Webmétricos realizados en la Universidad de las Ciencias Informáticas (UCI), que permiten extraer una serie de datos exhaustivos de los sitios web de dicho centro. Además, se ofrecen datos comparativos entre cada uno de los estudios, lo cual constituye la base fundamental para medir la evolución de la Web durante los últimos meses, tomando como punto de partida los Estudios Webmétricos realizados con anterioridad.

Se pueden mencionar, por ejemplo, los siguientes resultados generales:

- La Web de la UCI está compuesta por aproximadamente 160 sitios web, con más de 700000 páginas web, aunque sólo se han logrado descargar cerca de 462000.
- El 86% de las referencias a dominios externos al dominio de la UCI son al .CU, con más de 7700000 enlaces encontrados.
- Más del 90% de las páginas web analizadas se encuentran en idioma español.
- El sitio web con mayor contenido es el Portal del Proyecto ERP-CUBA (Planificación de Recursos Empresariales de Cuba), con más de 120000 páginas web.
- La Intranet de la UCI presenta el mayor Grado Interno y Externo de todos los sitios web.
- Más de un 97% de las páginas web analizadas están entre las profundidades 1 y 10.
- Más de un 93% de las páginas web analizadas fueron creadas o modificadas en el último año.
- En la Web de la UCI predominan las extensiones .PHP, .GIF, .PNG, .RPM, .CSS y .JPG.
- Predomina el Servidor de Aplicaciones Web Apache (en distintas variantes) y el Sistema Operativo GNU/Linux, del cual se encontraron cuatro distribuciones distintas.

Este estudio constituye una fotografía tomada a la Web. Y al medir sus resultados, en conjunto con otros estudios, se pueden establecer pronósticos y tendencias que ayudan a tener un mayor control y conocimiento de las tecnologías que actualmente son utilizadas por los usuarios y a establecer líneas de trabajo en función de mejorar el uso de las mismas y acelerar este proceso.

Referencia Bibliográfica

1. **Figuerola.C, Alonso.J**. Diseño de Spiders [En Línea] Marzo de 2006 [Citado el: 12 de Noviembre de 2008.] http://reina.usal.es/index2.php?option=com_content&do_pdf=1&id=65
2. Estadísticas sobre la Web chilena. [En Línea] 2002-2006 [Citado el: 15 de Noviembre de 2008.] http://www.todocl.cl/stats.phtml
3. **Baeza.R, Castillo.C. 2000.** Caracterizando la Web chilena 2000. [En línea] 2000. [Citado el: 17 de Noviembre de 2008.] http://www.todocl.cl/stats/jun2000/wcl2000.pdf
4. **Baeza.R, Poblete.B, Jean.F. 2003**. Evolución de la Web chilena 2001-2002. [En línea] Enero de 2003. [Citado el: 18 de Noviembre de 2008.] http://www.todocl.cl/stats/estudio2002.pdf
5. **Baeza.R, Castillo.C. 2005.** Caracterizando la Web chilena 2004. [En línea] 2005. [Citado el: 20 de Noviembre de 2008.] http://www.ciw.cl/webcl2004/Web_Chilena_2004.pdf
6. **Baeza.R, Castillo.C, Graells.E. 2007.** Características de la Web chilena 2006. [En línea] Marzo de 2007. [Citado el: 22 de Noviembre de 2008.] http://www.ciw.cl/material/web_chilena_2006/Web_Chilena2006.pdf
7. **Coutin.A, Valdés.M.** Estudio de las Estadísticas Web de accesos y visitas del Portal Cuba.cu 2002-2003. [En línea] 2003. [Citado el: 23 de Noviembre de 2008.] http://www.bibliociencias.cu/gsdl/collect/eventos/index/assoc/HASH0165.dir/doc.pdf
8. Sitio oficial del proyecto chileno WIRE. [En línea] 2000 [Citado el: 25 de Noviembre de 2008.] http://www.cwr.cl/projects/WIRE
9. Cursos de TI/Informática. Internet, Capítulo 3. [En Línea] 2002 [Citado el: 10 de Febrero de 2009.] http://www.mailxmail.com/curso/informatica/internet/capitulo3.htm
10. **Arroyo.N, Pareja.V.** Cibermetría. Estado de la cuestión. [En línea] Abril de 2005. [Citado el: 10 de Febrero de 2009.] http://digital.csic.es/bitstream/10261/4296/1/R-17.pdf
11. **Isidro F. Aguillo.** Cibermetría Introducción teórico-práctica a una disciplina emergente. [En línea] Febrero de 2003. [Citado el: 10 de Febrero de 2009.] http://internetlab.cindoc.csic.es/cursos/cibermetria.pdf
12. Documentación sin límites. Cibermetría y sucedáneos, primera parte. [En Línea] [Citado el: 10 de Febrero de 2009.] http://docunlimited.blogsome.com/2005/07/14/cibermetria-y-sucedaneos-primera-parte/
13. Definiciones de Robots de Búsqueda. [En Línea] [Citado el: 12 de Febrero de 2009.] http://www.abcdatos.com/buscadores/robot.html
14. Motores de Búsqueda. [En Línea] Septiembre de 2005 [Citado el: 12 de Febrero de 2009.] http://www.osmosislatina.com/aplicaciones/robots.htm
15. Web crawler. [En Línea] Diciembre de 2008 [Citado el: 15 de Febrero de 2009.] http://es.wikipedia.org/wiki/Web_crawler
16. **Baeza.R, Castillo.C, Tolosa.G, Bordignon.F. 2006.** Caracterización del Espacio Web de Argentina. [En línea] Febrero de 2006. [Citado el: 16 de Febrero de 2009.] http://eprints.rclis.org/archive/00009609/01/La_web_de_Argentina-Tolosa-Bordignon-Baeza-Castillo.pdf

17. Andreas Rauber**.** D-Lib Magazine, Volume 8, Number 12. [En línea] Diciembre de 2002. [Citado el: 16 de Febrero de 2009.] http://www.dlib.org/dlib/december02/rauber/12rauber.html
18. **Codenotti.B, Santini.M**. Propiedades estructurales de los Estados de África Web [En línea] Febrero de 2002. [Citado el: 16 de Febrero de 2009.] http://www2002.org/CDROM/poster/164/
19. **Baeza.R, Lalanne.F. 2004.** Characterization of the Korean Web. [En línea] Diciembre de 2004. [Citado el: 16 de Febrero de 2009.] http://quintay.dcc.uchile.cl/korea/korea.pdf
20. **Baeza.R, Castillo.C, Lopez.V. 2004.** Características de la Web de España. [En línea] Junio de 2005. [Citado el: 16 de Febrero de 2009.] http://www.catedratelefonica.upf.es/webes/2005/Estudio_Web_Espana.pdf
21. **Andras.A.** Buscando un pequeño dominio nacional – Informe preliminar. [En línea] [Citado el: 16 de Febrero de 2009.] http://www.ilab.sztaki.hu/websearch-data/Publications/p184-benczur.html
22. **Fernando.R, Tolosa.G, Bordignon.F. 2006**. Caracterización del Espacio Web de Perú. [En línea] Diciembre de 2006. [Citado el: 17 de Febrero de 2009.] http://eprints.rclis.org/7703/1/webpe.pdf
23. **Gomes.D.** A Characterization of the Portuguese Web. [En línea] 2003 [Citado el: 17 de Febrero de 2009.] http://xldb.fc.ul.pt/xldb/publications/webarchive2003.ppt
24. **Martínez.R.A 2006** Indicadores cibermétricos: ¿Nuevas propuestas para medir la información en el entorno digital? [Citado el: 17 de Febrero de 2009.]
25. **Mondelo.Y**. Primer Estudio Webmétrico en la UCI. Diciembre de 2008.
26. **Mondelo.Y, Amat.L**. Segundo Estudio Webmétrico en la UCI. Febrero de 2009.
27. **Mondelo.Y, Amat.L**. Tercer Estudio Webmétrico en la UCI. Abril de 2009.
28. **T. Berners-Lee, L. Masinter, M. McCahill.** RFC1738: Uniform Resource Locators (URL). Internet RFCs, 1994.

Recomendaciones

Tomar el presente Trabajo de Diploma como muestra de referencia para futuros estudios de la Web en la Universidad de las Ciencias Informáticas (UCI).

Extender la utilización de los Estudios Webmétricos no sólo a la universidad, sino a todas las regiones del país inmersas en el actual desarrollo tecnológico e informatización de la sociedad.

Investigar acerca de la utilización de los distintos indicadores webmétricos para el posicionamiento web.

Desarrollar una herramienta propia para el estudio de las características de la Web o contribuir al mejoramiento de la utilizada.

Anexo # 1: Guía de Instalación del WIRE.

Proceso de Instalación.

Lo primero es descargar el software:

- La página oficial de *WIRE* es: http://www.cwr.cl/projects/WIRE/
- La última versión se puede descargar de: http://www.cwr.cl/projects/WIRE/releases/

Requerimientos de Instalación.

Los siguientes paquetes son necesarios para proceder a instalar WIRE:

adns - DNS asincrónico de resolución.
xml2 – Biblioteca XML, incluyendo XML y analizadores XPath, versión 2.6 o posterior.
Swish-e - El motor de búsqueda de WIRE utiliza swish-e.

Los siguientes paquetes se sugieren para obtener mejores resultados:

LaTeX (con fullpage.sty, incluido en tetex-extras en algunas distribuciones) y **gnuplot** son necesarios para generar los informes.
djbdns - Útiles para el establecimiento de caché de *DNS* local.
docbook-xsl - Necesario para la generación de la documentación local.

Para instalar *WIRE*, luego de tener instaladas todas las dependencias, descompacte la distribución descargada del software, y mediante una consola muévase con el comando cd al directorio creado. Con privilegios de administración, ejecute los siguientes comandos de instalación:

```
% ./configure
% make
% make install
```

Es necesario realizar lo siguiente en el directorio de WIRE:

```
%aclocal
%autoheader
%automake -a
%autoconf
```

El núcleo del sistema tiene que ser ajustado para obtener un óptimo rendimiento. Esto puede hacerse fácilmente mediante el sistema de ficheros /proc en Linux (probado en el Kernel 2.4). Entre como root y ejecute los siguientes comandos:

```
% echo 32768 > /proc/sys/fs/file-max
% echo 131072 > /proc/sys/fs/inode-max
```

Esto establece el número máximo de archivos abiertos e ínodos. A veces, también debe establecer estos límites por usuario, en Linux, editar el archivo: **/etc/security/limits.conf** añadiendo al mismo las siguientes líneas:

- soft nofile 32000
- hard nofile 32000

Usted debe asegurarse de que el archivo **/etc/pam.d/login** incluye los límites de archivo: ***session required pam_limits.so***

Tenga en cuenta que los límites de los usuarios requieren que el mismo reinicie la sesión para aplicar los cambios.

Anexo # 2: Guía de Configuración del WIRE.

El archivo de configuración del *WIRE* es básicamente un fichero *XML*. La variable de entorno WIRE_CONF debe apuntar a este archivo de configuración. En el presente Anexo se ofrece una muestra de archivo de configuración que incluye varias de las opciones que por defecto se utilizan. De igual manera, hay algunos parámetros que se deben establecer antes de iniciar una recolección de datos.

Archivo de Configuración del *WIRE*: **consultar doc/sample.conf**

Antes de realizar un recorrido se deben realizar los siguientes ajustes obligatorios, en algunas variables de configuración:

collection/base

Este es el directorio donde la colección será almacenada.

collection/maxdoc collection/maxsite

Número máximo de documentos y sitios que pueden ser indexados. Estos valores no pueden ser modificados una vez que el Spider ha sido inicializado. El máximo real, es el 80% de estos valores.

seeder/accept/domain-suffixes

Los sufijos de dominio (código de país, nombres de dominio, etc) que van a ser descargados por el Spider.

harvester/resolvconf

La ubicación del Servidor DNS.

Los siguientes ajustes, son recomendados también:

harvester/nthreads

El número máximo de descargas simultáneas.

Sufijos numéricos:

- **K** = 1,000 Decimal kilo
- **M** = 1,000,000 Decimal mega
- **m** = Minuto
- **h** = Hora
- **d** = Día
- **w** = Semana

Otras variables de interés:

seeder/accept/protocol

El protocolo para el recorrido, actualmente solo soporta el protocolo http.

seeder/max-urls-per-site

No se adicionaran más de este número de *URL* por sitio. Algunos sitios de gran tamaño tienen infinita cantidad de direcciones *URL*. El valor por default es 10K.

seeder/add-robots-txt

Chequea la existencia del fichero de exclusión de robot: robot.txt. Valor por default 1, no se debe cambiar este valor. Es importante para respetar el protocolo de exclusión de robot.

seeder/extensions/download

Se definen los grupos de clasificación de las extensiones de los ficheros descargados. Como todas las extensiones son descargadas por default por el *Spider*, estas clasificaciones no serán usadas durante el recorrido, sino más bien a la hora de elaborar los informes.

seeder/extensions/log

Se definen los grupos de clasificación de las extensiones de los ficheros que no serán descargados, de estos solo se guardará la fuente donde se hizo referencia al mismo, nombre del fichero, y el *link* donde se encuentra.

seeder/extensions/stat

Se definen los grupos de clasificación de las extensiones de los ficheros que serán solamente contados. Se tomará la extensión y el número de veces que aparece en la colección.

seeder/extensions/ignore

Se definen los grupos de clasificación de las extensiones de los ficheros que serán ignorados durante el recorrido.

manager/maxdepth

Profundidad máxima a la que los documentos serán analizados. Tanto dinámicos como estáticos.

manager/max-errors

Cantidad máxima de errores permitidos para los sitios.

manager/score/pagerank

Aquí se establece como calcular el valor que tiene una página para la colección.

manager/score/siterank

Aquí se establece como calcular el valor que tiene un sitio para la colección.

manager/minperiod

Aquí se establece el tiempo que debe transcurrir para volver a visitar una página.

harvester/user-agent-comment

Descripción del Agente utilizado para el recorrido.

harvester/timeout

Tiempo para intentar descargar nuevamente una página cuando no está disponible.

harvester/maxfilesize

Tamaño máximo de los ficheros descargados.

harvester/wait

Tiempo de espera entre una descarga y otra dentro del mismo sitio.

gatherer/changetodynamic

El *Spider* puede detectar documentos dinámicos dentro de páginas estáticas. Si un documento cambia siempre luego de X visitas, entonces el mismo será considerado dinámico. Cantidad de veces que debe cambiar para ser considerado dinámico.

gatherer/maxstoredsize

Tamaño máximo en bytes de información.

gatherer/convert-to-utf8

Si está en 1, se codifica todos los documentos a UTF-8.

index

Parámetros para la indexación de la información.

analysis

Parámetros para el análisis de la información generada luego del recorrido.

Anexo # 3: Guía de Uso simple del WIRE.

En el presente anexo, se pretende explicar una guía simple para utilizar el *WIRE*. Algo así, como unas instrucciones paso a paso para correr el *Spider*. Es válido aclarar, que ya en este punto se debe tener correctamente instalada y configurada la herramienta.

Paso 1: Inicialmente se debe crear el directorio para almacenar la colección de información. Asumiendo que /opt/wiredata será el utilizado sería de la siguiente manera:

mkdir /opt/wiredata

Paso 2: Luego se procede a copiar el fichero de configuración en el directorio antes mencionado. El propósito principal de esto, es que permite utilizar distintos ficheros de configuración, en recorridos diferentes.

cp /usr/local/share/WIRE/sample.conf /opt/wiredata/wire.conf

Se puede, por otro lado, copiar directamente nuestro fichero de configuración particular para el recorrido que se quiere realizar.

Paso 3: Cambio de la variable de entorno WIRE_CONF, con la cual la herramienta hace referencia el fichero de configuración que debe utilizar.

Forma tcsh: ***setenv WIRE_CONF /opt/wiredata/wire.conf***

Forma bash/sh: ***WIRE_CONF=/opt/wiredata/wire.conf; export WIRE_CONF***

Paso 4: Editar el fichero de configuración, hasta que quede correctamente. Ver Anexo 2.

Paso 5: Crear una colección inicial en blanco.

wire-bot-reset

Paso 6: Adicionar la lista de partida de direcciones *URL* al *Spider*. Para ello se debe tener un fichero con estas direcciones *URL*. El mismo, debe contener *URL* absolutas a razón de una por línea; es recomendable incluir el fichero dentro de la colección de información. Por ejemplo, puede copiar el mismo en: /opt/wiredata/start_urls.txt

Ejemplo de Fichero:

http://www.sitio1.zzz/

http://www.sitio2.zzz/

http://www.sitio3.zzz/home

http://www.sitio4.zzz/modules/main

……

Note que “http” debe estar incluido en las direcciones del fichero de la lista de partida de direcciones *URL* brindada al *Spider*.

Si conoce además la dirección IP para algunos sitios, puede incluirla de la siguiente manera:

http://www.site1.zzz/ IP=10.35.18.235

http://www.site2.zzz/ IP=10.128.15.100

…….....

Luego, para pasar la lista de partida de direcciones *URL* al *Spider*, use:

wire-bot-seeder --start /opt/wiredata/start_urls.txt

Paso 7: Realice dos ciclos de prueba del software, esto le permitirá observar como va saliendo todo. El primer ciclo es especial, pues en este solo se descargarán los ficheros robot.txt, y se resolverán las direcciones IP de las direcciones *URL* proporcionadas. Use la siguiente secuencia de comandos para realizar dos ciclos de prueba:

```
wire-bot-manager
wire-bot-harvester
wire-bot-gatherer
wire-bot-seeder
wire-bot-manager
wire-bot-harvester
wire-bot-gatherer
wire-bot-seeder
```

Si ambos ciclos ocurren correctamente, lo más probable es que el fichero de configuración esté correcto también. Puede prestar especial atención a la salida del segundo ***wire-bot-harvester*** pues en el mismo podrá observar si se contactan o no los sitios proporcionados. Si este falla, entonces deberá utilizar el comando **wire-bot-manager --cancel** para cancelar el ciclo actual, antes de continuar el recorrido.

Paso 8: Realizar múltiples ciclos de la herramienta.

Para ello se puede utilizar el programa incluido ***wire-bot-run***; considerando que en cada ciclo son descargados por lo menos WIRE_CONF:config/manager/batch/size documentos, si esta variable tiene un valor de 100000 documentos, por ejemplo, en 50 ciclos se descargarían por lo menos unos 5000000 de páginas, aunque normalmente estará por debajo de este valor debido a errores en las páginas, servidores saturados, etc.

nohup wire-bot-run 50 >& /opt/wiredata/run.log &

Si se tienen privilegios de administración en el equipo utilizado, se puede correr el Spider con una máxima prioridad entre los procesos en ejecución. La siguiente línea, localiza el ID del proceso ***wire-bot-run***, y cambia su nivel de prioridad al máximo (Debe introducirse este comando como administrador del equipo).

nice -19 `ps -o "%p" --no-headers -C wire-bot-run`

¿Cómo detener el *Spider*?

Si el proceso actual es *wire-bot-harvester*, se puede detener de manera segura con la siguiente secuencia de comandos.

killall wire-bot-run
killall wire-bot-harvester
wire-bot-manager --cancel

En cualquier otro caso, una vez terminada su ejecución el proceso actual, solo bastaría con la última línea. De igual forma, si existe un fallo con el hardware, o la energía eléctrica falla.

¿Cómo generar los datos estadísticos y los reportes?

Todas las fases de análisis se realizan en un orden específico, y en cada una se debe introducir una línea de comando. La secuencia completa debe ser introducida en este orden (en algunos casos puede tardar varios minutos):

wire-info-analysis --pagerank
wire-info-analysis --hits
wire-info-analysis --sitelink-analysis
wire-info-analysis --doc-statistics
wire-info-analysis --site-statistics
wire-info-analysis --extension-statistics
wire-info-analysis --harvest-statistics
wire-info-analysis --lang-statistics

La herramienta cuenta con programas para generar gráficos y tablas en los reportes estadísticos. Los siguientes paquetes son necesarios para ello:

perl5 - Librería de lenguaje.
XML::LibXML - Módulo de Perl.
Latex – Instalación completa con fullpage.sty, incluido en tetex-extras en algunas distribuciones.
Gnuplot – Para generar los informes.

Luego solo basta con utilizar los comandos correspondientes para cada reporte:

Reporte sobre **Documentos**:
wire-info-analysis --link-analysis
wire-info-analysis --doc-statistics
wire-report-doc

Reporte sobre **Sitios**:
wire-info-analysis --sitelink-analysis
wire-info-analysis --site-statistics
wire-report-site

Reporte sobre **Grafo de Enlaces entre Sitios**:
wire-info-analysis --sitelink-analysis
wire-report-sitelink

Reporte sobre **Ciclos de Descarga**:
wire-info-analysis --harvest-statistics
wire-report-harvest

Reporte sobre **Idiomas**:
wire-info-analysis --lang-statistics
wire-report-lang

Reporte sobre **Extensiones**:
wire-info-analysis --extension-statistics
wire-report-extension

¿Donde encontrar información más detallada?

En ese caso, nada mejor que dirigirse al fichero de ayuda de la herramienta *WIRE*. Allí se encuentra todo lo mostrado aquí, y mucho más ;).

Anexo # 4: Distribución de sitios web por Dirección de IP.

NO	Dirección IP	Cantidad de sitios web	NO	Dirección IP	Cantidad de sitios web
1	10.0.0.12	52	27	10.128.50.54	1
2	10.0.013	43	28	10.128.50.155	1
3	10.208.0.14	13	29	10.128.50.221	1
4	10.3.10.45	10	30	10.128.50.223	1
5	10.128.50.121	9	31	10.128.50.224	1
6	10.128.50.122	8	32	10.128.50.225	1
7	10.0.0.10	7	33	10.128.50.226	1
8	10.0.0.11	5	34	10.32.30.103	1
9	10.210.0.6	4	35	10.3.10.23	1
10	10.128.50.120	3	36	10.3.10.25	1
11	10.3.10.41	3	37	10.3.10.27	1
12	10.209.0.9	3	38	10.3.10.28	1
13	10.0.0.9	2	39	10.3.10.31	1
14	10.0.0.30	2	40	10.208.0.4	1
15	10.0.0.8	1	41	10.208.0.9	1
16	10.0.0.20	1	42	10.208.1.201	1
17	10.0.0.31	1	43	10.209.0.11	1
18	10.0.0.32	1	44	10.209.0.18	1
19	10.0.0.48	1	45	10.209.0.20	1
20	10.0.0.49	1	46	10.209.0.21	1
21	10.0.0.91	1	47	10.209.2.10	1
22	10.0.0.92	1	48	10.209.12.2	1
23	10.0.0.170	1	49	10.210.0.4	1
24	10.0.0.171	1	50	10.210.0.8	1
25	10.128.21.90	1	51	10.210.0.10	1
26	10.128.50.52	1	52		

Tabla 11: Distribución de sitios web por Dirección de IP.

Anexo # 5: Extensiones Desconocidas más presentes en la Web.

NO	Extensión	Cantidad de Ficheros	NO	Extensión	Cantidad de Ficheros
1	com	13593	36	mspx	96
2	jigdo	11657	37	digests	90
3	meta	9662	38	uu	75
4	play	8601	39	xmanuals	73
5	metalink	6761	40	es	64
6	manifest	6596	41	inc	63
7	changes	2308	42	inf	62
8	tr	2154	43	use	57
9	src	1850	44	mask	55
10	hlp	1685	45	unmask	52
11	lzma	1246	46	dat	48
12	multiverse	900	47	fnt	48
13	db	898	48	mod	48
14	universe	898	49	keywords	47
15	main	896	50	tlk	47
16	restricted	890	51	contents	45
17	mht	840	52	orig	44
18	chm	785	53	legal	39
19	sign	756	54	asmx	37
20	sha1	645	55	jrexec	36
21	yast	574	56	mac	36
22	cu	487	57	macosx	36
23	eclass	480	58	os2	36
24	spl	432	59	dmg	34
25	pxf	388	60	org2	29
26	org	354	61	se	29
27	emz	284	62	dvd deltas	27
28	key	276	63	include	27
29	dtd	262	64	yaml	26
30	dos	252	65	plx	25
31	app	241	66	wsdl	25
32	squashfs	216	67	cmd	24
33	net	196	68	pisi	24
34	x	185	69	us	24
35	map	180	70	xpi	24

Tabla 12: Tabla de Extensiones Desconocidas más presentes en la Web.

Glosario de Términos.

Byte (B) Es la menor medida de almacenamientos de datos.

CGI Common Gateway Interface (Interfaz de Entrada Común).

CIBA Grupo de Cibermetría Aplicada.

CITMATEL Empresa de Tecnologías de la Información y Servicios Telemáticos Avanzados.

CIW Centro de Investigación de la Web (Chile).

CMS Content Management System (Sistema de Gestión de Contenidos).

ERP Los Sistemas de Planificación de Recursos Empresariales (*Enterprise resource planning, ERP* por sus siglas en inglés) son sistemas de información gerencial que integran y manejan muchos de los negocios asociados con las operaciones de producción y de los aspectos de distribución de una compañía comprometida en la producción de bienes o servicios. La Planificación de Recursos Empresariales, permite la integración de ciertas operaciones de una empresa, especialmente las que tienen que ver con la producción, la logística, el inventario, los envíos y la contabilidad.

GEWEB Generador de Estudios Webmétricos.

Gigabyte (TB) Una medida utilizada para el almacenamiento de datos de alta capacidad. Representa 1024 megabytes.

Heterogéneo Lo que no pertenece a un mismo género. Se dice de lo que está compuesto por cosas o partes diferentes.

Hipertexto El hipertexto es una tecnología que organiza una base de información en bloques distintos de contenidos, conectados a través de una serie de enlaces cuya activación o selección provoca la recuperación de información.

Hipervínculo (hyperlink) Una conexión entre un elemento de un documento de hipertexto como una palabra, frase, símbolo o imagen y un elemento diferente del documento, otro documento de hipertexto, un archivo o un guión.

Homogéneo Lo que pertenece a un mismo género. Se dice del compuesto cuyos elementos son de igual naturaleza condición.

HTTP El protocolo de transferencia de hipertexto o *HyperText Transfer Protocol* es el protocolo usado en cada transacción de la Web. Es un protocolo orientado a transacciones y sigue el esquema petición-respuesta entre un cliente y un servidor.

IMAP *Internet Message Access Protocol* o Protocolo de Acceso a Mensajes de Internet, es un protocolo de red de acceso a mensajes electrónicos almacenados en un servidor. Mediante *IMAP* se puede tener acceso al correo electrónico desde cualquier equipo que tenga una conexión a Internet. *IMAP* tiene varias ventajas sobre *POP*, que es el otro protocolo empleado para obtener correo desde un servidor. Por ejemplo, es posible especificar en *IMAP* carpetas del lado servidor. Por otro lado, es más complejo que *POP* ya que permite visualizar los mensajes de manera remota y no descargando los mensajes como lo hace *POP*.

Internet Red internacional que conecta miles de redes más pequeñas. “Internet” con mayúscula se refiere a la red que actualmente se usa, mientras que “internet” con minúscula es el concepto de interconectar varias redes.

Kilobyte (KB) Una medida utilizada para el almacenamiento de datos. Representa 1024 bytes.

Link (enlaces) Link, hipervínculo, vínculo, hiperenlaces Conexión entre dos equipos o nodos. Conexión de una página web con otra mediante una palabra.

Linux, Windows, UNIX Sistemas operativos.

Megabyte (MB) Una medida utilizada para el almacenamiento de datos. Representa 1024 kilobytes.

NNTP *Network News Transport Protocol* o Protocolo de Transferencia de Noticias en la Red, es un protocolo inicialmente creado para la lectura y publicación de artículos de noticias.

Página Toda entidad en la Web que tiene asociada una *URL*. En este documento usamos una definición un poco más restrictiva que no considera como páginas a imágenes, video, música y otros archivos multimedia o comprimidos.

Página estática Toda página que existe previamente a ser solicitada.

Página dinámica Toda página que es creada en el momento en que es solicitada.

PHP - *Hypertext Preprocessor* - Es un lenguaje interpretado de alto nivel embebido en páginas *HTML* y ejecutado en el servidor.

POO Programación orientada a objetos.

POP3 En informática se utiliza el *Post Office Protocol* en clientes locales de correo para obtener los mensajes de correo electrónico almacenados en un servidor remoto. La mayoría de los suscriptores de los proveedores de Internet acceden a sus correos a través de *POP3*.

Servidor Un computador que está conectado a Internet y presta algún servicio.

SINI Polo Productivo de Soluciones Informáticas para Internet.

sitio web Nombre de un ordenador que presta el servicio de proveer páginas web.

SNMP Es el Protocolo Simple de Administración de Red. Es el Protocolo parte de *TCP/IP* para el manejo y la administración remota de los recursos de la red. Es un conjunto de estándares de comunicación entre dispositivos conectados a la red sobre *TCP/IP*. *SNMP* permite a los administradores supervisar el desempeño de la red, buscar y resolver sus problemas, y planear su crecimiento.

TIC Tecnologías de la Información y las Comunicaciones.

UCI Universidad de las Ciencias Informáticas.

URI Aunque se acostumbra llamar *URL* a todas las direcciones Web, *URI* es un identificador más completo y por eso es recomendado su uso en lugar de la expresión *URL*. Un *URI* (*Uniform Resource Identifier*) se diferencia de un *URL* en que permite incluir en la dirección una subdirección, determinada por el “fragmento”. Esto se comprende mejor analizando la estructura de un *URI*.

URL Estándar para referirse a una dirección en la Web, ejemplo: “http://www.sitio.cl/pagina.html”.