Universidad de las Ciencias Informáticas

Facultad 10

Propuesta de pronósticos y tendencias sobre el comportamiento de la Web en la UCI.

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

Autores: Marledis Pupo Mulgado

Yisel Reyna Castro

Tutores: Ing. Yonny Mondelo Hernández

Lic. Raydel Zumeta Fernández

Ciudad de la Habana, junio de 2010

"Año 52 de la Revolución"

Declaración de autoría

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los __16__ días del mes de junio del año 2010

Marledis Pupo Mulgado

_____
Firma del Autor

Yisel Reina Castro

_____
Firma del Autor

Ing. Yonny Mondelo Hernández

_____
Firma del Tutor

Lic. Raydel Zumeta Fernández

_____
Firma del Tutor

"Podemos tener todos los medios de comunicación del mundo, pero nada, absolutamente nada, sustituye la mirada del ser humano"

Paulo Coelho

Agradecimientos

De Marledis:

Primero que todo se me hace imprescindible agradecerle a esta revolución por haberme dado la oportunidad junto a la UCI de graduarme como ingeniera en ciencias informáticas.

A mi papá quien a pesar de no conocer mucho sobre mi carrera siempre me apoyó y me dió aliento para no ser derrotada por las adversidades del destino y poder llegar a donde estoy hoy. A él que yo sé que se siente extremadamente orgulloso del camino que yo decidí seguir.

A mi madre que toda la vida se la ha pasado apoyándome y dándome todo lo necesario e incluso más, para que yo siguiera adelante. Ella que a pesar de los problemas siempre trata de darme lo mejor de sí y que me quiere tanto, ojalá y nunca me falte.

A Jorge que es mi segundo padre por apoyarme y tener tanta confianza en mí, por enseñarme a ser responsable de mis actos y por inculcarme siempre buenas cosas para la vida.

A Pedro que también se ha portado excelente conmigo, siempre me ha apoyado en los buenos y malos momentos, y aunque es un gruñón yo sé que me quiere mucho y desea lo mejor para mí.

No puede faltar mi familia que se sienten muy orgullosos de mí; mi bisabuela Nina que la adoro y es muy importante en mi vida, mi abuela Nia que es mi abuelita querida, mis tíos y tías (Joel, Juan Carlos, Luis, Melba y Eunice (Cuca como cariñosamente le decimos)), por hacerme pasar ratos agradables en mi vida y molestarme tanto en mi infancia. También a Caridad que se ha portado maravillosa conmigo todo estos años.

A mis tías Berta y Noelia que se han portado como mis segundas madres, a Nicolás, Nicola como cariñosamente le decimos, mis primas Puchy y Yanela que siempre me han dado buenos consejos.

A mi hermano que es el único que tengo y lo quiero mucho.

Agradecimientos

A todos mis primos; Lissette, que es más que mi prima, mi hermana, Yoandris, Lisbet, Leanet, Yarlenis, Sheila, Norgito y Yoxania.

Claro que no puede faltar Marcel que aunque es un peleón lo quiero mucho, y en estos dos años solo ha querido lo mejor para mí, aunque muchas veces no le haga caso. Te adoro mi vida y gracias por el amor que me brindas aunque a veces de formas extrañas.

A mis suegros (Migdalia y Rodolfo) y mi cuñada (Anyel) que desde que los conocí solo he recibido afecto y cariño de su parte, acogiéndome en su familia como una hija más.

A mis grandes amigas, María y Ariannys que son dos hermanas para mí, que hemos pasado buenos y malos ratos, gracias por escuchar mis problemas y darme fuerzas para seguir adelante, las quiero mucho.

A Clary y Armín por ayudarme cuando lo necesité y por quererme tanto.

A la profe Graciela que se ha preocupado por nosotras y por el desarrollo del trabajo de diploma de forma incondicional, que nos atendió siempre que fuimos a molestarla, y que si no fuera por ella tal vez no estaría escribiendo los agradecimientos ahora. De verdad muchas gracias profe. A la profe Cristina por sus consejos y ayuda.

A mi compañera de tesis Yisel, por soportarme todo este tiempo y ser comprensiva conmigo, nunca te voy a olvidar Yise te quiero mucho, a su novio Alberto por todo lo que hizo por nosotras en el transcurso de la tesis, gracias Albertini.

A mis compañeros de aula en estos cinco años de carrera, a Mayrel y Angela que han estado conmigo en las buenas y malas en estos 5 años, escuchando mis problemas, apoyándome y dándome consejos, por todo esto y más, amigas es una palabra que no es suficiente para describir nuestra relación, espero que no perdamos el contacto, las quiero mucho. A Dianelis por su ayuda en la confección del documento, negra si no fuera por ti Yisel y yo no lo hubiésemos logrado, te quiero mucho negra culona.

A todos los que estuvieron y están involucrados en mi vida de alguna forma u otra les agradezco su apoyo en algún momento que lo necesité.

Agradecimientos

De Yisel:

Primeramente agradecerles a mis padres, por haberme apoyado tanto en estos cinco años de carrera, por ser ante todo mis guías y mi ejemplo, por confiar tanto en mí, por estar a mi lado en todo momento y por ser el sostén que todo hijo necesita para seguir adelante.

A una persona muy importante en mi vida, mi hermanito Danilo, gracias por quererme tanto, y por ser para mí el hermano más maravilloso del mundo, no tengo palabras para decirle cuanto lo quiero.

A mi abuelita Ana que desde pequeña estuvo a mi lado y a mi otra abuela Miriam, gracias a las dos por darme todo su cariño.

A mi tía Nilvia y a Demetrio, gracias por ser mis segundos padres en todo este tiempo, y ayudarme en todo lo que me hizo falta.

A mi tío Jorge (Negro), a mi tío Eddy (Jabao), a mis otros tíos Angelito, Jorge, Julio, Magalis, Rebeca, a mis primos: Angélica, Anniélica, Jorgito, Rebequita, Álden, Alexey, Enriquito y a toda mi familia en general, gracias por estar siempre orgullosos y pendientes de mí, y por ser partes de mi vida.

Agradecer también a María Elena, papá Alfredo, mamita Olga, Delmis, Lourdes Vega, Sonia, Juana, Maribel, Iskra, que aunque no sean familiares míos de sangre, se han comportado conmigo como si lo fuesen, y siempre han estado pendientes de mí.

A mi compañera de tesis, Marledis (Loty) que más que eso es mi amiga, y sin ella no hubiese podido estar aquí hoy, pues en todo momento me supo apoyar como una verdadera compañera y siempre me dio mucho aliento para seguir adelante.

Agradezco a los tutores, Yonny Mondelo y Raidel Zumeta, por el apoyo y la ayuda brindada durante el desarrollo de la Tesis.

A todos los integrantes del tribunal y muy en especial a la profesora Graciela, pues siempre estuvo a nuestra disposición para ayudarnos con todo lo que nos hiciera falta, y a María Cristina, pues también se portó muy bien con nosotras.

Agradecimientos

A mi novio Alberto, por no tener momentos malos para mí, por haberme apoyado en todo momento y también por ser tan bueno y comprensible conmigo.

A todos mis amigos de antaño y a los que llegué a hacer a lo largo de la carrera: Ané, Alberto Garnache, Denier, Diana, Yusel, Zulia, Rosy, Graciela, María de Lourdes, Juan Manuel, Enmanuel, Elián, gracias a todos pues fueron los que estuvieron en las buenas y en las malas, apoyándome en todo y siendo incondicionalmente mis verdaderos amigos.

A mis compañeros de aula, todos los varones: Isleam, Adnier, Yunier, Merallo, Charly, Javier, Miguel Angel, el Chino, Carlos, Reynier, Luis, a todos gracias por ayudarme en los momentos que los necesité, y principalmente las hembras: Darianne, Marisol, Yilena, Susel, Eglis, Yuliet, Karina, Angela, Dianelys, Lisandra Cala, Mayrel, Evelyn, gracias a todas, pues con ellas pasé momentos muy buenos y estuvieron conmigo todo este largo tiempo, soportándome y dándome todo su apoyo. También agradecerles a Roxana, Anay y Laura, pues también conviví con ellas y fueron muy buenas conmigo en todo ese tiempo.

A todas mis compañeras de deporte: Yordi, Dalvis, Mayi (la fanática), Made, a la profesora Yunelsis, y en especial a las niñas del Voly: Ivette, Yailén San Juan, Estela, Mavis, Lisbet, gracias a todas ya que se portaron muy bien conmigo y me apoyaron mucho, y puedo decir que una de las cosas que más disfruté en estos cinco años fue haber podido ser partícipe de todos los juegos deportivos de mi facultad.

A todos mis vecinos, allá en mi casa, Ana Celia, Magdalena, Sandra, Tía Juana, Lourdita, Tía Mayra, Tío José, Herminia, Isora, Idannis, Idelaine, Isorita, Magalis, Rosa, María, que aunque hoy no puedan estar aquí, les agradezco todo el apoyo y confianza que me dieron, y sé que siempre esperaron lo mejor de mí.

Y en general a todos los que están presentes en este día tan especial para mí, muchas gracias.

Dedicatoria

De Marledis:

Este trabajo se lo dedico a mi madre por ser como es conmigo, te mereces lo mejor del mundo (mama) como cariñosamente le digo.
A mis padres, Gilberto, Jorge y Pedro que me han dado su granito de arena en todo el transcurso de mi vida.
A toda mi familia por el amor y el cariño que me brindan.
A la revolución cubana por permitirme hacerme ingeniera.
En fin, se la dedico a todos los que tuvieron y está involucrado en mi vida.

De Yisel:

Esta tesis va dedicada primeramente a tres personas que fueron muy especiales para mí, y que en este momento por desgracia de la vida no pueden estar conmigo.

A mi abuelo Yuri, por ser un hombre increíble y por brindarme muy buenos consejos los cuales me dieron mucha fuerza para seguir adelante, en los momentos más difíciles de mi carrera.

A mi gran amigo y hermano Michel, la vida le jugó una mala pasada, pero yo siempre lo tendré presente, por todos los buenos momentos que pasamos juntos y lo recordaré toda mi vida como uno de mis mejores amigos.

A mi tía Nancy, que más que a una tía la quise como a una abuela, y fue una de las personas que más esperó y anhelaba este momento, pero sé que donde esté, va a sentirse muy orgullosa de mí.

Y por último a mis padres y mi hermano, pues ellos son la razón de mi vida, y todos los días les doy gracias a dios por tenerlos a mi lado.

Resumen

El avance tecnológico y los análisis cuantitativos se ven facilitados y al mismo tiempo obligados a encontrar nuevos campos de acción, como es el caso de los estudios que se están desarrollando actualmente sobre el contenido y estructuras de las páginas Web. Por lo que estudiar las características de la Web permite establecer parámetros sobre el desarrollo de la misma, así como observar tendencias de los usuarios y/o desarrolladores a partir de la información que nos proporciona. La ciencia encargada del estudio de la Web; es la *webmetría*, ciencia bastante joven en el campo científico; enfocada fundamentalmente en la bibliometría e informetría, ciencias estas más antiguas y relacionadas con la información ya sea digital o en copia dura. La información cuantitativa y cualitativa de la Web se puede obtener a través de indicadores webmétricos, estos se pueden dividir en cuatro grupos fundamentales: de conectividad, impacto, densidad y descriptivos. El presente trabajo de diploma tiene como propósito caracterizar la Web de la UCI y luego a través de las comparaciones realizadas entre los cinco estudios webmétricos dar una visión del comportamiento que puede seguir dicha Web.

Palabras claves: Web, webmetría, indicadores webmétricos, estudios webmétricos, información.

Índice

Índice de figuras.

Índice de tablas

Índice de anexos

Introducción

En el mundo contemporáneo el desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC) ha encaminado las sociedades contemporáneas hacia un mundo digitalizado; constituyendo la World Wide Web (WWW) el elemento que fomenta este acelerado progreso tecnológico, lo cual es provocado por el inmediato intercambio de información, se hace imprescindible por tanto definir ¿ qué es la Web?

Aunque muchos lo consideren fruto del marketing y del auge que ha tomado internet como plataforma de negocios, se pueden encontrar diversos conceptos y definiciones del término.

Tim Berners-Lee describe la Web como "una red que trajo como resultado a la sociedad poder para el individuo, la eficacia social, la comprensión y la armonía y el funcionamiento de la potencia de la informática en la vida real" [1]. Aunque es muy cierto que la Web proporciona mucho conocimiento a la sociedad, no pierden validez aquellos criterios que plantean que muchas veces ese conocimiento no es usado para el bienestar del mundo, ni en función de la armonía o comprensión que debe existir. Unos pocos aprovechan esta enorme ventana que se les abre a la información para beneficio de sí mismos sin considerar las consecuencias negativas que traen para la humanidad. Siendo muchas veces poco confiable e incluso con poco nivel científico la información brindada en la WWW, utilizada también para promocionar en cierto modo las campañas mediáticas llevadas a cabo en todo el mundo con fines belicistas.

Para Cronin y McKim, "La web funciona como un foro mundial, un espacio compartido que crea nuevas formas de interacción social" [2]. Ciertamente la Web se abre al mundo como el ciberespacio ideal donde se puede encontrar ofertas de variados servicios que muchos solo se podían ofrecer de forma tradicional y que ahora la Web se encarga de proporcionar sin muchas dificultades, además fomenta el intercambio cultural imprescindible para este mundo contemporáneo. Pero lamentablemente las ventajas que posee no son para el beneficio de toda la humanidad, ya que muchos de los seres humanos que habitan este planeta carecen del acceso necesario y muchos incluso de la tecnología para poder ser partícipes de esta nueva era de la información.

Wolton plantea que "la Web es un subconjunto de Internet que conecta a las páginas por las estructuras de hipertexto" [3]. Las autoras no concuerdan con lo expresado pues aunque sea solo un elemento

tecnológico que ha evolucionado al mundo se deben ver sus beneficios más allá de lo superficial. Una gran parte de los servicios de conexión proporcionan recursos como enciclopedias, noticieros, acceso a bibliotecas y otros materiales educativos de valor, servicios de gran utilidad para la formación de las sociedades modernas. Esta es la cara más amable, útil y pedagógica de la red y la que convierte a la Web en una herramienta de alto valor educativo. Como expresaría Castells, "la Web es mucho más que sólo una tecnología, es un medio de comunicación, integración y organización social" [4].

El mundo actual es una sociedad de redes de información electrónica, de tecnologías basadas en archivos de sonido, vídeo y animación que cambian la forma de comunicación tradicional. La aparición de Internet, donde surgen nuevas formas de literatura gris como los foros de discusión y las publicaciones electrónicas, representan un modo revolucionario de comunicación, que conduce a la desaparición de soportes y medios considerados como clásicos en los procesos de transferencia de información [5].

La Web se encuentra a la cabeza de los medios de difusión masivos, proporcionando información y diversos servicios a los millones de usuarios que día a día interactúan con la misma. Debido al desarrollo constante que posee, el estudio de sus características nos brinda variada información sobre su comportamiento, por lo que la importancia de conocer estas características radica en que permite establecer parámetros sobre el desarrollo de la web así como observar tendencias de los usuarios y/o desarrolladores. En virtud de la información que nos proporciona se han realizado varios trabajos relacionados con la caracterización de la Web.

Los indicadores webmétricos son elementos que se tienen en cuenta para realizar un estudio de la Web, los mismos ayudan a caracterizar el comportamiento de la Web. Algunos de los países que han incursionado en estos estudios son: Argentina, Austria, Chile, Brasil, Corea, España, Hungría, Perú, Cuba [6].

En los diferentes artículos sistematizados relacionados con los estudios webmétricos, el término utilizado para referirse a la webmetría es webometrics (en inglés) y al hacer la traducción al español se traduce como webometría o webmetría. Por lo que se utilizará el término **webmetría** en todo el desarrollo del mismo.

En Cuba, las incursiones que se han hecho en el tema han sido estrictamente teóricas, lo que ha posibilitado forjar las bases para realizar ya en la práctica un estudio de la Web cubana. La Empresa de Tecnologías de la Información y Servicios Telemáticos Avanzados (CITMATEL) es responsable de la publicación del artículo titulado "Estudio de las Estadísticas Web de accesos y visitas del Portal Cuba.cu"[1], el cual ofrece importante información al respecto.

En la Universidad de las Ciencias Informáticas se han realizado cuatro estudios webmétricos, que aportan importantes resultados estadísticos sobre la Web de la UCI a partir de los sitios y páginas que son exhaustivamente analizados. Algunos de los datos arrojados por estos estudios son, por ejemplo: cantidad total de sitios y páginas web existentes en el dominio UCI.CU, promedios de texto, profundidad, edad e idiomas por cada sitio web analizado, entre otros datos.

A partir de esta información, se cuenta con varias fotografías de la Web en la UCI durante los últimos años; que permiten conocer el estado real de la misma, bajo determinadas circunstancias objetivas y subjetivas, independientemente de las variaciones considerables que pueda sufrir durante el paso del tiempo.

Sin embargo, no se cuenta con una visión que permita establecer pronósticos y tendencias que ayuden a tener un mayor conocimiento y control de las tecnologías que son utilizadas (o de aquellas que pueden comenzar a ser obsoletas), y trazar líneas de trabajo en función de mejorar el uso de las mismas. De lo anterior se puede establecer el siguiente **problema a resolver**:

¿Cómo establecer futuros pronósticos y tendencias de evolución de la Web en la Universidad de las Ciencias Informáticas?

Por tanto el **objeto de estudio** de la investigación son los Estudios Webmétricos, siendo el **campo de acción** el proceso de sistematización de la información de la Web en la Universidad de las Ciencias Informáticas (UCI).

[1] http://www.bibliociencias.cu/gsdl/collect/eventos/index/assoc/HASH0165.dir/doc.pdf

Se establece como **Objetivo General**:

Proponer pronósticos y tendencias sobre el comportamiento de la Web en la UCI.

Lo cual se desglosa en los siguientes **objetivos específicos**:

- Sistematizar teóricamente sobre los indicadores webmétricos y las caracterizaciones de la Web en el mundo y en la UCI.
- Comparar las características de la Web en la UCI, y sus antecedentes.
- Proponer pronósticos y tendencias sobre el comportamiento de la Web en la UCI a partir de las comparaciones previas.

Para dar cumplimiento a estos objetivos se proponen las siguientes **tareas de investigación:**

- Sistematización de los antecedentes de los indicadores webmétricos en los estudios de la Web en la UCI y de las caracterizaciones que se han realizado anteriormente.
- Comparación de las características de la Web en la UCI, y sus antecedentes.
- Propuesta de una visión de pronósticos y tendencias sobre el comportamiento de la Web en la UCI a partir de las comparaciones previas.

Posible Resultados:

Propuesta de pronósticos y tendencias sobre el comportamiento de la Web en la UCI.

Dentro de los **métodos de investigación** utilizados se encuentran los siguientes:

Analítico-Sintético

Se aplicará para lograr el entendimiento a partir del análisis de caracterizaciones realizadas con anterioridad sobre la Web tanto de la Universidad de las Ciencias Informáticas como otras realizadas en el mundo y los indicadores webmétricos utilizados en los mismos.

Histórico-Lógico

Permitirá una mayor comprensión de los estudios webmétricos a través del análisis de su evolución a nivel mundial y además constataremos teóricamente cómo ha evolucionado la Web de la UCI en este último año.

El presente trabajo de diploma consta de dos capítulos distribuidos de la siguiente manera:

Capítulo 1: Fundamentación Teórica, centrado en el estado del arte relacionado con el objeto de estudio, abordándose la existencia y resultados obtenidos en los diferentes estudios webmétricos desarrollados a nivel nacional e internacional. Además de los principales conceptos tratados entorno a la investigación.

Capítulo 2: Caracterización, pronósticos y tendencias sobre el comportamiento de la Web en la UCI, enfocado en la caracterización realizada a la Web de la UCI, a través de los diferentes indicadores webmétricos utilizados para obtener resultados cuantitativos y cualitativos que permitan en una segunda parte del capítulo establecer posibles pronósticos y tendencias a seguir por la Web de la universidad.

Capítulo 1: Fundamentación Teórica.

1.1. Introducción.

El desarrollo de este capítulo tiene como objetivo definir conceptos importantes para entender la investigación. En el mismo se pretende dar una introducción al tema que será centro de atención en el transcurso de la investigación, o sea, los estudios webmétricos a nivel nacional e internacional, haciendo énfasis en los estudios realizados a la Web de la UCI anteriormente y la evolución que presenta a través de un estudio actualizado de la misma.

1.2. Estudios Realizados.

Estudiar la web de manera global es algo complejo, pues se debe tener un control actualizado de la misma en todos los países que al menos poseen visibilidad en la web mundial. Hay algunas instituciones que se han encargado de esta tarea hace algunos años. Ranking Mundial de Universidades en la Web, es una iniciativa del Laboratorio de Cibermetría que pertenece al Centro de Ciencias Humanas y Sociales (CCHC) que a su vez es parte del mayor centro nacional de investigación de España, el CSIC.

El Ranking Mundial de Universidades en la Web fue lanzado oficialmente en el año 2004, y es actualizado cada 6 meses (los datos son recolectados durante los meses de Enero y Junio y publicados un mes más tarde). Los indicadores Web utilizados están basados y se correlacionan con los tradicionales indicadores bibliométricos y cienciométricos. El objetivo del proyecto es el de convencer a las comunidades académicas y políticas de la importancia de la publicación web no sólo para la diseminación del conocimiento académico sino también como una forma de medir la actividad científica, el rendimiento y el impacto [7].

Este centro de investigación usa 4 indicadores webmétricos, y ofrecen un RANKING de Universidades de acuerdo a ellos. Los mismos son: **Tamaño, Visibilidad, Ficheros Ricos, Académicos.** Los valores obtenidos a través de estos indicadores son combinados basados en la siguiente fórmula:

WEBOMETRICS RANK		
VISIBILITY (external inlinks) 50%	SIZE (web pages)	20%
	RICH FILES	15%
	SCHOLAR	15%

Figura 1. Principales indicadores webmétricos utilizados para establecer el ranking.

Ranking Mundial de Universidades en la Web, siendo la siguiente tabla una muestra de las 10 primeras entre las 8000 que están registradas en el sitio.

			POSICIÓN			
RANKING MUNDIAL	UNIVERSIDADES	PAÍS	TAMAÑO	VISIBILIDAD	FICHEROS RICOS	SCHOLAR
1	Harvard University		2	3	20	1
2	Massachusetts Institute of Technology		1	1	1	5
3	Stanford University		6	2	5	17
4	University of California Berkeley		7	4	28	27
5	Cornell University		4	5	14	33
6	University of Washington		12	7	3	68
7	University of Minnesota		9	12	4	16
8	Johns Hopkins University**		40	21	42	2
9	University of Michigan		8	8	32	21
10	University of Wisconsin Madison		3	9	12	53

Figura 2. Ranking Mundial de las primeras diez Universidades

1.2.1. Estudio Latinoamericano.

Utilizando los datos cibermétricos de las 500 primeras universidades latinoamericanas de acuerdo a indicadores Web obtenidos mediante motores de búsqueda se analiza la comunicación de sus actividades científicas y académicas a través de internet. Se presta atención a los llamados ficheros ricos, normalmente asociados a documentos ligados a la publicación científica. Además el uso de estos formatos para la comunicación informal se está generalizando entre las instituciones líderes de la región.

Resultados Obtenidos

Un total de 25 países están representados en la muestra, siendo Brasil con 133 universidades y México con 76 los que tienen más instituciones entre las 500 primeras (figura 3). La región representa solo el 2.0% del total de las 1000 primeras universidades del mundo, aunque dicho porcentaje sube al 6.8% cuando consideramos las 5000 primeras.

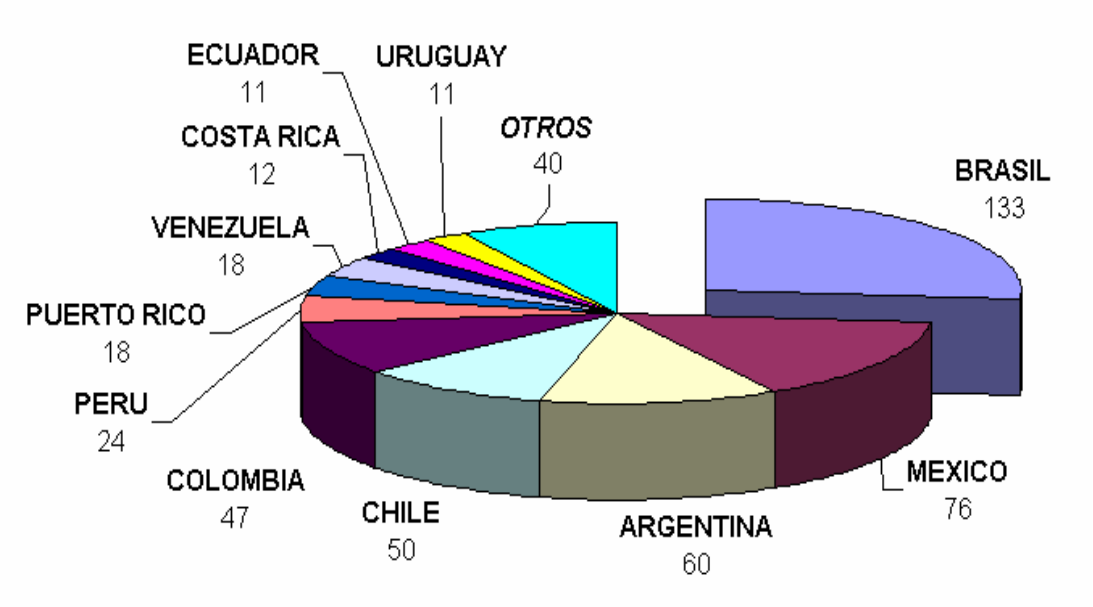

Figura 3. Distribución por países de las 500 primeras universidades latinoamericanas de acuerdo a indicadores Web (Rank webmétricos, enero 2006).

Las medidas de tamaño se obtuvieron de la combinación de datos de los cuatro motores pero dada la irregularidad de su comportamiento, se excluyeron los valores máximo y mínimo. La visibilidad se obtuvo a través de los motores Yahoo y MSN Search, dada la imposibilidad de derivar esta medida desde Google.

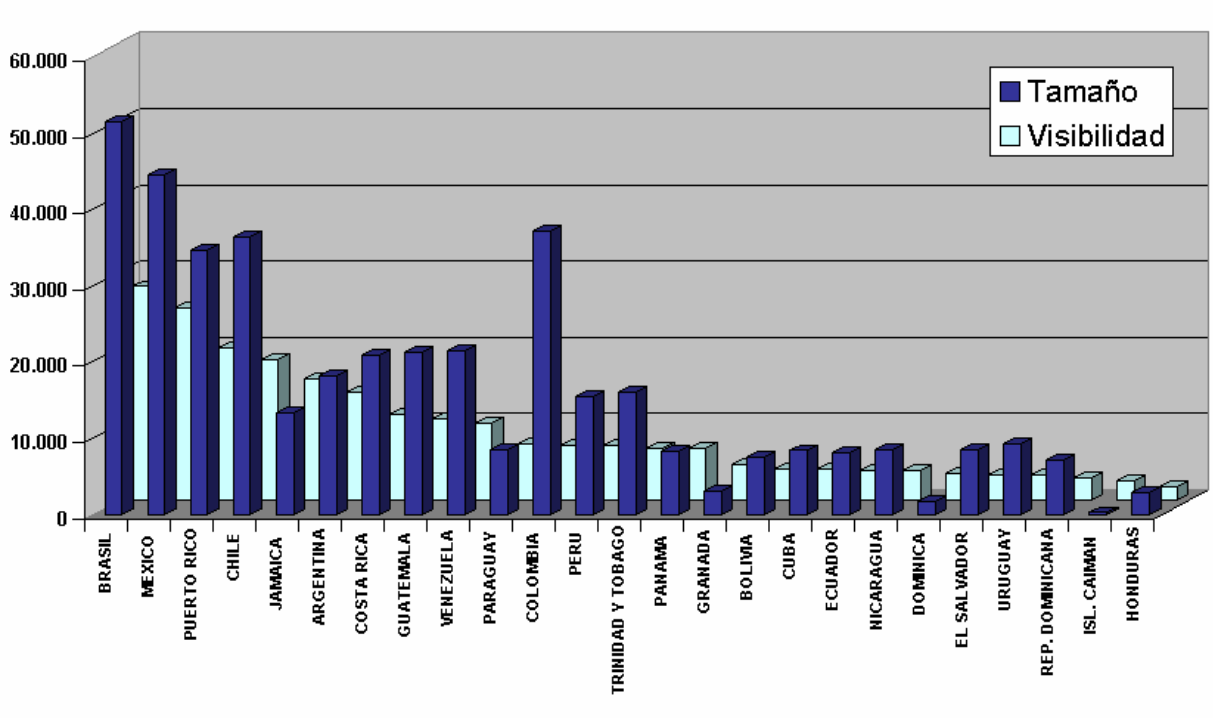

Figura 4. Media de páginas (tamaño) y enlaces recibidos (visibilidad) por universidad en los países representados en la muestra (enero 2006).

Tal como se muestra en la figura 4 Brasil no solo es el país mejor representado en la muestra, sino que también es el que mayor tamaño medio por universidad y más enlaces externos recibe, clasificación esta última donde destacan los países que utilizan el inglés (Jamaica y Puerto Rico). Esto significa que las universidades brasileñas además de numerosas entre las 500 primeras, ocupan en dicha lista posiciones de liderazgo. Hay que destacar el escaso volumen de páginas de las universidades argentinas respecto a lo esperado y la baja visibilidad de las colombianas, a pesar de su gran tamaño medio.

El estudio de la producción más ligada a procesos de comunicación tanto académica como científica se centra en el volumen de ficheros ricos y su reparto según tipología entre las universidades de la región.

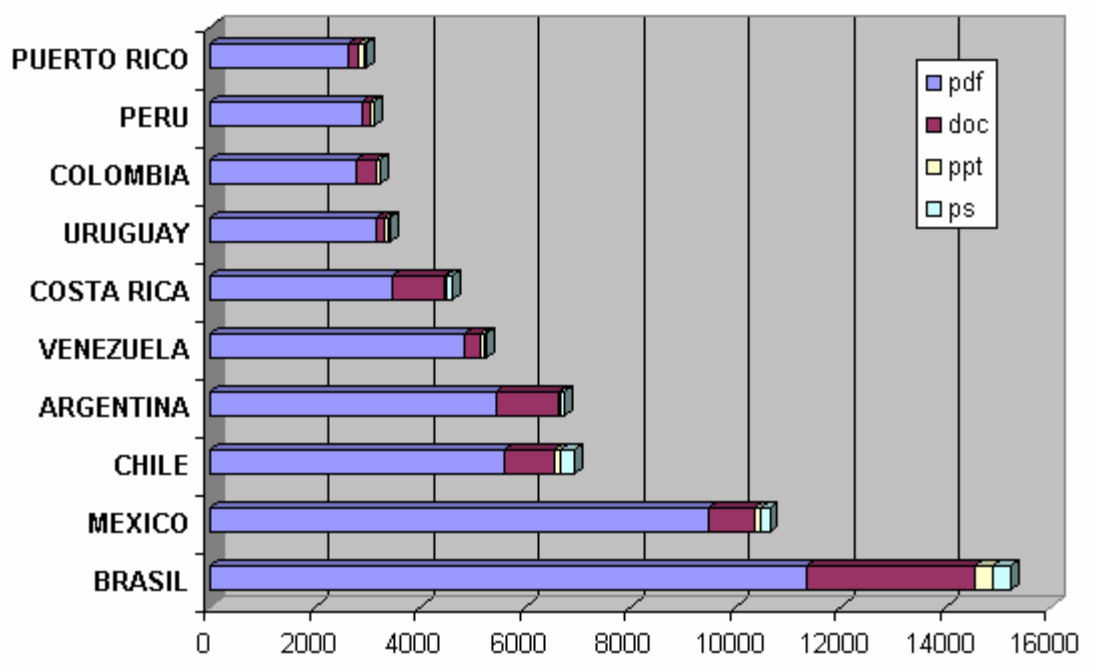

Figura 5. Distribución de los ficheros ricos según formato y país (media de universidades, enero 2006).

El formato Adobe Acrobat (.pdf) es consistentemente el más utilizado (figura 5). La edición de este tipo de documento requiere el manejo de un programa habitualmente no disponible en los paquetes ofimáticos convencionales (tales como MS Office). Aunque es posible convertirlo desde otros formatos, esta dificultad nos indica que se reserva para situaciones donde se valora su universalidad de facto como formato de comunicación de documentos por la Web.

Un grupo reducido de universidades latinoamericanas compiten en igualdad de condiciones con instituciones del resto del mundo en lo que respecta a su compromiso con la publicación y diseminación de conocimiento a través de la Web. Se trata fundamentalmente de grandes instituciones nacionales de carácter público, aunque también hay una buena representación de universidades católicas. Destacan por número (entre las mejor clasificadas) las universidades brasileñas y mejicanas (Sao Paulo, UNAM), aunque también otros parámetros las sitúan en las primeras posiciones.

La comunicación informal a través de formatos ricos parece asumida en todo este grupo de universidades de élite, posiblemente por iniciativas individuales de autoarchivo de documentos o mediante la creación de repositorios institucionales o temáticos de artículos. Nuevamente son las universidades brasileñas y

mejicanas las que presentan un mayor número de documentos en formato ricos tales como pdf o doc, lo que indica que efectivamente publican de manera abierta e intencionadamente parte de sus contenidos de alto contenido científico.

Es posible que las políticas e iniciativas que promueven la publicación en la Web no penetren todavía a todos los niveles académicos, pero ya son responsables de contribuir significativamente a la mejora de posición en internet, atrayendo visitas y enlaces externos [8].

1.2.2. Estudios realizados en la UCI.

En la UCI se han realizados cuatro estudios webmétricos que han aportado gran cantidad de información sobre el comportamiento, visibilidad y funcionamiento de dicha Web. Permitiendo hacer un análisis comparativo entre ellos, y así medir su evolución en años anteriores, forjando además las bases para este nuevo estudio que servirá para actualizar la información que se tiene de la Web de la UCI.

Todos los estudios han sido desarrollados por el Proyecto Generador de Estudios Webmétricos (GEWEB) del Grupo de Proyectos de Cibermetría Aplicada (GPsCIBA), perteneciente al Departamento Productivo de Soluciones Informáticas para Internet (SINI). Se utilizó en todos los casos como Spider, el sistema WIRE[2], desarrollado en el Centro de Investigación de la Web (CIW) de Chile.

Otras regiones que han realizado estudios webmétricos son: Argentina, Austria, África, Chile, Corea, Cuba, España, Hungría, Perú y Portugal. Basados en aspectos tales como: tamaño de la página, términos más utilizados, grado entrante y saliente de páginas, PageRank (Ranking de las páginas), códigos de respuestas HTTP, longitud de las URL, profundidad de los documentos, tipos de archivo, tamaño de los archivos, direcciones IP, tipos de Red, sistema operativos.

También se tomaron en cuenta los datos obtenidos del análisis sintáctico como nivel de páginas HTML, los tipos de cabeceras HTTP, tipos de servidores, cifras globales y estudio de vocabulario, tipo de documento e idioma, número de referencia hacia y desde un dominio, representación de la estructura global de hipervínculos entre dominios y preferencias de los usuarios, Documentos no indexables por el

[2] Web Information Retrieval Environment (Entorno de Recuperación de Información en la Web).

buscador, total de accesos, total de archivos, total de páginas, total de visitas, total de clientes, entre otros [6].

1.3. Webmetría.

El avance tecnológico y los análisis cuantitativos se ven facilitados y al mismo tiempo obligados a encontrar nuevos campos de acción, como es el caso de los estudios que se están desarrollando actualmente sobre el contenido y estructuras de las páginas Web. En este sentido Cronin y McKim, mencionan que a medida que la Web se va convirtiendo en un medio cada vez más importante para la comunicación de la ciencia y la educación es lógico que los estudios cuantitativos se enfoquen en este medio también. Todo esto demuestra que la Web es un fértil campo de investigación para la bibliometría, cienciometría e informetría [9].

Ingwersen y Christensen, relacionan la webmetría con el uso de aplicaciones bibliométricas tradicionales en toda la Web. Para estos autores el nuevo método tiene como objetivo investigar los modelos de comunicación, la identificación de áreas de investigación, los estudios históricos sobre el desarrollo de una disciplina o campo de la investigación y la evaluación por países, instituciones o individuos [10].

Lennart Björneborn, define webmetría como "el estudio de los aspectos cuantitativos de la construcción y uso de recursos de información, estructuras y tecnologías de la Web, utilizando enfoques bibliométricos e informétricos” [11].

1.3.1. Otras denominaciones de Webmetría.

Esta nueva área de estudios también ha sido presentada por otros autores con distintas denominaciones. De acuerdo con Cronin, Elisabeth Davenport utilizó el vocablo influmetría para referirse al estudio cuantitativo de las líneas imperceptibles y difusas de la influencia académica en el medio electrónico, expresadas en notas de agradecimiento o reconocimiento incluidas en publicaciones, trabajos e investigaciones científicas [12]. En otro artículo, utilizando esta misma terminología, abusan de la relación entre autores, agradecimientos y citaciones en el contexto de la validación académica, relación esta denominada “triángulo de reconocimiento”.

Shiri, por su parte, cita un estudio realizado en 1997 por La Real Escuela de Biblioteconomía, cuyo objetivo era analizar cuantitativamente la página web danesa, bajo el nombre de internetmetría [13]. Ya en 1995 se utilizó el término netometría para referirse a la aplicación de cienciometría de la Web. En encuestas realizadas por los investigadores Almind y Ingwersen en 1996 y otra de Rostaing y Quoniam en 1997 el término adoptado para referirse a estudios similares a los realizados por La Real Escuela de Biblioteconomía fue internetmetría. Otra expresión utilizada fue Bibliometría web, por Chakrabarti en el año 2002.

Es importante señalar que los mismos objetos investigados por webmetría se han estudiado también en otras áreas, un ejemplo de esto es la ciencia Informática o de la computación como se le conocía anteriormente. Por esta razón, una serie de enfoques han surgido desde mediados de los 90 con nombres como "Web de Ecología, "Inteligencia Web", "Exploración Web" y "Análisis gráfico de la Web".

Thelwall afirma que "la razón de ser de la webmetría es denotar una herencia en la bibliometría y la informetría destacando su visión general de la Ciencia de la Información para el estudio de la Web" [14]. Este término es el que mejor expresa la naturaleza de los estudios cuantitativos aplicados a la Web y además, es el que ha adquirido mayor difusión en la literatura internacional producida sobre el tema.

1.3.2. Indicadores Webmétricos.

La Web desempeña un papel cada vez más importante en la investigación. La aparición de la Internet ha llevado a cambios en el proceso de la publicación académica y la comunicación, en la manera científica y académica de buscar y encontrar información. El modo de ser exacto de la interacción entre la nueva información y las tecnologías de la comunicación y la investigación científica y académica muchas veces no es muy claro.

El desarrollo de indicadores de Internet es un área de investigación cada vez mayor. La primera conferencia de Webmetría tuvo lugar en Roorkee, India, por Informetrics and Scientometrics & Fifth COLLNET en una reunión en el 2004. La Unión Europea ha reconocido la importancia de la Webmetría mediante la financiación de dos grandes proyectos:

- Web de Indicadores de Ciencia, Tecnología e Innovación de Investigación (WISER por sus siglas en inglés).
- Indicadores europeos, Sistema de Economía de la Ciencia-Tecnología y el Ciberespacio (EICSTES por sus siglas en inglés).

Otros proyectos han producido mediciones adicionales a la Web, incluyendo:

Los indicadores estadísticos de evaluación comparativa de la Sociedad de la Información (SIBIS por sus siglas en inglés) en la Unión Europea y el Centro Online de Librerías de Computadoras (OCLC por sus siglas en inglés) en Estados Unidos.

Los indicadores webmétricos se pueden dividir en cuatro grupos fundamentales: de conectividad, impacto, densidad y descriptivos. Los descriptivos son los encargados de contabilizar el tamaño o número de objetos en un espacio Web, ya sean páginas, archivos o vínculos y se utilizan para medir el nivel de aceptación que posee determinado espacio Web en países, regiones, organizaciones o grupo de persona con respecto al contenido que posee el mismo.

Las medidas de conectividad, impacto y densidad están estrechamente relacionadas con la forma de hipertexto que posee la Web, teniendo como objetivo examinar las conexiones entre las páginas y sitios a través de sus enlaces. Las medidas de densidad específicamente tienen como propósito estimar cuanto una población se relaciona entre si dentro de una red o comunidad virtual, a partir del número máximo de posibilidades de relacionamientos. Estos últimos indicadores son de gran utilidad para los estudios comparativos.

Algunos de los indicadores utilizados para la realización de estudios webmétricos son: Tamaño de los Sitios Web, Visibilidad o Popularidad, Factor de Impacto Web, Luminosidad, Densidad Media por Links, Densidad de Red, entre muchos otros.

Tamaño de los Sitios Web

Existen dos formas de medir el tamaño de los Sitios Web: la primera; para obtener el tamaño de un sitio Web se realiza un cálculo utilizando la suma de todas las páginas que forman parte de un mismo domino

sin importar el formato, ya sea HTML o cualquier otro. Este indicador es importante para determinar Ranking de páginas Web o para el cálculo del Factor de Impacto Web. El segundo; es obtener el tamaño de los Sitios Web por el número de bytes que contiene, cálculo adoptado por los autores Almind y Ingwersen en su artículo Análisis informétricos en la World Wide Web: enfoques metodológicos para Webmetría, de 1997. Sin embargo, en dicho trabajo se aprobó el primer criterio o forma de calcular el tamaño de los sitios Web.

Visibilidad o Popularidad

La visibilidad es un atributo de un sitio que indica cuan "visible" o cuán bien posicionado está en las listas de salida de los motores de consulta cuando se consulta por temas relevantes a él. Para calcular este valor se utiliza el comando *búsqueda avanzada* en los motores de búsqueda tales como Google, Yahoo! o MSN y una expresión de búsqueda que dependiendo del motor que se utilice incluye operadores lógicos, signos gráficos, etc.

En trabajos empíricos que incluyen la visibilidad de los sitios Web es de preferencia por lo general, el uso de Alta Vista por ser un motor de búsqueda que ofrece los mejores operadores de delimitación a la hora de filtrar y contabilizar los vínculos. Este es el caso por ejemplo de investigaciones llevadas a cabo por Alastair G. Smith y el proyecto Indicadores Europeos, Ciberespacio y el sistema de Economía-Ciencia-Tecnología (EICSTES por sus siglas en ingles), proyecto que logró con éxito la tarea de promover una estructura integrada de conceptos derivados de internet, basados en indicadores de actividades científicas.

La popularidad o visibilidad es una forma para saber en qué medida un sitio es popular o no y se obtiene mediante el cálculo de la tasa de visibilidad en línea, incluido el número de visitas recibidas y la presencia de un sitio en la Web en comparación con sus competidores (cuando es un sitio comercial). Estos datos ayudan a determinar la importancia de un sitio, su tránsito en la Web y la posición que alcanza en los principales motores de búsqueda tales como Google, Yahoo!, Alta Vista o MSN.

Factor de Impacto Web (FIW)

Según Thelwall, el factor de impacto es en esencia el número de páginas dirigidas a determinado sitio o área en internet dividido por el número de páginas en este sitio o área [15]. De acuerdo con el artículo publicado por Nadia Vanti en el año 2002, relacionado con el tema el factor de impacto Web puede ser representado por la siguiente fórmula:
FIW = número de páginas que conectan (enlazan) determinado sitio / número de páginas del sitio conectado (enlazado) [16].

Este indicador sirve para medir y comparar el atractivo y la influencia que pueden alcanzar los diferentes sitios Web. La dinámica que posee la red sugiere que la medición del factor de impacto Web puede ser útil para complementar las mediciones tradicionales. Permite reconocer el grado de reconocimiento que poseen los países o sitios de investigación en la Web en un determinado período de tiempo.

Básicamente existen dos tipo de factor de impacto Web los externos que representan el número de páginas visitadas fuera del sitio Web que está siendo analizado y los internos que representan los vínculos dentro del mismo sitio Web que está siendo analizado.

Es importante resaltar las palabras de Ingwersen en 1998 cuando expresa que “en comparación con citas en revistas científicas, instituciones o personas, que pueden ser estables o aumentar; el número de vínculos que se refieren a un objeto en particular en la Web puede disminuir y hasta desaparecer, esto es a causa, muchas veces del cierre o la restructuración de páginas que estaban disponible en la red y que han cambiado o ya no existen. Lo que imposibilita en esos casos el cálculo retrospectivo del factor de impacto Web” [17].

Luminosidad

Luminosidad puede ser definido como el número de vínculos externos que posee un sitio Web, apuntando hacia otras URLs que por lo general son instituciones similares. Este indicador mide el grado de conectividad en la Web. Puede ser utilizado también para comparar los sitios con enlaces al resto de la Web.

Densidad Media por Link

Como explica Almind y Ingwersen la densidad media por link consiste en la relación que puede ser establecida entre los números de páginas de un sitio Web y la cantidad de enlaces que este sitio posee como un todo (esto incluye todos los enlaces, es decir, externos e internos). Al realizar la división entre el número de todos los enlaces y el número total de páginas de un sitio se obtiene el número medio de enlaces por cada página, este resultado corresponde a la densidad media por link o enlace. Esta medida proporciona la reunión o la estandarización en un solo valor dos datos: el tamaño de página y la cantidad de enlaces [18].

Densidad de Red

Para calcular la Densidad de una Red se dividen el número de enlaces de la red por la suma total de nodos multiplicado por el mismo número menos 1(dado que no se tienen en cuenta las relaciones entre los mismos nodos). La relación entre el número de conexiones efectivas entre los pares de nodos (enlaces) y el número máximo de conexiones posibles entre los pares de nodos reflejan la Densidad de la Red. Un ejemplo en una red con 10 nodos la máxima posibilidad de combinación será de 90 combinaciones, si de estas 90 posibilidades solo 18 son establecidas, la densidad de la red será de 0.20 o 20%. [19]

1.3.3. Aplicaciones de la Webmetría.

La Webmetría dirige su atención a cubrir las metodologías y los resultados de las investigaciones bibliométricas, cienciométricas e informétrica, con énfasis en los aspectos relacionados con la Web.

Según Thelwall, “Webmetría son los aspectos cuantitativos de la construcción y uso de la Web, incluyendo cuatro áreas principales de investigación: *análisis de contenido de páginas web, análisis de la estructura de enlaces*, *análisis de uso web* (explotación de los programas que el comportamiento de registro de búsqueda y de búsqueda Web) y el *análisis de las tecnologías Web* (incluido el rendimiento de los motores de búsqueda)” [20].

También existe la posibilidad de realizar un estudio webmétrico utilizando formas híbridas, explorando las técnicas de análisis web para la categorización automática, utilizando la gráfica de la topología de los enlaces y el contenido de la similitud de texto y metadatos, así como el uso de los datos.

Entre las medidas que se pueden realizar en el ámbito de la Webmetría podemos encontrar por ejemplo, en cuanto a la distribución de frecuencias de páginas en el ciberespacio, esta medida indica el estudio o análisis de la comparación de la presencia de varios países en la red, la proporción de páginas personales, comerciales e institucionales. Como destaca Almind y Ingwersen, es importante que las clasificaciones pueden ser establecidas a partir del tipo de páginas, que puede medir el peso de la red pública y privada [18]. Tarea que es más fácil de ver cuando los nombres de dominio son: .Edu y .Com.

Según estos mismos autores se pueden hacer valoraciones más profundas, como son las clasificaciones para entender categorías como páginas personales, institucionales u organizacionales, páginas de índice cuya función principal es poner a disposición una serie de link o hipervínculos y por último las páginas recurso que no son más que aquellas que proporcionan datos en texto, sonido o imagen.

En la Web se hace posible así como en el formato tradicional evaluar el grado de cobertura de determinado tema. Se pueden utilizar ambos formatos para realizar un estudio comparativo con el fin de determinar en qué medida se superponen o coinciden en el alcance de ambos. Sin embargo, su diferencia no radica en las formas de acceder a los mismos, ni en sus políticas de actualización sino en la función de los medios de comunicación. No siempre la red refleja con plena fidelidad la situación, los avances o procesos que experimentan una institución o centro de investigación, ni los cambios a un sujeto, tema o asunto fuera de la Web. De hecho, hay zonas donde la visibilidad es mayor en la Web y otros en los que una mayor visibilidad existe en formato tradicional o en copia dura como también se le llama [19].

Las técnicas de medición pueden aplicarse a un grupo fundamental de 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 Páginas 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 período de tiempo determinado.
- El número de citas recibidas por cada Página Web.
- Ordenar las Web más citados 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 [6].

1.4. Herramientas Utilizadas para Realizar Estudios Webmétricos.

Los estudios webmétricos también requieren ciertos mecanismos para realizar la búsqueda, extracción, cuantificación, representación y visualización de la información disponible en la Web. Los motores de búsqueda, los programas cartográficos y programas para la representación y visualización de las redes, son los métodos utilizados para este propósito.

1.4.1. Motores de Búsqueda.

Como afirma Smith, el principal instrumento para realizar estudios webmétricos son los motores de búsquedas ya que permiten trabajar con grandes volúmenes de información. También señala que estos buscadores permiten contar el número total de páginas en un espacio Web así como el número total de enlaces a esos espacios, entiéndase también por el término *espacio Web* si el dominio es de un país o institución [21].

También favorecen la investigación de los link o la relación entre documentos, de manera que se puede establecer una analogía entre el análisis de los hipervínculos y el tradicional análisis de citas en formato duro. A través de motores de búsqueda es posible contar el número de subdominios, su visibilidad, el factor de impacto y el rango de página de la mayoría de los sitios más visitados.

1.4.2. Programas Mapeadores.

Considerado como método de segunda generación basan su unidad de análisis es más pequeña que la de los motores de búsqueda haciendo más difícil el trabajo cuando se trata de un gran volumen de información. Este tipo de programa permite complementar los datos obtenidos a través de los motores de búsqueda y llegar a etapas más avanzadas de cuantificación, abriendo nuevas opciones para el trabajo analítico. Según Aguillo, estos programas muestran la información relativa al tamaño y número de sitios, el tipo de recurso que contienen y también son útiles para calcular la luminosidad de los sitios web [22].

Existen grupos y proyectos que utilizan estas herramientas para realizar sus propios estudios, ejemplo de esto es el Grupo de Investigación de Cibermetría Estadística (SCRG, por sus siglas en inglés) con sede en la Universidad de Wolverhampton, Reino Unido, dedicado entre otras cosas, al desarrollo de software y metodologías para explotar los recursos disponibles en Internet que pueden ser utilizados en la investigación que se enfoca más bien en las ciencias sociales.

El grupo, dirigido por Mike Thelwall, uno de los investigadores más reconocidos en el campo de la Webmetría / Cibermetría, es responsable de la base de datos del Proyecto Link. El principal objetivo de este proyecto es proporcionarles a los investigadores herramientas que pueden utilizar para realizar un análisis en la estructura de enlaces o link. Recursos desarrollados en este proyecto se encuentran gratuitamente en el sitio del grupo (http://cybermetrics.wlv.ac.uk/).

Otro importante mapeador que se utiliza para la medición webmétrica y cibermétrica es el Sitio de Analistas y Analizador de Contenido, programa comercial que ofrece Microsoft para el paquete Back Office basado en el viejo programa de Webmapper, funcionando muy bien para el análisis de las páginas dinámicas. Este fue una de las herramientas utilizadas por el equipo del Centro de Información y Documentación Científica del Consejo Superior de Investigaciones Científicas en España (CINDOC/ CSIC), para el estudio de 4000 sitios Web universitarios españoles, con el propósito de calcular una serie de indicadores cibermétricos importantes.

Existen otras herramientas tanto académicas como comerciales que han sido empleadas por instituciones públicas y privadas con el objetivo de evaluar sus sitios Web. En un estudio realizado por Arroyo (2004, online) la investigadora profundiza en las características y aplicaciones de algunos de estos programas, realizando también una comparación entre ellos.

Entre los software analizados se encuentran: Astra SiteManager, COAST WebMaster, Funner Profiler, Webcount, WebKing Lite, Web Treendes y Xenu Link Sleuth. Todos mostrando un empate en la mayor o menor sofisticación, información sobre la estructura, contenido y semántica de los sitios, presentando estadísticas que ayudan en la contabilidad y la corrección de problemas que puedan surgir eventualmente.

1.5. Dificultades para la Realización de Estudios Webmétricos.

Según lo expresado por Bar-Ilan y Olvera Lobo, los documentos desaparecen, ocurren cambios continuos, nuevas páginas relevantes son constantemente agregadas y los buscadores demoran un tiempo hasta incorporar tales cambios, tornando más difícil el proceso de análisis e indexación de estas páginas en la red [23] y [24]. A esto, se suma el problema planteado por Lynch en el año 1997, derivado del carácter mutante de la propia estructura de muchas páginas, las cuales no trabajan con archivos estáticos, sino con contenidos que varían con alta frecuencia, como es el caso de revistas electrónicas o de bases de datos interactivas, o que constituye un inconveniente más para el análisis y la cuantificación de estos sitios.

Según Almind e Ingwersen, una solución para los problemas y dificultades para encontrar lo que se está buscando en internet, es usar las bases de datos indexadas de la (www) [18]. De cualquier forma, la indexación y cobertura de estos es muy desigual. No existe la estandarización de la información, cada autor escoge caminos diferentes.

Con respecto a los motores de búsqueda, estos presentan una serie de dificultades lógicas que entorpecen la medición de los datos contenidos en sus bases de datos. Como explica Judit Bar-Ilan y Thelwall, los motores acostumbran a perder información, existen URLs recuperadas en cierto momento por un determinado motor de búsqueda no son encontradas por ese mismo motor algún tiempo después (aunque sigan existiendo). También el contenido muchas veces se pierde, ya que las URLs recuperadas una segunda vez no contienen la misma información que la primera vez. Además cuando son utilizados varios motores de búsqueda y comparados entre sí para evaluar su funcionamiento, se puede percibir que la sobre posición de los resultados mostrados por ellos es sumamente pequeña, pudiendo afectar de algún modo la confiabilidad de los análisis webmétricos [15] y [23].

Las contradicciones van más allá de la incapacidad de contabilidad, incluidos los problemas relacionados con el procesamiento de la sintaxis de búsqueda. La transformación puede conducir a resultados erróneos debido a que los recursos utilizados para desarrollar estrategias de búsqueda – truncamiento, búsqueda por campos y operadores booleanos no siempre funcionan satisfactoriamente, menos aún cuando se usan en combinación.

Para realizar análisis webmétricos satisfactorios, es necesario que a la hora de escoger el motor de búsqueda se deben tratar de elegir motores que reúnan las siguientes características o que la combinación de más de un motor de búsqueda cumplan estos criterios:

1. Contar con una base de datos actualizada, que incluyan las páginas nuevas y que elimine las que ya no existen o se encuentren fuera de servicio.
2. Cubrir la mayor proporción de Web posible.
3. Ser capaz de delimitar la búsqueda por dominios.
4. Ofrecer la posibilidad de recuperar las páginas que contengan enlaces para un sitio Web particular.
5. Permitir la combinación de resultados de búsqueda con operadores booleanos para por ejemplo, contar el número de páginas que se encuentran enlazadas a un sitio Web particular excluyendo sus enlaces internos. [21]

1.6. Tecnología a Utilizar.

Los programas o herramientas de software que se utilizan para los estudios de la Web, conocidos como *agentes o robots web, spiders, wanderers o, incluso, gusanos (worms)*, muchas veces no están disponibles para su uso pleno. Un ejemplo de esto son varios servicios de búsqueda en Internet, cada uno de los cuales tiene su propio spider o robot web; pero solo se puede contar con los datos que los mismos facilitan sin ir más profundo, lo cual muchas veces no resuelve el problema al cual se desea dar solución. Por otro lado, el valor económico de muchos de estos servicios hace que, en ocasiones, la información que se puede obtener de ellos sea bastante pobre acerca de detalles concretos sobre el funcionamiento de muchos de ellos [6].

Algunas de las herramientas que se utilizan para realizar estudios webmétricos son:

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).

Dentro de las herramientas libres existentes para lograr el mismo propósito, es decir, estudiar el comportamiento de la Web se encuentran las siguientes:

- **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*/1.18 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-type9 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 swordview, 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, 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 él 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. Una 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 opera 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”.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. 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.

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 otras herramientas no se lograría de manera eficiente. 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. [2]

La herramienta seleccionada por las autoras del presente trabajo es el **WIRE Crawler** por ajustarse a las necesidades del estudio Web de la UCI. Además para compatibilizar los resultados de este estudio, los anteriores que se han realizado en la UCI y lograr posteriormente la comparación de los mismos.

1.7. Conclusión.

En este capítulo se abordaron los principales conceptos relacionados con el objeto de estudio, realizándose un breve recorrido por esta joven ciencia que es la webmetría: encargada de estudiar el comportamiento de la Web. Exponiendo sus fronteras, características, herramientas para su desarrollo, ventajas y principales definiciones, además de un pequeño recorrido por los estudios webmétricos realizados a nivel mundial, incluyendo los realizados en nuestro país y principalmente los realizados en la UCI.

Capítulo 2: Caracterización, pronósticos y tendencias sobre el comportamiento de la Web en la UCI.

2.1. Introducción.

En el presente capítulo se exponen los resultados obtenidos en el quinto estudio webmétrico, que aunque existen ya varios estudios en la UCI, es importante tener información actualizada para realizar una posterior comparación de los resultados obtenidos con los resultados ya existentes de los estudios anteriores. Finalmente se proponen pronósticos y tendencias a seguir por la Web de la UCI en tiempos futuros.

2.2. Nivel Colección.

Actualmente la Web de la UCI está compuesta por aproximadamente 156 sitios, que contienen cerca de 1 200 000 páginas. Aunque solo se pudieron descargar 1 042 571. La siguiente tabla muestra datos importantes en cuanto a la distribución de las páginas Web analizadas, aunque estos datos serán analizados con más profundidad en el nivel correspondiente a las páginas.

Total de páginas descargadas	1 042 571	
Páginas únicas	977 536	93.76%
Páginas duplicadas	65 035	6.24%
Páginas estáticas	303 324	29.09%
Páginas dinámicas	739 247	70.91%

Tabla 1. Datos generales del quinto estudio webmétrico.

2.2.1. Enlaces a dominios externos.

De acuerdo al análisis realizado después de culminado el estudio, se pudo detectar 120 dominios externos al dominio de la universidad (.uci.cu), dentro de los cuales los que más prevalecen son: .CU (Cuba) con 11 528 448 enlaces en nuestra web, .ORG (Organizaciones) con 1 015 764, .COM (Comerciales) con 587 041, .NET con 65 802, representando un porciento de 87.12; 7.68; 4.44; 0.50 respectivamente del total de enlaces a dominios externos. También estos enlaces conducen a páginas web, ficheros de textos, multimedia entre otras extensiones.

Dominio	Cantidad de Enlaces	Porcentaje (%)
CU - Cuba	11 528 448	87.12
ORG	1 015 764	7.68
COM	587 041	4.44
NET	65 802	0.50
.GOV – Gobierno	13 272	0.10
ES - España	6 337	0.05
AR - Argentina	2 478	0.02
DE - Alemania	1 622	0.01
RU – Federación Rusa	1 381	0.01
US – Estados Unidos	1 326	0.01

Tabla 2. Dominios externos referenciados en páginas de la UCI.

Los elevados números que muestran la cantidad de enlaces a otros dominios fuera del uci.cu en gran medida representan el alto nivel de acceso que poseen los usuarios de la universidad a los grandes

volúmenes de información que son necesarios tanto para la docencia así como para la investigación y la producción que se llevan a cabo en el centro.

2.2.2. Software utilizado como servidor Web.

Al concluir 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.

En algunos casos la información es bastante completa, incluyendo el nombre del servidor, la versión del software instalada, el sistema operativo utilizado y las extensiones instaladas; no obstante, muchas veces la respuesta no contiene toda la información requerida. A continuación se muestran dos gráficas que representan los tipos de servidores Web utilizados y los sistemas operativos respectivamente.

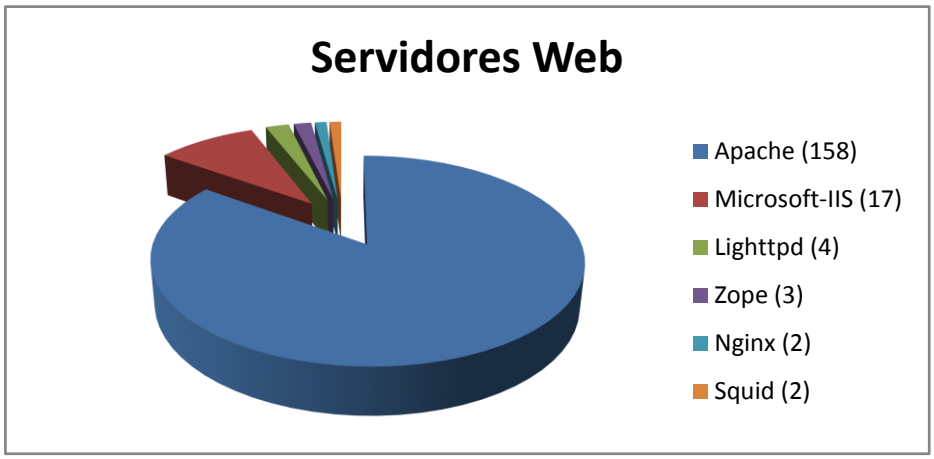

Figura 6. Distribución de servidores web por dirección IP.

Figura 7. Distribución de sistemas operativos por dirección IP.

La información reflejada en las gráficas da fe de la persistencia entre los servidores Web del Apache, usado en más de los 150 sitios Web encontrados con este tipo de servidor, ubicándose por tanto nueve veces por encima de los servidores de Windows (Internet Information Server) encontrando solamente 17 sitios que utilizan este tipo de servidor.

En la última gráfica se muestra de forma general la representación del sistema operativo Linux, pero las distribuciones de dicho sistema que se encontraron en la Web son: Debian, Ubuntu, CentOS y Red Hat que en conjunto suman 165 direcciones IP que utilizan Linux como sistema operativo. Dejando solo 17 direcciones para los sistemas operativos Windows y solamente 4 que no está disponible el sistema operativo sobre el cual está trabajando. Al observar los datos y la gráfica es evidente el aumento en la comunidad universitaria del software libre; alternativa por la cual la universidad y todo el país está optando.

2.2.3. Sitios Web por dirección IP.

Es común encontrarse varios sitios Web hosteados en una misma dirección IP, en la siguiente tabla se muestran las distribuciones de los sitios Web por dirección IP encontradas.

Dirección IP	Cantidad de sitios	Dirección IP	Cantidad de sitios

10.0.0.8	1	10.0.0.91	2
10.0.0.9	2	10.0.0.170	1
10.0.0.10	6	10.0.0.186	2
10.0.0.11	4	10.128.60.118	1
10.0.0.12	43	10.0.0.213	1
10.0.0.13	6	10.3.10.41	3
10.0.0.210	30	10.3.10.45	5
10.0.0.214	1	10.3.10.47	1
10.0.0.222	3	10.128.60.4	1
10.0.0.16	1	10.128.60.7	1
10.0.0.17	1	10.128.50.221	1
10.0.0.22	3	10.208.0.6	23
10.0.0.48	1	10.208.0.9	2
10.0.0.49	1	10.209.0.11	2
10.0.0.70	1	10.209.0.22	5
10.0.0.77	2	10.209.0.9	3
10.209.0.18	1	10.209.0.20	2

10.209.12.2	1	10.210.0.4	1
10.210.0.6	5	10.128.50.121	1
10.128.60.4	5	10.128.60.112	1
10.210.0.20	1	10.210.0.8	1
10.208.1.25	1	10.208.1.24	2
10.128.60.4	1	10.209.0.22	1

Tabla 3 . Distribución de sitios Web por dirección IP.

Esta tabla refleja que un total de 184 sitios Web poseen dirección IP conocida sumando un total de 46 direcciones, de estas 10 direcciones son nuevas y además la dirección de mayor cantidad de sitios es (10.0.0.12) con 43 sitios hosteados. Existen también 24 direcciones con un solo sitio Web, entre 2 y 6 sitios Web hay 19 direcciones IP, con 23 y 30 respectivamente se encontraron 2 direcciones (10.208.0.6) y (10.0.0.210). Vale señalar que las direcciones IP de mayor número de sitios Web hosteados son direcciones conocidas y que por los resultados de este y los otros estudios son bastante usadas en la Web de la UCI.

2.3. Nivel Sitios.

En este epígrafe se expondrá toda la información relacionada con los sitios Web, algunas de las características que se detallarán más adelante, se encuentran a continuación.

Total de Sitios OK	156
Promedio de Enlaces Internos	10 320 170
Promedio de Páginas Web por Sitio	768 751

Promedio de Páginas Dinámicas	531 989
Promedio de Páginas Estáticas	236 762
Tamaño promedio en MB	12 635
Promedio de Profundidad Máxima	592
Promedio de Grado Interno	283
Promedio de Grado Externo	283

Tabla 4. Datos generales del nivel Sitios.

2.3.1. Tamaño total de la colección de la información analizada.

La colección alcanzó un tamaño total de 36.5 G, dentro del mismo se encuentran el 19,7 G que es el total de texto plano descargado, además de los ficheros generados y los reportes que recogen toda la información necesaria para realizar el análisis de la información y datos de la Web. Toda esta información compone el tamaño total de la colección alcanzado en este último estudio que supera a los cuatro anteriores en más de 10 G.

2.3.2. Tamaño promedio de los sitios en MB.

El tamaño aproximado de un sitio web promedio en la UCI es de 126,35 MB; esta cifra se ve claramente como aumentó con respecto a la cifra del cuarto estudio que fue de 102,41 MB, y esto se debe en gran medida a la cantidad de sitios que se analizaron, 156 en total, 16 más que el estudio anterior. Hay sitios en estos momentos en la universidad que se actualizan y aumentan su información constantemente, lo cual permite que crezca de inmediato su capacidad.

2.3.3. Distribución de páginas Web por sitios.

A continuación se muestra una tabla, la cual presenta una serie de sitios, que resultaron ser los que poseen mayor cantidad de páginas web descargadas, entre ellos se encuentran: mirror.prod.uci.cu, gforge.f10.uci.cu, ubuntu.uci.cu, debian.uci.cu, entre otros.

La gran mayoría de estos sitios, utilizan CMS (Content Management Systems) en español Sistemas de Gestión de Contenidos, los cuales están constituidos por páginas dinámicas de contenidos, y esto hace que crezca el número de páginas que presenta el sitio. Este aumento viene dado en parte por la facilidad que brindan los CMS para la publicación de nueva información, o la modificación de la ya existente. Su dinamismo es otro factor importante, pues permite editar los contenidos mediante el navegador, sin la necesidad de hacerlo en la máquina donde está montado el servidor.

Dirección de los Sitio	**Cantidad de Documentos**
mirror.prod.uci.cu	150 090
gforge.f10.uci.cu	150 000
ubuntu.uci.cu	128 661
facultad7.uci.cu	124 969
debian.uci.cu	113 086
softwarelibre.uci.cu	98 420
cpav.uci.cu	79 797
drupaleros.uci.cu	79 585
onlinejudgef8.uci.cu	47 337
servicio.hab.uci.cu	28 191
comunidades.uci.cu	25 176
gpi.uci.cu	21 341

ucipedia.uci.cu	17 160
php.uci.cu	16 467
softwarelibre.hab.uci.cu	15 729
primavera.uci.cu	10 654
portal.centalad.prod.uci.cu	9 050
postgresql.uci.cu	8 690
comedor.hab.uci.cu	8 154
seriecientifica.uci.cu	7 783

Tabla 5. Sitios con mayor cantidad de páginas Web descargadas.

Existen 40 sitios que solo presentan una sola página de contenido, en la mayoría de los casos estas páginas cumplen la función de re-direccionar a otros sitios, o mostrar información estática. O sea se puede decir que en la universidad hay sitios que no presentan mucha información que mostrar al usuario, y que solo poseen una o dos páginas en el mismo. Al existir un gran número de sitios que presentan pocas páginas de contenido, esto trae consigo un consumo innecesario de los servidores disponibles en la Universidad.

2.4. Nivel Páginas.

En este último epígrafe se abordarán datos detallados de las páginas descargadas. Esta información es observada independientemente de los sitios analizados, es decir, como elementos que al igual que los sitios en si contienen información de ellas en sí. Algunas de las características que se abordan son la edad, profundidad, idioma, extensiones, entre otras.

2.4.1. Cantidad de páginas únicas/duplicadas de la colección.

Aunque ya estos datos fueron mostrados anteriormente, en esta sección se profundizará en el significado que tiene cada uno de ellos. Como resultado de este quinto estudio se descargaron 977 536 páginas únicas que representan el 93.76% del total de páginas descargadas. Dejando solamente el 6.24% que representa 65 035 de las páginas duplicadas.

Los datos arrojados en el estudio pueden tomarse como información positiva o negativa, depende solamente de qué forma se desee interpretar, en el caso de las páginas duplicadas si la información que muestran tiene un alto grado de importancia para toda la comunidad universitaria pues entonces es lógico que se tome positivamente la duplicación de las páginas que poseen dicha información pero si no es tan importante tener una información determinada en varias páginas para que la comunidad universitaria la conozca entonces las páginas duplicadas se toman como páginas negativas que solamente sobrecargan la red.

2.4.2. Cantidad de páginas dinámicas/estáticas de la colección.

El promedio de páginas estáticas es de 2 367 62 y de páginas dinámicas 5 319 89, representando un 30. 79 y un 69.20 por ciento respectivamente del total de páginas web. En este último estudio se sigue evidenciando la presencia de las páginas dinámicas en la Web como las más representativas, esto se debe a que estas páginas poseen una mayor aceptación dentro de los usuarios que visitan las páginas o los sitios Web existentes en la red de la UCI. De forma general, se refleja también el nivel de avance alcanzado desde la Web 1.0 hasta la Web 2.0, que esta última incorpora entre sus mejoras el desarrollo de herramientas de colaboración como son: wikis, blogs, entre otros.

2.4.3. 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 clic en un vínculo, para llegar a la misma sin abandonar el sitio web y comenzando desde la portada del mismo. 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 clic y sencillamente abandona el sitio web [25].

En la tabla de profundidad que generó el spider se encuentran profundidades desde la 1 hasta la 48, pero solamente entre la 1 y la 10 se encuentran 1 041 965 páginas Web que representan el 99.94% de la colección analizada, esto es algo muy positivo en la Web ya q cuando se realizaron los primeros estudios estos datos no era ni semejante a los resultados actuales. Entre las profundidades 11 y 48 se encuentra el resto de las páginas Web equivalente a 606 para representar el 0.04%. Es notable que a medida que aumenta la profundidad disminuye la cantidad de páginas Web.

2.4.4. Edad de las páginas de la colección.

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 [25].

En este quinto estudio se evidencia el aumento en la Web de páginas creadas o modificadas recientemente, representando el 96.53%, equivalente a 242 364 documentos presentes en la Web, dejando solamente el 3.47% a las páginas que no han sido modificadas en este último período.

Figura 8. Edad de las páginas de la Web en la UCI.

2.4.5. Idioma de las páginas de la colección.

El idioma predominante en la Web de la UCI sigue siendo el español, ¿a qué se debe esto?, desde el primer estudio webmétrico se ha evidenciado la fuerte presencia del idioma español en la Web, tal vez porque es nuestra lengua materna, pero a pesar de eso también existe gran cantidad de páginas en diversos idiomas siendo el inglés el de mayor predominio, entre esos idiomas se encuentran el italiano, alemán, portugués, francés, entre otros. A continuación se muestra una tabla con la relación de los idiomas encontrados.

Idiomas	Total de documentos	Porcentaje (%)
Español	280 583	95.66
Inglés	8 278	2.82
Noruego	3 896	1.33
Francés	188	0.06
Danés	152	0.05
Italiano	76	0.03
Catalán	58	0.02
Alemán	25	0.01
Irlandés	19	0.01
Portugués	17	0.01
Sueco	6	0
Turco	2	0
Griego	0	0

Tabla 6. Idiomas encontrados en la Web de la UCI.

2.4.6. Extensiones encontradas.

Con el uso de este indicador se puede tener una idea bastante acertada de las extensiones más utilizadas en la Web de la UCI, además de que posibilita conocer o descubrir nuevas extensiones y el nivel de representatividad que poseen los mismos en la Web. Algunas de las extensiones más conocidas son: extensiones de audio, video e imagen; extensiones de entrada común y código fuente; extensiones de software; extensiones fuera del HTML y TXT; extensiones de ficheros comprimidos; otras extensiones y extensiones desconocidas hasta este momento.

Extensiones de imagen, video y audio

En la Web de la UCI se encuentra como extensión de video predominante .MOV con 3621 ficheros que representa el 53.46% del total de extensiones de video encontradas, seguido del .DAT con 2229 ficheros y

un 32.91% de presencia en la Web. Estas extensiones son las más representativas de la Web, aunque también aparecen .FLV; .WMV; .AVI; .MPG; .MP4 y .QT respectivamente. Ver anexo 1.

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, la compañía que fabrica las computadoras "Imac", 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 [25].

Las extensiones de imagen han sido prácticamente las mismas desde el primer estudio webmétrico realizado en la UCI, desde .PNG hasta .PBM pasando por .JPG; .GIF; .ICO; .IMG; .BMP y .WMF. Aunque hasta el cuarto estudio la extensión predominante de imagen era la .GIF en el quinto estudio esta extensión fue desplazada a un segundo puesto en la tabla de las extensiones más usadas en la Web, siendo la .PNG la extensión de mayor presencia con 8 198 093 ficheros para un 46.84% del total de extensiones encontradas y dejando la .GIF con 7 295 092 ficheros que representa un 41.68%. Ver anexo 2.

PNG (*Portable Network Graphics*) es un formato gráfico basado en un algoritmo de compresión sin pérdida para bitmaps o mapas de bit no sujeto a patentes. Este formato fue desarrollado en buena parte para solventar las deficiencias del formato GIF y permite almacenar imágenes con una mayor profundidad de contraste y otros importantes datos [26]**.** Al hacer doble click sobre un archivo con extensión .*P*NG se abre con el Visor de imágenes y fax de Windows. Al hacer click con el botón derecho del ratón y elegir Editar, se edita con Paint.

A pesar del aumento de ficheros de audio .OGG y luego de haber sido la extensión de audio más usada en la Web según los resultados arrojados por el cuarto estudio webmétrico, en el quinto estudio la extensión .MP3 es la que vuelve a prevalecer en la Web como la más usada con 759 ficheros que representan un 91.34% tomando gran diferencia sobre la extensión .OGG que solo posee el 7.34% de representatividad.

MP3 (MPEG Audio layer 3) es un formato de compresión de datos de audio con pérdida, desarrollado por la Organización Internacional de Normalización (ISO). Este formato se utiliza para comprimir formatos de

audio normales (WAV o CD o audio) en una relación de 1:12. Permite almacenar el equivalente a 12 CD-ROM de álbumes de música en el espacio de un solo CD. Es más, el formato mp3 casi no altera la calidad del sonido para el oído humano [27].

Extensiones CGI y software

Las extensiones CGI son aquellas de entrada común y código fuente, en el quinto estudio se evidencia la permanencia en la Web de los ficheros .PHP con una representatividad de 19 721 282 para un 99.41% casi el total de los ficheros CGI de la Web en la UCI, luego le siguen .ASP (0.44%), .CGI (0.07%), .PL (0.04%), .JS (0.03%), .JSP (0.01%) respectivamente. Estas son las extensiones que se encuentran con mayor presencia aunque existen otros ficheros como .BIN, .TPL, .PM, entre otros que pueden verse en el anexo 3.

PHP es un lenguaje de programación interpretado, diseñado originalmente para la creación de páginas web dinámicas. Es usado principalmente en interpretación del lado del servidor (server-side scripting) pero actualmente puede ser utilizado desde una interfaz de línea de comandos o en la creación de otros tipos de programas incluyendo aplicaciones con interfaz gráfica usando las bibliotecas Qt o GTK+.

PHP es un acrónimo recursivo que significa **P**HP **H**ypertext **P**re-processor (inicialmente PHP Tools, o *P*ersonal *H*ome *P*age Tools). Fue creado originalmente por Rasmus Lerdorf en 1994; sin embargo la implementación principal de PHP es producida ahora por The PHP Group y sirve como el estándar de facto para PHP al no haber una especificación formal. Publicado bajo la PHP License, la Free Software Foundation considera esta licencia como software libre. [28]

Las extensiones de software no han sufrido muchos cambios en el orden en el cual aparecen en la Web permaneciendo los .DEB con 817 454 que representa 86.88% del total, variando solo en 0.13% de los resultados obtenidos en el cuarto estudio. En la siguiente tabla se muestran todas las extensiones de software encontradas con sus respectivos porcentajes y total de ficheros encontrados.

Extensiones de software	Total de ficheros	Porcentaje (%)

DEB	817 454	86.88
RPM	119 974	12.75
EXE	1 596	0.17
JAR	910	0.10
ISO	570	0.06
DIFF	241	0.03
PATCH	117	0.02

Tabla 7. Extensiones de software.

Extensiones de documentos que no son html y de compresión

Dentro de los documentos que no son html encontrados en la Web se encuentran .PDF, .DOC, .TXT, .PPT, .ODS, .XML, .README, .XLS, entre otros que pueden verse con más detalle en el anexo 4. Pero a pesar de que en los estudios anteriores la extensión .PDF era la de mayor presencia en la Web, los resultados arrojados en este quinto estudio muestran un nuevo monarca dentro de los documentos que no son html, con 277 002 equivalente a 84.67% la extensión .README es la de más presencia en la Web de la UCI dejando a la extensión .PDF con solamente 8.68% de representación.

Mientras navega por la Web, seguramente se encuentra archivos de texto, sonido y vídeo que se pueden descargar. Particularmente, los archivos de multimedia pueden ser muy grandes, lo que significa que pueden desplazarse muy lentamente a través de la red. Descargar estos archivos puede, algunas veces, llevar hasta horas. Para hacer un uso eficiente del espacio y acelerar las cosas, la mayoría de los archivos de gran tamaño están comprimidos. La compresión de archivos puede reducir considerablemente el tamaño y el tiempo de los mismos a la hora de descargarlos. ¿Cómo funciona?

El software de compresión usa ecuaciones matemáticas complejas para buscar en el archivo patrones que se repiten en los datos. Reemplaza los datos con códigos más pequeños que ocupan menos espacio. Por

ejemplo, una manera en la que funciona el software de compresión es reemplazar caracteres que se repiten con un código que también anota la posición de esos caracteres en los datos. Con una imagen, encontraría todas las partes rojas, por ejemplo y las reemplazaría con un código.

Para ver datos descomprimidos, se necesita un programa compatible de descompresión que pueda leer esos códigos y convertir los datos a su forma original. La mayoría de los archivos que encuentra en la Red son de vídeo, texto, gráficos o sonido. Algunos pueden ser comprimidos, otros no. Los archivos comprimidos más comunes que encuentra en la red son los que tienen extensiones como .ZIP, .SIT y .TAR. Pueden ser archivos unitarios o grupos de archivos que han sido reunidos en un archivo comprimido. Un archivo comprimido, a veces, puede contener archivos de vídeo y gráficos, y muchas veces contiene programas con la documentación relacionada con ellos.

En el quinto estudio webmétrico no se notó mucha variabilidad dentro de las extensiones de compresión existentes en nuestra Web. A continuación se muestra una tabla con las extensiones, total de ficheros y porcentaje representativo en la Web.

Extensiones de compresión	Total de ficheros	Porcentaje (%)
GZ	321 731	97.01
ZIP	5 328	1.61
TAR	2 431	0.73
BZ2	1 420	0.43
RAR	728	0.22
SIT	2	0

Tabla 8. Extensiones de compresión.

Extensiones extras

Dentro de esta categoría se encuentran las extensiones .css con la mayor representación 845 431 y un porcentaje de 90.20% evidenciando el gran uso de las hojas de estilo (CSS) para el diseño de las páginas Web, seguido de .SWF con 91 391 que representa el 9.75% referentes a ficheros flash, también se encuentran .GPG (0.04%), .PSD (0.01%) y .NET_FILES (0%) respectivamente.

Extensiones desconocidas

Las extensiones que aquí se exponen son las que no están en la relación inicial de extensiones existentes que posee el Spider. Existe una gran posibilidad de que muchas de estas extensiones si sean conocidas pero no habían sido identificadas con anterioridad en la Web. En el anexo 5 se muestra la relación de estas extensiones, vale decir que en este estudio se encontraron poco más de 240 extensiones desconocidas.

2.4.7. Código de estado de las páginas descargadas.

El código de estado de las páginas descargadas es la respuesta que brinda la página solicitada, es decir, si existe o no y el motivo por el cual no puede ser entregada la solicitud hecha en caso de que la página no responda. A continuación se muestra una tabla con los diferentes estados de las páginas descargadas.

Estado http	Código http	Documentos	Porciento (%)
OK	200	908 041	87.10
Moved	301	48 898	4.69
Error Connect	97	34 826	3.34
Internal Error	500	24 640	2.36
Found	302	16 719	1.60
Forbidden	403	5 039	0.48
Not Found	404	2 272	0.22
See Other	303	831	0.08
Error DNS	98	517	0.05
Partial	206	330	0.03
Error Timeout	95	238	0.02
No Content	204	59	0.01
Bad Request	400	124	0.01
Unauthorized	401	34	0
Not Acceptable	406	2	0
Unavailable	503	1	0

Tabla 9. Código de estado de las páginas descargadas.

Existe una gran cantidad de códigos de estado, aunque se pueden agrupar de la siguiente manera:
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)**.

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

Para la realización de este último estudio se contó con 300 sitios conocidos pero solamente 156 de ellos contó con al menos una página con estado OK. La SCC más grande es de 53 sobrepasando el cuarto estudio en 14, con SCC-id 184. También hubo una disminución dentro de las SCC con solo un sitio Web existiendo en este estudio solo 95.

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.**
- **Island:** sitios que no son accesibles ni hacia ni desde **Main**.
- **Tentacles:** sitios que sólo se conectan con **In** (**Tin**) u **Out** (**Tout**), pero en el sentido inverso de los enlaces.
- **Tunnel:** una componente que une las componentes **In** y **Out** sin pasar por **Main**.[6]

Nombre de los componentes	Número de sitios	Porcentaje (%)
Main-Norm	18	11.54

Main-Main	6	3.85
Main-In	2	1.28
Main-Out	27	17.31
In	4	2.56
Out	64	41.03
Tin	7	4.49
Tout	4	2.56
Tunnel	1	0.64
Island	23	14.74
Tentacles	0	0

Tabla 10. Componentes de las SCC.

2.6. Pronósticos y tendencias.

A continuación se muestran datos generales de los principales indicadores webmétricos, tomando información de los cinco estudios webmétricos que se han realizado en la Universidad de las Ciencias Informáticas.

Variable\Indicador	**Primer Estudio Webmétrico**	**Segundo Estudio Webmétrico**	**Tercer Estudio Webmétrico**	**Cuarto Estudio Webmétrico**	**Quinto Estudio Webmétrico**
Total de Páginas Analizadas	127 718	458 457	461 831	756 980	1 042 571
Texto Total de la Colección	3 641.76 MB	8 834.68 MB	8 307.75 MB	14 336 MB	19710.6 MB

Texto Promedio por Página	0.029 MB	0.019 MB	0.018 MB	0.019 MB	18.9 KB
Total de sitios web	108	164	159	140	156
Páginas Promedio por Sitio	1 356	3 336	4 414	5 709	7 687
Texto Promedio por Sitio	33.72 MB	53.87 MB	52.25 MB	102.41 MB	126.35 MB

Tabla 11. Resumen de los cinco estudios webmétricos.

A partir de la tabla anterior se puede establecer que en la actualidad la Web de la UCI está compuesta por aproximadamente 300 sitios web, solo se analizaron 156 sitios, que representa un total de 1 042 571 páginas descargadas, los datos antes mencionado fueron del quinto estudio, comparándolos con los estudios anteriores se ve el aumento en cuanto a la cantidad de sitios, pues desde el primer estudio hasta el quinto, se evidencia su crecimiento y a medida que pasen los años, se pronostica que la UCI crezca en este aspecto considerablemente.

2.6.1. Enlaces a dominios externos.

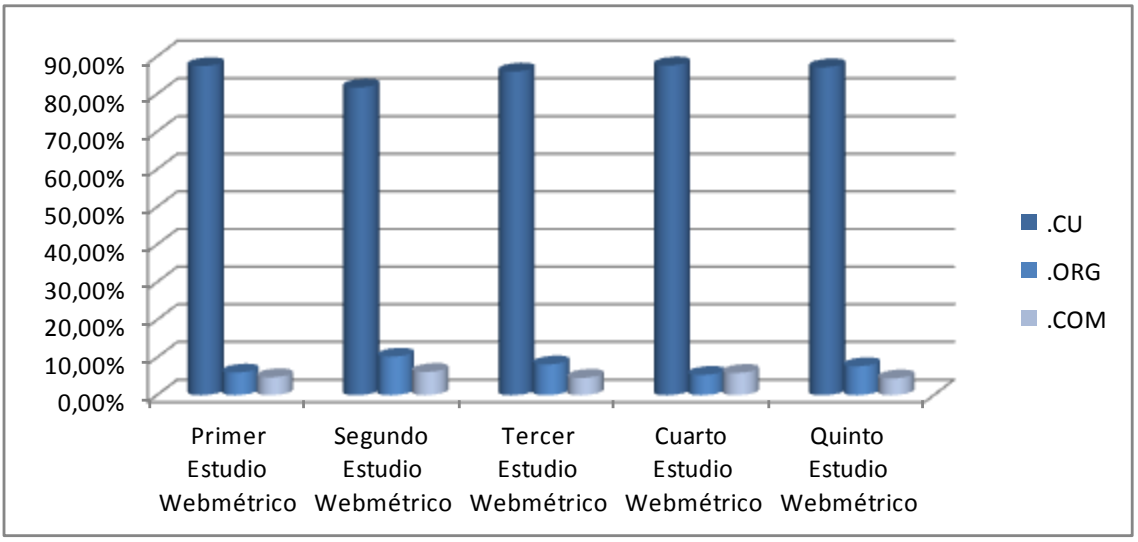

Figura 9. Enlaces a dominios externos de los cinco estudios webmétricos.

Los registros de los enlaces externos muestran como ha habido un aumento de los vínculos al dominio cubano .CU, con respecto a los demás dominios, lo que muestra la confiabilidad de la información que se encuentra presente en los sitios del país. En los próximos años la Web de la universidad pudiera no mostrar alteraciones significativas principalmente en el dominio (.CU), pues el acceso a los sitios cubanos se hace mucho más rápido, las cuentas de todos los usuarios en la universidad tienen acceso pleno a los mismos, no siendo así con los sitios internacionales ubicados principalmente en los dominios (.ORG y .COM) donde su acceso es un poco más limitado; también realizando una comparación entre los cinco estudios, los porcentajes de la figura 9 evidencian el comportamiento estable que ha tenido este indicador a lo largo de todos los estudios realizados.

2.6.2. Idioma de las páginas.

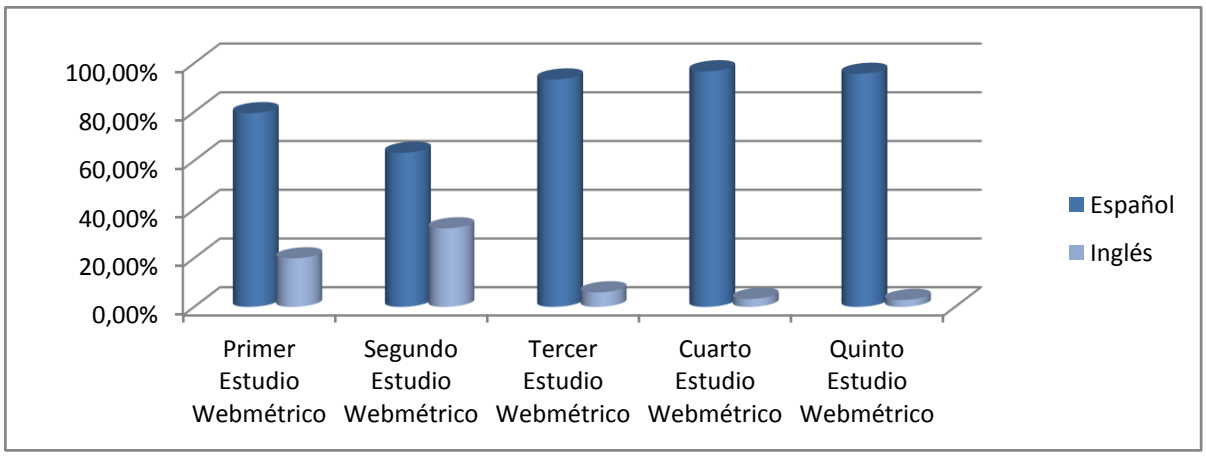

Figura 10. Idiomas más usados en los cinco estudios webmétricos realizados.

De acuerdo a esta gráfica, el mayor porcientos de las páginas web de la UCI están en idioma español, se puede ver cómo ha aumentado este número a partir del tercer estudio webmétrico, donde actualmente ya son pocos los sitios que presentan otro idioma que no sea el español, Como muestra la gráfica hay cierto uso del idioma inglés pero en menos medida. Por lo que se puede decir que considerando los porcentajes de los cinco estudios, y valorando los datos de los tres últimos, se puede pronosticar que la web de la UCI seguirá utilizando el idioma español con un mayor grado y a medida que los sitios vayan aumentando van a ir decayendo variablemente los porcentajes de los demás idiomas, principalmente el inglés, que es el segundo que más se utiliza en la UCI.

2.6.3. Software utilizado como Servidor Web y Sistemas Operativos.

En la universidad se usan con más frecuencia los sistemas operativos Windows y Linux, siendo utilizados mayormente como servidores Web, el Apache y el Microsoft IIS (Internet Information Server), aunque estos son los más usados no se puede descartar el uso de otros software como son Zope y Lighhtpd. A continuación se muestran tablas y gráficas que representan los por ciento de los distintos softwares y sistemas operativos utilizados en la universidad a partir del tercer estudio webmétrico realizado.

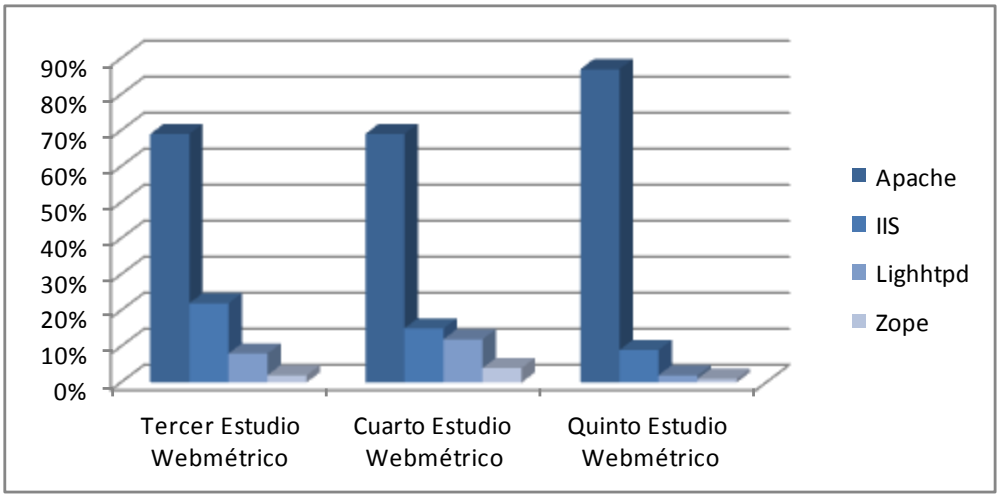

Figura 11. Software utilizado como servidor Web.

Sistemas Operativos.

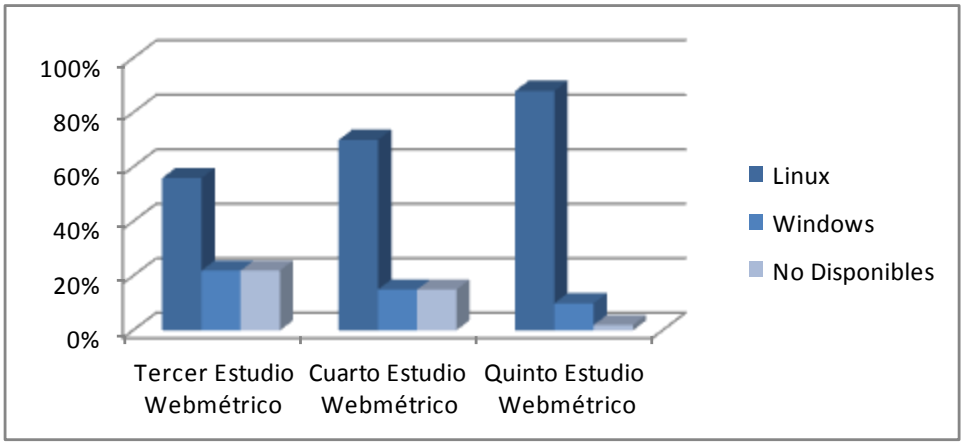

Figura 12. Sistemas operativos más usados.

A partir de los datos y las figuras anteriores es fácil darse cuenta como el uso del software libre va en ascenso en la universidad, es una prueba contundente que demuestra como en la UCI el empleo de herramientas y sistemas basados en software libre son prácticamente la primera opción para la utilización de servidores Web y sistemas operativos como Debian, Ubuntu, Red Hat, entre otros. No se puede pasar por alto también que en la universidad se desarrolló una distribución de GNU/Linux llamada Nova, que es la única distribución nacional conocida hasta este momento no solo en la UCI sino también en todo el país,

vale decir que existen otras distribuciones desarrolladas pero lamentablemente no se posee la documentación necesaria como para que se dé a conocer tanto a nivel nacional como internacional. Por todo lo anteriormente dicho la Web de la UCI va a seguir las siguientes tendencias:

- Aumentará el uso de sistemas operativos basados en software libre haciéndose más visible en la Web las distribuciones Debian, Ubuntu y otras, disminuyendo cada vez más el uso de Windows como sistema operativo, no se afirma que desaparezca totalmente pero si disminuirá considerablemente.
- Los software utilizados como servidores Web están estrechamente relacionados a los sistemas operativos, por tanto, va a existir una proporcionalidad en este aspecto, aumentando cada vez más el uso de Apache como servidor Web y dejando atrás el Microsoft IIS (Internet Information Server), como uno de los tipos de servidores más usados, sin contar claro está el aumento también de la visibilidad de los servidores menos conocidos hasta este momento.

2.6.4. Cantidad de páginas únicas/duplicadas.

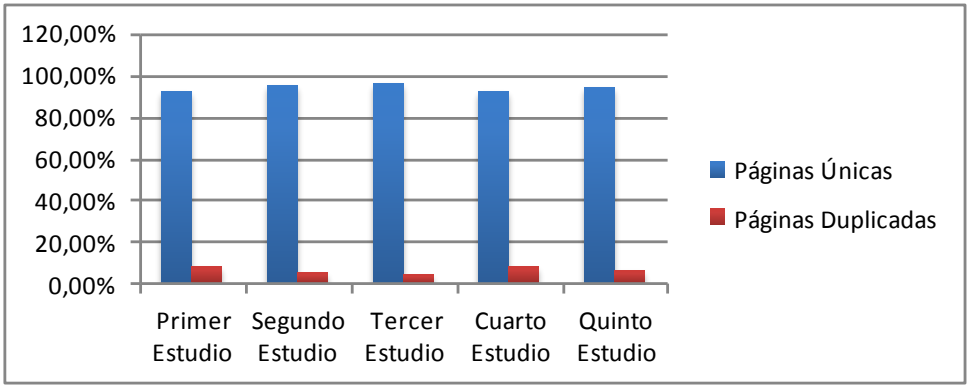

Figura 13. Cantidad de páginas únicas/duplicadas.

Este indicador ha mostrado un descenso de las páginas únicas en los últimos dos estudios comparado con el segundo y tercer estudio, no siendo así las páginas duplicadas que han mostrado un aumento, evidenciándose la reproducción de la información. En el futuro el comportamiento de las páginas duplicadas tendrían una tendencia a disminuir y las páginas únicas a aumentar, pues los resultados del cuarto estudio y del quinto, realizados con diferencia de pocos meses derivan este comportamiento y

también la universidad actualmente está adoptando la iniciativa de centralizar toda la información respecto a un mismo tema ya sea de docencia, investigación, producción, comunidades de desarrollo, entre otros.

2.6.5. Cantidad de páginas dinámicas/estáticas.

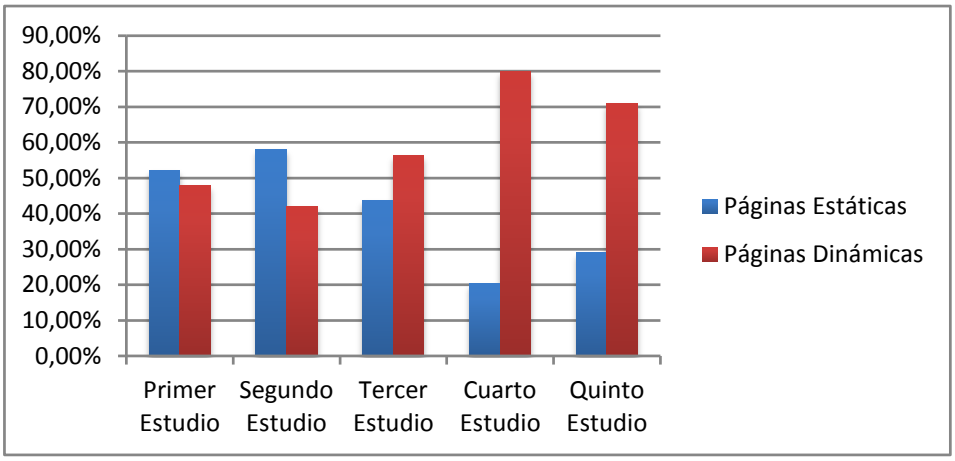

Figura 14. Cantidad de páginas dinámicas/estáticas.

De acuerdo a los datos anteriores se evidencia un predominio de las páginas dinámicas con respecto a las páginas estáticas en la Web de la UCI. Este aumento se debe en gran medida al uso de CMSs a la hora de implementar los sitios, pues estos incorporan una serie de ventajas o facilidades como son: la interactividad con el visitante, la utilización de bases de datos las cuales organizan y reducen el tamaño de la información almacenada. Por los planteamientos anteriores se puede pronosticar que el desarrollo de la Web de la UCI tendrá una tendencia a que las páginas dinámicas sigan creciendo considerablemente con relación a las páginas estáticas, reafirmándose estos datos en los últimos tres estudios realizados.

2.6.6. Extensiones encontradas.

A continuación se muestra una tabla con las extensiones más conocidas, con la representación de los porcientos que las sitúan como las extensiones más usadas en la Web de la UCI. La tabla contiene la información de los cinco estudios realizados pero basados en las extensiones con mayor representatividad que arrojó el último estudio.

	Extensiones	Primer Estudio	Segundo Estudio	Tercer Estudio	Cuarto Estudio	Quinto Estudio
Video	MOV	69.59%	57.31%	70.90%	84.95%	53.46%
Audio	MP3	98.78%	80.17%	68.52%	48.56%	91.34%
Imagen	PNG	27.92%	38.24%	26.69%	38.99%	46.84%
CGI	PHP	99.34%	99.32%	99.61%	99.16%	99.41%
Software	DEB	90.08%	88.17%	0.46%	86.75%	86.88%
Compresión	GZ	74.92%	90.40%	88.27%	38.76%	97.01%

Tabla 12. Principales extensiones de los cinco estudios webmétricos.

Mucho de los datos mostrados no sufrieron cambios en los cinco estudios webmétricos, tal es el caso de las extensiones CGI con los ficheros PHP que se han mantenido siempre como las más usadas de su tipo en la Web, oxilando siempre entre el 99.16% y 99.61%, a partir del poco cambio que muestra esta extensión se puede decir que este tipo de extensión referente al código fuente o entradas comunes no presentará muchos cambios en el futuro. Esto se debe a que la mayoría de los proyectos productivos de la universidad desarrollan portales o sitios Web utilizando como lenguaje principal el PHP, por tanto, estas extensiones no han de sufrir muchos cambios en su comportamiento futuro.

El resto de las extensiones en todo el recorrido webmétrico que se ha realizado han sufrido altibajos que hacen casi imposible la realización de predicciones sobre su comportamiento ya que muchos de ellos no logran mantenerse como el más visible en la Web entre dos estudios consecutivos. A pesar de que se hace un poco difícil y puede que no sean exactos los pronósticos las extensiones de video, audio, imagen, compresión y software no deben tener mucha variabilidad, por ejemplo, tomando las extensiones de software como caso de estudio, las más usadas son los .DEB que son archivos usados en Linux o lo que es lo mismo en sistemas basados en software libre, por lo que si se enlaza la tendencia a seguir por los sistemas operativos y las extensiones .DEB es notable su permanencia en la Web con menos altibajos e

incluso con tendencias a aumentar su porcentaje. Las extensiones de imagen sorpresivamente sufrieron cambios como se pudo ver en el epígrafe 2.4.6 tal vez este cambio esté propiciado por la interconexión existente con los sistemas operativos utilizados aunque este tipo de extensión es multiplataforma su gran visibilidad en este último estudio se puede considerar interesante e incluso puede que en próximos estudios siga siendo el nuevo monarca de las extensiones de imagen.

2.7. Conclusiones

Los resultados obtenidos en este capítulo brindan el escalón más alto para lograr valorar la evolución que ha tenido la Web en la UCI. A través de tablas y gráficas se muestran los resultados logrando que sean un poco más entendibles para cualquier lector. Se desglosaron los indicadores utilizados, siendo explicados detalladamente. Además se establecieron pronósticos sobre el comportamiento futuro de la Web de la UCI, teniendo como base los resultados de los cinco estudios webmétricos realizados.

Conclusiones Generales

Con la culminación del presente trabajo de diploma se cumple con el objetivo trazado en el mismo, se ofrecen datos comparativos entre cada uno de los estudios, basados en indicadores webmétricos como la edad de las páginas, cantidad de páginas únicas/duplicada, extensiones encontradas en la Web, entre otros indicadores que constituyen la base fundamental para medir la evolución de la Web durante los últimos años.

A través del estudio de los principales conceptos relacionados con el objeto de estudio se profundizó en esta joven ciencia que es la Webmetría, encargada de estudiar el comportamiento de la Web. Exponiendo sus fronteras, características, herramientas para su desarrollo, ventajas y principales definiciones.

Los resultados obtenidos en el trabajo de diploma brindan el escalón más alto para lograr valorar la evolución que ha tenido la Web en la UCI. Esto se puede constatar a través del más del millón de páginas descargas, los 156 sitios analizados, la cantidad de páginas que se encuentran disponibles representando el 87.10% del total de páginas, entre otros datos de interés. Se establecieron pronósticos sobre el comportamiento futuro de la Web, tales como; el crecimiento futuro de páginas y sitios Web, el aumento del uso de Linux como sistema operativo llevando consigo el aumento de los servidores Apache y disminuyendo el uso de Windows y de Microsoft IIS como sistema operativo y servidor Web respectivamente.

De manera general, los resultados obtenidos son de un gran valor para toda la comunidad universitaria, y fundamentalmente para la dirección de la UCI, ya que la información que aquí se proporciona es de vital importancia e incluso se puede decir que hasta estratégica, principalmente si se trata de la renovación tecnológica de la universidad.

Recomendaciones

Por la propia naturaleza cambiante de las páginas y sitios Web de la UCI se recomienda continuar realizando estudios webmétricos cada cierto período de tiempo para conocer los cambios que pueden ocurrir en la Web.

Tomar el presente trabajo como referencia para la toma de decisiones en diferentes aspectos, fundamentalmente los relacionados con la tecnología utilizada en el centro, que afectan a toda la comunidad universitaria.

Llevar el presente trabajo de diploma fuera de las fronteras de la UCI hacia el resto de las instituciones cubanas que de una forma u otra están inmersas dentro del desarrollo tecnológico, para aprovechar las ventajas que poseen los indicadores webmétricos ya sea en redes internas o externas.

Publicar los datos obtenidos de todos y cada uno de los estudios webmétricos realizados a través de un sitio Web y hacerlo visible a toda la comunidad universitaria.

Referencias Bibliográficas

[1] **Berners-Lee, T, y otros.** *The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities*. [En línea] 2010. [Citado el: 2 de 12 de 2009].. Disponible en: http://www.sciam.com/2001/0501issue/0501berners-lee.html.

[2] **Cronin, Blaise, Mckim, Geofrrey.** *Internet. In: A Informação: tendências para o novo milênio*. Brasília: IBICT, 1999.

[3] **Wolton, Dominique.** *Internet, e depois?: uma teoria das novas mídias*. Porto Alegre: Sulina, 2003.

[4] **Castells, Manuel.** *Internet y la sociedad red: lección inaugural del programa de doctorado sobre sociedade de la información y del conocimento en Universitat Oberta de Catalunya – UOC*. . [En línea] 2010. [Citado el: 4 de 12 de 2009]. Disponible en: http://www.uoc.edu/web/esp/articles/castells/castellsmain.html.

[5] **MsC. Martínez, Rodríguez, Ailín,** *Indicadores cibermétricos: ¿Nuevas propuestas para medir la información en el entorno digital?*, [En línea] 2010. [Citado el: 30-11-2009]. Disponible en: http://bvs.sld.cu/revistas/aci/vol14_4_06/aci03406.htm

[6] **Mondelo, Hernández, Yonny y Díaz, Madruga, Yuley.** *Características de la Web de la Universidad de las Ciencias Informáticas.* Ciudad de la Habana : s.n., 2009.

[7] **Ranking Web de universidades del Mundo.** *Ranking Web de universidades del Mundo* [En línea] 2010. [Citado el: 10 de 02 de 2010.]. Disponible en: http://www.webometrics.info/about_es.html.

[8] **Aguillo, Isidro F., y otros.** *Indicadores Web de actividad científica formal e informal en Latinoamérica.* Madrid : s.n., 2006.

[9] **Cronin, Blaise, Mckim, Geofrrey.** *Science and scholarship on the World Wide Web: a North American perspective*. Journal of Documentation, v. 52, v. 2, 1996.

[10] **Ingwersen, P; Christensen, F.H.** *Data set isolation for bibliometric online analyses or research publications: fundamental methodological issues.* Journal of the American Society for Information Science, v. 48, n. 3, 1997.

[11] **Bjórneborn, Lennart.** *Small-world structures across an academie web space: a library and information science approach.* PHD dissertation. Copenhagen, DK: Department of Informations Studies, Royal School of Library and Information Science, 2004.

[12] **Cronin, Blaise.** *Bibliometrics and beyond: some thoughts on web-based citation analysis*. Journal of Information Science, v. 27, n. 1, 2001.

[13] **Shiri, A.A.** *Cybermetrics: a new horizont in information research. 1998*. Paper presented at the 49th FID conference and congress held in India, New Delhi. 11-17 Octubre. 1998.

[14] **Thelwall, Mike.** *What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation*. Information Research, v. 8, n. 3 apr. 2003.

[15] **Thelwall, Mike.** *Web impact factors and search engine coverage*. Journal of Documentation, v. 56, n. 2, 2000.

[16] **Vanti, Nadia.** *Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do con hecimento. Ciência da Informação*, v. 31, n. 2, mayo/agosto. 2002.

[17] **Ingwersen, Peter.** *The calculation of Web impact factors*. Journal of Documentation, v. 54, n. 2, p. 236-243, 1998.

[18] **Almind, Tomas C.; Ingwersen, Peter.** *Informetric analyses on the world wide web: methodological approaches to 'Webometrics'.* Journal of Documentation, v. 53, n. 4, 1997

[19] **Vanti, Vitullo, Nadia Aurora.** *Links Hipertextuais na Comunicação Científica: análise webométrica dos sítios acadêmicos latino-americanos em Ciências Sociais.* Porto Alegre : s.n., 2007.

[20] **Thelwall, Mike.** *What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation*. Information Research, v. 8, n. 3 apr. 2003.

[21] **Smith, Alastair**. *A tale of two web spaces: comparing sites using web impact factors*. Journal of Documentation, v. 55, n. 5, p. 577-592, dez. 1999.

[22] **Aguillo, Isidro F., y otros.** *Factor de impacto y visibilidad de 4.000 sedes web universitarias españolas*. [En línea] 2010. [Citado el: 10 de 01 de 2010.]. Disponible en: http://www.cindoc.csic.es/estudios_ea2004_0020_informe.doc

[23] **Bar-Ilan, Judit.** *Search engine results over time: a case study on search engine estability*. Cybermetrics, v. 2/3, n. 1, 1998/99. . [En línea] 2010. [Citado el: 15 de 01 de 2010.] Disponible en: www.cindoc.csic.es/cybermetrics/vol2iss1.html

[24] **Olvera Lobo, María Dolores.** *Métodos y técnicas para la indización y la recuperación de los recursos de la World Wide Web*. Boletin de la Asociación Andaluza de Bibliotecários, n. 57, 1999. [En línea] 2010. [Citado el: 18 de 01 de 2010.] Disponible en: http://www.aab.es/51n57a4.htm

[25] **Mondelo, Hernández, Yonny.** *Cuarto Estudio Webmétricos en la Universidad de las Ciencias Informáticas.* Ciudad de La Habana, Cuba. Enero de 2010.

[26] **Portable Network Graphics (PNG).** [En línea] 2010. [Citado el 28 de 04 de 2010]. Disponible en: http://es.wikipedia.org/wiki/Portable_Network_Graphics

[27] **MP3.** [En línea] 2010. [Citado el 28 de 04 de 2010]. Disponible en: http://es.kioskea.net/contents/audio/mp3.php3

[28] **PHP.** [En línea] 2010. [Citado el 28 de 04 de 2010]. Disponible en: http://es.wikipedia.org/wiki/PHP

Bibliografía

- **Aguillo, Isidro F.** *Cibermetría: la métrica de la Web*. In: SEMINÁRIO BUSQUEDA: DEL ARCHIVO A LA RED. Madrid: Residencia de Estudiantes Fundación Francisco Giner de los Ríos, 2003. Disponíble en: http://www.archivovirtual.org/seminario/busqueda.htm Consultado el: 8-01-2010.

- **Aguillo, Isidro F.** *Posicionamiento en el web del sector académico iberoamericano*. Interciencia, Caracas, v.30 n.12 dic. 2005a. Disponible en: http://www.scielo.org.ve/scielo.php?pid=S0378-18442005001200003&script=sci_arttext Consultado el: 10-01-2010.

- **Almind, Tomas C; Ingwersen, Peter.** *Informetric analyses on the world wide web: methodological approaches to 'Webometrics'.* Journal of Documentation, v. 53, n. 4, p. 404-426, 1997.

- **Arroyo, Natalia.** *Métodos y herramientas para la extracción de datos en Cibermetría: el software académico y comercial*. Universidad de Salamanca. Departamento de Biblioteconomía y Documentación. Director: José A. Frías Montoya. 2004.

- **Lynch, Clifford.** *Searching the Internet: combining the skills of the librarian and the computer scientist may help organize the anarchy of the Internet*. Scientific American, marzo 1997. Disponible en: http://www.sciam.com/0397issue/0397lynch.html. Consultado el: 18-01-2010.

- **Notess, Greg R.** *On the net: search engine inconsistencies*. Online, v. 24, n.2, 10 mayo 2004.

- **Quoniam, L; Rostaing, H.** *From Scienciometrics, Informetrics to Internetometrics, Cybermetrics or is it possible to neglect Internet nowadays?* In: CYBERMETRICS'97. Jerusalém, Israel, 1997. Comunicação Científica. Jerusalém, Israel, 1997. Disponible en: http://www.cindoc.csic.es/cybermetrics/cybermetrics.html Consultado el: 5-01-2010.

- **Vanti, Vitullo, Nadia Aurora.** *Links Hipertextuais na Comunicação Científica: análise webométrica dos sítios acadêmicos latino-americanos em Ciências Sociais*. Porto Alegre : s.n., 2007.

- **Web Indicators Portal**. [En línea] [Consultado el: 8-02-2010]. http://www.webindicators.org.

Anexos

Anexo 1. Extensiones de video.

Nombre de la extensión	Total de documentos	Porcentaje (%)
MOV	3 621	53.46
DAT	2 229	32.91
FLV	421	6.22
WMV	307	4.53
AVI	91	1.34
MPG	84	1.24
MP4	17	0.25
QT	3	0.04

Anexo 2. Extensiones de imagen.

Nombre de la extensión	Total de documentos	Porcentaje (%)
PNG	8 198 003	46.84
GIF	7 295 052	41.68
JPG	1 728 412	9.88
ICO	279 390	1.6
IMG	216	0
BMP	57	0
WMF	28	0
PBM	8	0

Anexo 3. Extensiones CGI.

Nombre de la extensión	Total de documentos	Porcentaje (%)
PHP	19 721 282	99.41
ASP	87 280	0.44
CGI	12 963	0.07
PL	7 195	0.04
JS	5 939	0.03
JSP	1 507	0.01
PY	831	0
SHTML	591	0
TPL	88	0
CFM	87	0
BIN	83	0
PM	63	0
CFG	54	0
JHTML	37	0
PCGI	3	0

Anexo 4. Extensiones que no son HTML.

Nombre de la extensión	Total de documentos	Porcentaje (%)
README	277 002	84.67
PDF	28 387	8.68
DOC	11 607	3.55
XML	2 579	0.79
TXT	2 211	0.68
PPT	2 117	0.65
YML	935	0.29
XLS	682	0.21
ODT	582	0.18
DTD	283	0.09
RTF	154	0.05
ODP	108	0.03
INI	92	0.03
TTF	81	0.02
ASC	79	0.02
ASM	59	0.02
CONF	50	0.02
LOG	37	0.01
TEX	27	0.01
LIST	15	0
SQL	15	0
MSO	12	0
XSL	10	0
CHM	8	0
EL	4	0
ODS	3	0
TORRENT	2	0
AUTORUN	1	0
OTT	1	0
RSS	1	0
TEMPLATE	1	0

Anexo 5. Extensiones desconocidas.

Nombre de la extensión	Total de documentos	Porcentaje (%)	Nombre de la extensión	Total de documentos	Porcentaje (%)	Nombre de la extensión	Total de documentos	Porcentaje (%)
META	187 787	50.48	MSI	94	0.03	PRELOAD	34	0.01
SHOWPROBLEM	60 439	16.25	WSS	92	0.02	PUBLISHER	34	0.01
STATUS	44 622	12	POT	90	0.02	SEAM	34	0.01

ATOM	17 626	4.74	THMX	77	0.02	SERVICE	34	0.01
CHANGES	13 500	3.63	EML	71	0.02	CACHEDESCRIPTORS	33	0.01
PLAY	13 170	3.54	YAML	68	0.02	DOTTEDNAME	33	0.01
XPI	12 176	3.27	DB	63	0.02	EXCEPTIONS	33	0.01
CSV	4 392	1.18	XSD	62	0.02	I18NMESSAGEID	33	0.01
X	3 075	0.83	CHERRY	59	0.02	SCHEMA	33	0.01
COM	2 014	0.54	NEO	58	0.02	CONFIGURATION	32	0.01
ORG	1 779	0.48	TOOL	55	0.01	LOCKFILE	32	0.01
APP	1 544	0.42	XPATH	52	0.01	TESTING	32	0.01
DMG	1 175	0.32	ISPELL	48	0.01	TRAVERSING	32	0.01
MO	836	0.22	PROPERTIES	47	0.01	COMPONENT	31	0.01
NET	766	0.21	CTL	46	0.01	DATA	31	0.01
PO	689	0.19	DO	46	0.01	COPY	30	0.01
PAR	629	0.17	NSF	41	0.01	HOOKABLE	30	0.01
CU	621	0.17	SECURITY	41	0.01	US	30	0.01
MSPX	328	0.09	AUTHENTICATION	38	0.01	LEGAL	29	0.01
MHT	266	0.07	TESTHROWSER	38	0.01	PROXY	29	0.01
INCLUDE	264	0.07	II	37	0.01	EVENT	28	0.01
PXF	194	0.05	WSDL	37	0.01	POD	28	0.01
MSG	180	0.05	LSS	36	0.01	INTERFACE	27	0.01
ANNOUNCE	169	0.05	DEVTOOLS	35	0.01	STATS	27	0.01
PT	150	0.04	I18N	35	0.01	PPD	26	0.01
REQUIRE	148	0.04	LOCATION	35	0.01	SENDMAIL	26	0.01
PISI	102	0.03	SQLALCHEMY	35	0.01	BROWSER	24	0.01
PYDEB	24	0.01	BUILDO	14	0	CREOLE	9	0

			UT					
AIR	22	0.01	EDU	14	0	FAMFAMFAM	9	0
FRAMEWORK	22	0.01	ADMIN	13	0	FILES	9	0
URI	22	0.01	ICPC	13	0	HTC	9	0
WHAT	22	0.01	PCAP	13	0	ICU	9	0
RESTFULCLIENT	21	0.01	ES	12	0	LIME	9	0
WHO	21	0.01	PRO	12	0	MAP	9	0
ACTION	20	0.01	UK	12	0	PHING	9	0
CD	19	0.01	CONTENTTYPE	11	0	PHPMAILER	9	0
MIGRATED	19	0.01	RULES	11	0	PRADO	9	0
STUFF	19	0.01	BRITAIN	10	0	PROPEL	9	0
ASHX	18	0	EYEPACKAGE	10	0	PROTOTYPE	9	0
CAT	18	0	KEY	10	0	PSPIMAGE	9	0
DE	18	0	NET_REMOTIONG	10	0	TMPL	9	0
GSG	18	0	P	10	0	TMPROJ	9	0
JSB	18	0	RU	10	0	CA	8	0
NEW	18	0	STM	10	0	KEEP_SYS	8	0
OGV	18	0	UTIL	10	0	KEYWORDS	8	0
TM2	18	0	AGAVI	9	0	MASK	8	0
PAGE	17	0	CONFIGFILES	9	0	PHPS	8	0
RMDIR	8	0	ASIA	2	0	BIND	1	0
UNMASK	8	0	BIZ	2	0	CHAMBERY	1	0
USE	8	0	BR	2	0	CPKG	1	0
ANYMORE	7	0	BROWN	2	0	CRT	1	0
BY	7	0	BUISSE	2	0	CVS	1	0
ENTRY	7	0	CACHE	2	0	DAA	1	0
POST	7	0	CARREIRA	2	0	DIR	1	0
QRC	7	0	CH	2	0	EC	1	0
RECENT	7	0	CN	2	0	ENT	1	0

STALLMAN	7	0	DK	2	0	FROM	1	0
TGS	7	0	DUMP	2	0	GFGF	1	0
BENLI	6	0	FI	2	0	GIT	1	0
CX	6	0	FM	2	0	GNE	1	0
LXP	6	0	GEO	2	0	HIRSEH	1	0
MOBI	6	0	IDB	2	0	HTACCESS	1	0
RB	6	0	INFO	2	0	IDL	1	0
SYSLOG	6	0	ISS	2	0	JSON	1	0
UA	6	0	IT	2	0	LOCALNAME	1	0
ARS	5	0	KDEVELOP	2	0	LSM	1	0
PROBLEMSTATUS	5	0	KDEVSES	2	0	LSM	1	0
SUBMITPAGE	5	0	KUBINA	2	0	MANIFEST	1	0
VERSION	5	0	LAWSON	2	0	PAT	1	0
XUL	5	0	LY	2	0	PG	1	0
BUNDLE	4	0	MIL	2	0	SHARE	1	0
IDX	4	0	MINE	2	0	SOL_PREINSCRIPCION	1	0
INSTALL	4	0	MODULE	2	0			
MACOS_BINARIES	4	0	NL	2	0			
MINIX	4	0	PCS	2	0			
PLANNER	4	0	PEHOFER	2	0			
RAZR3	4	0	PETRAK	2	0			
REMOTE	4	0	PF	2	0			
SIG	4	0	PHILIPS	2	0			
CAIRO	3	0	SESSION	2	0			
CATALOGUELIST	3	0	TO	2	0			
CHANGELOG	3	0	VACKLIN	2	0			
EXAMPLE	3	0	WMZ	2	0			

PATCHES	3	0	ADS	1	0			
SHOCK WAVEFLASH	3	0	ALERT	1	0			
SS	3	0	APP3	1	0			
VSPX	3	0	APPS	1	0			

Glosario de Términos

TIC: Tecnologías de la Información y las Comunicaciones.

www: World Wide Web. También conocida como “la Web “o” la Red“. Sistema mundial de servidores Web conectados a Internet. No todos los ordenadores conectados a Internet forman parte de la (WWW).

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

UCI: Universidad de las Ciencias Informáticas.

CCHC: Centro de Ciencias Humanas y Sociales.

CSIC: centro nacional de investigación de España.

GEWEB: Generador de Estudios Webmétricos.

CIBA: Grupo de Proyectos de Cibermetría Aplicada.

SINI: Polo Productivo de Soluciones Informáticas para Internet.

CIW: Centro de Investigación de la Web.

WISER: Web de Indicadores de Ciencia, Tecnología e Innovación de Investigación.

EICSTES: Indicadores europeos, Sistema de Economía de la Ciencia-Tecnología y el Ciberespacio.

SIBIS: Los indicadores estadísticos de evaluación comparativa de la Sociedad de la Información.

OCLC: Online de Librerías de Computadoras.

SCRG: Grupo de Investigación de Cibermetría Estadística..

CINDOC/CSIC: Centro de Información y Documentación Científica del Consejo Superior de Investigaciones Científicas en España.

W3C: World Wide Web Consortium. Consorcio internacional de compañías y organizaciones involucradas en el desarrollo de Internet y en especial de la WWW. Su propósito es desarrollar estándares y ”poner orden“ en Internet.

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

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

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

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.

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

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