Universidad de las Ciencias Informáticas.

Facultad 10.

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

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

***Autor* (es):**

Annie Hernández Sánchez

Yordani Molina Peña

***Tutor(es)*:**

Msc. Prof. Aux. Orlando Cárdenas Fernández

Ciudad de La Habana, Junio 2008.

El futuro pertenece a quienes creen en la belleza de sus sueños.

Eleanor Roosevelt.

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 firmo la presente a los ____ días del mes de _____ del año 2008.

_____	_____	_____
Firma del Autor	Firma del Autor	Firma del Tutor
Annie Hernández Sánchez	Yordani Molina Peña	Orlando Cárdenas Fernández

OPINIÓN DEL TUTOR DEL TRABAJO DE DIPLOMA.

Título: Caracterización de la Web de la Universidad de las Ciencias Informáticas

Autores: Annie Hernández Sánchez y Yordani Molina Peña.

Por todo lo anteriormente expresado considero que el estudiante está apto para ejercer como Ingeniero Informático; y propongo que se le otorgue al Trabajo de Diploma la calificación de ____ .

_____ ____ De julio del 2008.

Firma

Agradecimientos

Agradezco a:

A mis padres y familiares por tanto amor y tanto sacrificio, por estar siempre ahí dándome todo su apoyo y todo su amor. Por impulsarme hacia el camino correcto y hacer que me sienta hoy satisfecha por el esfuerzo que he hecho.

En especial a mi padre Cheito que sabes que sin ti nada seria posible, gracias por haberme hecho sentir mejor aun en la distancia y por haberme acogido como tu hija, por estar al lado mío en los buenos y en los momentos que más lo necesite, un beso. A mi padre del alma papi Raúl, por ayudarme dándole apoyo a mí mamá y haciéndome feliz en los pequeños momentos en los que estuvimos juntos, no sabes cuanto te quiero. Mima para ti mi agradecimiento se hace pequeño.

A Lázaro mi novio del alma, por tu apoyo e incondicionalidad, a Conchy gracias. A Yordanis Molina mi compañero de tesis, tu sabes que te amo como amigo.

A los profesores que me han ayudado a lo largo de la carrera en especial a Clara Gisela, por sus consejos y la confianza que me hizo sentir en mi misma, al tutor, a mis amistades por no dejarme caer y a los que de una forma u otra hicieron posible este milagro como José Ramón.

Annie Hernández Sánchez

Les agradezco a mis padres y a mi hermano por darme fuerzas para seguir adelante. A toda mi familia en general y a Tere que me ha dado su apoyo incondicional. A mis compañeros de estudio, especialmente a Annie por siempre estar preocupándose por mí. A mis amistades, que han compartido conmigo esta etapa tan linda de la vida y en especial a Gio, Dania, Mayre y a Yalena. A todos los profesores que a lo largo de la carrera han contribuido con mi formación.

En fin a todos aquellos que de una forma u otro me ayudaron desinteresadamente para que fuera posible este maravilloso sueño.

Yordani Molina Peña

Dedicatoria

Mi tesis va dedicada a las tres mujeres que han significado todo en mi vida, y a quienes amo con toda la fuerza de mi corazón, a mi madre querida, a mi hermana del alma y a la memoria de mi abuela.

A mi mami alla, porque te me fuiste aun sin yo haber podido cumplir tu sueño que era verme graduada, se que significaba mucho para ti. Porque tus brazos siempre se abrían cuando quería un abrazo. Tu corazón comprendía cuando necesitaba una amiga. Tus ojos tiernos se endurecían cuando me hacía falta una lección. Tu fuerza y tu amor me guiaron, y me dieron alas para volar.

A ti mi negra por ser mi vida más que nada en el mundo, por ser mi mayor alegría, orgullo y mi mayor preocupación, por ti es que me empeño día a día para salir adelante, para ser tu mayor ejemplo a seguir, tu y yo somos un alma que habita en dos cuerpos; un corazón que habita en dos almas.

A mi mamá del alma, por darme la vida, por haberme ayudado a tener un futuro haciéndome una profesional, por aguantar conmigo todos estos años, por ser la luz que me guía para seguir adelante y no tener miedo aun en la distancia, por ser mi mejor amiga ahora y siempre, por haberme dado mi mayor felicidad que es mi hermana. Mami para ti lo más grande del mundo que es mi amor, mi cariño y mi devoción.

…Annie…

Dedico la tesis a mis padres por ser mis guías aun cuando no estuvimos cerca y a mi hermano por saber darme confianza cuando más lo necesite.

…Yordy…

RESUMEN

En el presente trabajo de investigación se caracteriza el espacio web de la Universidad de las Ciencias Informáticas (UCI) a partir del análisis de una muestra tomada en el mes de mayo de 2008. El estudio contempla tanto características particulares de las páginas, como el conjunto de ellas en el nivel sitio. Para la recolección de las páginas se utilizó el robot de búsqueda Wire Crawler, el cual recolectó un total de 856 páginas de 76 sitios que eran los que se encontraban enlazados en ese momento en los servidores. Para el análisis del nivel página se tomó en cuenta la muestra de las páginas descargadas mientras que para el nivel sitio se tomaron todas las páginas visitadas.

Los resultados obtenidos son consistentes con respectos a otros espacios Web. Del estudio surgen las siguientes observaciones:

El 90.48% de las páginas han sido creadas o modificadas en menos de un año, lo que indica que la Web UCI está creciendo aceleradamente. Con referencia a las tecnologías, el 47.08% de las páginas de la muestra son estáticas y el 52.92% dinámicas. Se encontró que el sitio más representativo para la UCI es softwarelibre.uci.cu, por ser el que más enlaces internos y mayor cantidad de documentos presenta. El dominio más referenciado desde la UCI es .cu, proveniente de Cuba.

De los indicadores anteriores se desprende que existe un importante desarrollo tecnológico y de la infraestructura de comunicaciones de la UCI relacionada con la Web.

Palabras Claves

Web universitaria, indicadores cibermétricos, páginas web, sitios web.

ÍNDICE

INTRODUCCIÓN 1

Fundamentación Teórica. 7

1.1 La Web. Su composición. 7
1.1.1 Tipos de Web. 9

1.2 Cibermetría y Webmetría. Surgimiento y objetivo. 11
1.2.1 Aplicaciones de la Cibermetría y la Webmetría. 12

1.3 Indicadores cibermétricos. 13
1.3.1 Ventajas de los indicadores cibermétricos. 16

1.4 Herramientas cibermétricas. 17
1.3.2 Robots de búsqueda. Principio de su funcionamiento. 17

1.5 Estudios internacionales y nacionales de la Web. 20

1.6 Conclusión. 30

Selección de los elementos cibermétricos para la investigación. 32

2.1 Análisis de indicadores. 32
2.1.1 Indicadores más frecuentes empleados para el análisis métrico. 33
2.1.2 Selección de los indicadores para la investigación. 35

2.2 Herramientas tecnológicas empleadas para desarrollar estudios cibermétricos. 38
2.2.1 Características de los robots de uso abierto. 40

2.3 Conclusión. 48

Análisis cualitativo y cuantitativo de la Web. 49

3.1 Nivel colección. 49

3.2 Nivel Página. 51

3.3 Nivel sitio 59

3.4 Conclusiones. 64

CONCLUSIONES GENERALES 65

RECOMENDACIONES 66

REFERENCIA BIBLIOGRÁFICA 67

BIBLIOGRAFÍA 68

Anexos 73

GLOSARIO DE TÉRMINOS 75

INTRODUCCIÓN

El desarrollo acelerado de los procesos sociales en la actualidad implica no solo un cambio de paradigmas económicos, políticos e ideológicos, también presupone un vertiginoso desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC).

El desarrollo tecnológico global implica el incremento del papel de los nexos informacionales a todos los niveles, constituyendo la red informática (Web) premisa, condición y consecuencia de este inesquivable progreso de la sociedad humana.

La Web presenta relaciones de información entre los documentos de hipertexto mediante enlaces que se establecen entre ellos. Esta se presenta como un espacio público utilizado por múltiples usuarios con objetivos diferentes. En un principio su objetivo era permitir compartir información, aunque no ha dejado atrás este objetivo, en la actualidad es un medio de publicación y servicios para múltiples usos como publicidad, entretenimiento, educación, comercio entre otros.

La Web se caracteriza en su definición por un conjunto inusual y heterogéneo de elementos. Las mismas características que la hacen un importantísimo medio de difusión y comunicación, hacen muy complejo su análisis. Por encontrarse en constante crecimiento y por ser tan dinámica en su desarrollo, el estudio de características y tendencias, brinda interesante información para entender su estructura.

La Web ha sido objeto de constantes estudios y análisis. A nivel mundial, se han hecho diferentes caracterizaciones a través de indicadores cibermétricos que ayudan a conocer su comportamiento. En varios países se han realizado investigaciones con determinada profundidad sobre los contenidos y estructuras de la Web: Argentina, Austria, África, Chile, Corea, Cuba, España, Hungría, Perú y Portugal.

El país que más experiencia tiene en el estudio de las características de su Web es Chile, que ha sido objeto de estudio durante los años 2000, 2002, 2004, y la más reciente es la del 2006.

En nuestro país se han realizado acercamientos teóricos que constituyen la base para los estudios del comportamiento de la Web cubana, fundamentalmente en la Empresa de Tecnologías de la Información y Servicios Telemáticos Avanzados (CITMATEL), llamado Estudio de las estadísticas Web de accesos y visitas del Portal Cuba.cu, el cual aportó significativos datos para haber sido el primer acercamiento a la Web cubana.

El análisis cualitativo y cuantitativo de Internet desde la Universidad de las Ciencias Informáticas (UCI), constituye el objetivo fundamental del grupo de proyecto Grupo de Cibermetría Aplicada (CIBA), circunscrito al Polo Científico Productivo Estudios de Internet.

En los cinco años de desarrollo de esta Ciudad Digital no se ha realizado una caracterización de la Web universitaria, por lo que no se dispone de suficiente información acerca de su estructura y contenidos, tampoco se cuenta con probadas técnicas cuantitativas para la descripción y evaluación de sus contenidos.

La ausencia de la caracterización implica también la carencia de estudios de indicadores que permitan medirla en cifras para la obtención de estadísticas rigurosas, pertinentes y actualizadas sobre el impacto de la red en distintas actividades de interés científico-tecnológico, económico y social y el análisis de los patrones de comunicación científica a través de la Web y su trascendencia para el proceso de informatización de la sociedad cubana.

Al no disponer de una investigación de la Web local que sirva de plataforma para indagaciones más abarcadoras, no se puede definir las tendencias de su desarrollo histórico y proyecciones futuras, algo tan necesario para el grupo de proyecto a mediano plazo.

Se puede inferir como **Problema Científico** de la investigación: ¿Cómo resolver la carencia de información sobre el comportamiento de los indicadores cibermétricos en la Web de la Universidad de las Ciencias Informáticas?

El **Objeto de estudio** es el proceso de control de la información sobre el comportamiento de los indicadores en la Web.

El **Campo de acción** está dado por el proceso de control de la información sobre el comportamiento de los indicadores de la Web en la Universidad de las Ciencias Informáticas.

Se persigue como **Objetivo general:**

1. Caracterizar integralmente el comportamiento de la Web de la Universidad de las Ciencias Informáticas.

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

1. Definir los indicadores de interés para la caracterización de la Web de la Universidad de las Ciencias Informáticas.
2. Monitorear el comportamiento de la Web de la Universidad de las Ciencias Informáticas.
3. Analizar cualitativa y cuantitativamente la información recuperada sobre la Web de la Universidad de las Ciencias Informáticas.

La investigación debe partir de la siguiente **Hipótesis**: Si se caracteriza integralmente la Web de la Universidad de las Ciencias Informáticas, entonces se resolvería la necesidad de información sobre el comportamiento de los indicadores cibermétricos en ella.

Dentro de esta respuesta anticipada al problema científico se pueden identificar las siguientes **variables:**

1. Caracterización integral de la Web de la UCI. (Variable independiente)
2. Necesidad de información sobre el comportamiento de los indicadores cibermétricos en la Web de la UCI. (Variable dependiente).

Operacionalización de la variable dependiente:

VARIABLE	DIMENSIONES	OPERACIONALIZACION	INDICADORES	INDICES
Necesidad de información sobre el comportamiento de los indicadores cibermétricos en la Web de la UCI	Eficiencia	Rapidez en las búsquedas de información.	Escala de puntuación	0-10
		Integralidad de la información.	Satisfacción de necesidades	0-10
	Competitividad	Confiabilidad.	Exactitud de los datos	0-10
		Competitividad tecnológica	Conocimiento de las tecnologías de punta	0-10

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

1. Revisar caracterizaciones de Web realizadas en otros países y Universidades Tecnológicas.
2. Analizar los indicadores cibermétricos utilizados en caracterizaciones anteriormente realizadas.
3. Definir los indicadores de interés para la caracterización de la Web de la Universidad de las Ciencias Informáticas.
4. Seleccionar el Spyder adecuado para la investigación.
5. Recuperar la información de la Web de la Universidad de las Ciencias Informáticas utilizando el Spyder.
6. Analizar el comportamiento de la Web de la Universidad de las Ciencias Informáticas.

Se espera que una vez concluido el trabajo se pueda contar con los siguientes **resultados concretos:**

- Metodología para el monitoreo y caracterización de la Web de la Universidad de las Ciencias Informáticas.

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

Para el alcance de estos resultados se utilizarán los siguientes métodos de investigación:

Métodos teóricos:

- Análisis - Síntesis.

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

- Inducción – Deducción

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

- Lógico – Histórico

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

Métodos Empíricos:

- Entrevista.

Se planifica entrevistar a los especialistas y directivos que tienen la responsabilidad del funcionamiento de la Web de la UCI, con el propósito de obtener información de primera mano sobre el objeto de estudio.

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

Primer capítulo: Fundamentación del tema, estado del arte a nivel internacional, nacional y en la UCI. Valoración de las caracterizaciones realizadas en otros países.

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

Tercer capítulo: Se realizará el estudio del comportamiento de la Web de la UCI, mediante un análisis cuantitativo y cualitativo de la misma.

1 CAPÍTULO
Fundamentación Teórica.

En el presente capítulo se brinda una visión general de los temas referentes al estado del arte del problema existente en la universidad, debido a la carencia de información sobre el comportamiento de los indicadores cibermétricos de su Web, por lo que se hace necesario realizar un estudio tanto a nivel nacional como internacional, para poder comprender la investigación de una forma más objetiva.

1.1 La Web. Su composición.

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

La Web es un conjunto de documentos entrelazados en un sistema de hipertexto. El usuario entra en la Web a través de una página de inicio. [1]

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

por Timothy Bernes-Lee en 1989 para el laboratorio europeo, para la física de partículas (CERN). [2]

Usando la Web, se tiene acceso a millones de páginas de información agrupadas en sitios. La apariencia de un sitio web puede variar ligeramente dependiendo del explorador que se use. Así mismo, las versiones más recientes disponen de una funcionalidad mucho mayor tal como animación, realidad virtual, sonido y música. La exploración en la Web se realiza por medio de un software especial denominado Explorador o Browser.

Su característica sobresaliente es el texto remarcado a través de hipervínculos, apareciendo palabras o frases en texto de un color diferente al resto del documento. Por lo general, este texto es subrayado. Al seleccionar una palabra o frase, se es transferido al sitio o página relacionada a esa frase. En algunas ocasiones hay botones, imágenes, o porciones de imágenes que pueden activarse mediante un clic. Esta característica proporciona relaciones de información mediante documentos, sirviendo como enlace entre ellos.

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

Composición de La Web.

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

Los **sitios web:** son un conjunto de archivos electrónicos y páginas web referentes a un tema en particular, que incluye una página inicial de bienvenida, generalmente

denominada página de inicio o home page. Que presentan un nombre de dominio y una dirección en Internet especifica, empleada por instituciones, organizaciones e individuos para comunicarse en el mundo.

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

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

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

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

1.1.1 Tipos de Web.

La evolución de la Web comienza con el surgimiento de la Web 1.0 (La Web estática), luego se le da paso a la Web 1.5 (Web dinámica), y por terminado con la Web 2.0 (Web colaborativa). Todo se debe a los avances tecnológicos y al empuje de muchas compañías que se han puesto a aprovechar de todas las potencialidades que presenta la Web, y así implementar un valor agregado a sus clientes.

La Web estática: *Son aquellos sitios enfocados principalmente a mostrar una información permanente, donde el navegante se limita a obtener dicha información, sin que pueda interactuar con la página web visitada, las Web estáticas están construidas*

principalmente con hipervínculos o enlaces (links) entre las páginas web que conforman el sitio, este tipo de Web son incapaces de soportar aplicaciones web como gestores de bases de datos, foros, consultas online, e-mails inteligentes.

Esta es una opción más que suficiente para aquellos sitios web que simplemente ofrecen una descripción de su empresa, quiénes somos, dónde estamos, servicios. Ideal para empresas que no quieren muchas pretensiones con su sitio web, simplemente informar a sus clientes de sus productos y su perfil de empresa. [3]

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

El desarrollo de este tipo de Web es más complicado, pues requieren conocimientos específicos de lenguajes de programación así como creación y gestión de bases de datos, pero la enorme potencia y servicio que otorgan este tipo de páginas hace que merezca la pena la inversión y esfuerzo invertidos respecto a los resultados obtenidos. [4]

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

Es una etapa que ha definido nuevos proyectos en Internet y está preocupándose por brindar mejores soluciones para el usuario final. Muchos aseguran que hemos reinventado lo que era el Internet, otros hablan de burbujas e inversiones, pero la realidad es que la evolución natural del medio realmente ha propuesto cosas más interesantes como lo analizados diariamente en las notas de Actualidad. [5]

1.2 Cibermetría y Webmetría. Surgimiento y objetivo.

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

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

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

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

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

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

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

Cibermetría: *Es la disciplina dedicada a la descripción cuantitativa de los contenidos y procesos de comunicación que se producen en el ciberespacio*. [6]

Webmetría: *Es el estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías de una parte concreta de Internet, por regla general a una Web o portal*. [7]

1.2.1 Aplicaciones de la Cibermetría y la Webmetría.

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

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

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

1.3 Indicadores cibermétricos.

Uno de los frentes abiertos en el campo de la Cibermetría es el estudio de aspectos importantes de la Web, a través de indicadores cibermétricos por ser *una medida de relevancia en la toma de decisiones, que cuantifica aspectos de creación, difusión y aplicación de la ciencia y la tecnología en la medida en que están representadas en Internet o el World Wide Web*. [8].

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

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

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

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

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

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

Entre los que se encuentran:

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

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

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

Medidas de visibilidad e impacto: Se basan en el carácter hipertextual de la Web y exploran los patrones de enlace entre páginas y sedes de distintas procedencias. El número y diversidad de enlaces externos recibidos, su volumen respecto a los contenidos y objetos de enlace (llamado Factor de Impacto Web) o índices que se construyen de acuerdo con el peso relativo de las sedes de origen de los enlaces, como el PageRank de Google, aunque en otras instituciones se puede visualizar como Ranking. Este indicador ayuda a que se sepa cuan bueno es el contenido o estructura del sitio o de la página al haber creado una buena impresión al visitante. Esta medida permite establecer listas ordenadas, según la jerarquía numérica de estos indicadores.

Entre estos indicadores se encuentra el factor de impacto (FI). El factor de impacto web (FIW) es uno de los primeros indicadores examinados en los trabajos webmétricos.

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

Medidas de popularidad: El consumo de información medido en términos de número y características de las visitas desde la Web resulta un atractivo, aunque extremadamente complejo de implementar, es un método de evaluación, que ayuda a saber la popularidad que puede presentar la información o la forma de representarla. Es ciertamente interesante para estudios temporales, en los que la medida de la evolución resulta prioritaria para los correspondientes informes. Como se indica es complicado obtener

valores absolutos, pero ciertos valores relativos con valores importantes pueden, no obstante, utilizarse en análisis comparativos.

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

1.3.1 Ventajas de los indicadores cibermétricos.

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

1. Mayor potencia crítica para la realización de análisis de patrones globales y sectoriales, dentro de las que están:
    - Mejor tratamiento con técnicas estadísticas.
    - Nuevas unidades: Mayor finura en el análisis.
    - Perspectiva cuantitativa y objetiva.
2. Mejores resultados esperados:

    - Presentando una batería de indicadores más amplia.
    - Las medidas combinadas.
    - Visualización más espectacular.
    - Seguimiento individualizado.
    - Medidas directas e indirectas.
    - Comparación con descriptores “offline”.

3. Ventajas políticas:

    - Medida de la producción del conocimiento.
    - Incremento de los contenidos.

1.4 Herramientas cibermétricas.

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

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

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

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

1.3.2 Robots de búsqueda. Principio de su funcionamiento.

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

robot simplemente visita los sitios y extrae los enlaces que están incluidos dentro de éstos.

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

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

Principio de funcionamiento.

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

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

Entre las tareas más comunes de los Web Crawlers tenemos:

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

Cómo decide un robot qué visitar.

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

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

Cómo decide un robot qué Indexar.

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

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

1.5 Estudios internacionales y nacionales de la Web.

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

- **Argentina.**

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

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

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

Enlaces: grado entrante y saliente de páginas, PageRank (Ranking de las páginas), grado entrante y saliente del Hostgraph, componentes fuertemente conectados y la estructura microscópica.

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

Dando como resultado que la cantidad media de páginas por sitio es 65. Existen un total de 66.021 componentes fuertemente conectados. Dentro de la distribución de los códigos HTTP para dar respuesta el que predomina es el OK con un 78.66%. Se observa una longitud promedio de 68 bytes sin incluir la parte correspondiente al protocolo, lo que la

incrementaría en 7 bytes. El mayor por ciento dentro de la profundidad de los documentos está dado por el número 4 que la presentan 4.964.279 documentos, para un 40,44%. Dentro de la distribución de sitios por país se encuentra Argentina con 18.177 sitios para un 75,87%; siguiéndole en la escala Estados Unidos con 4.700 sitios, para un 19.62%.

- **Austria.**

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

Los indicadores utilizados para hacer la caracterización fueron los siguientes:

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

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

Lo que dio como resultado encontrar más de 200 000 tipos diferentes de archivos sobre la base de sus extensiones, y más de 200 tipos de información. Se encontraron diferencias significativas en la manera de vídeo, se proporcionó información en relación con el tipo de servidor web utilizado. Contaban con servidores web Apache, Netscape, Stronghold. De 10 dominios 9 están estrechamente relacionados entre sí. Los distintos estudios realizados han mostrado estadísticamente, por ejemplo, los primeros rastros de documentos XML a principios de 1999 y revelan que la forma en documentos XML han ido aumentando en número y como parte de la Web documental que estará disponible a partir de 1999 hasta la fecha. Otro fascinante ejemplo de ello es la sorprendente victoria repentina del PDF sobre el archivo PostScript.

- **África.**

Los datos para el estudio de la Web africana se recolectaron con el robot de búsqueda UbiCrawler. Para la realización de la caracterización se tomaron en cuenta puntos de partida los datos obtenidos del análisis sintáctico como nivel de páginas HTML, los tipos de cabeceras HTTP, tipos de servidores, ultima fecha de modificación, tamaño de la

página, lenguaje natural, lenguaje del scripting, extensiones de archivos, protocolos en los números URL y los gráficos de la Web de África, donde está la distribución, el derecho del poder, la estructura de componentes fuertemente conectados y las interconexiones.

Con la guía de estos indicadores se llegó a la conclusión de que la mayoría de las páginas en el dominio de África no tienen un tipo de documento definido, pues la mayoría de los documentos están realizados en HTML 4 y es el 7.71% de toda la Web. La distribución de las cabeceras de las páginas no parece diferir significativamente de los datos que se conocen de manera general las investigaciones en la Web. La mayoría de los sitios en el dominio de África usa de tecnología de Microsoft-IIS, Apache y Netscape-Enterprise. Más de 600000 páginas están en los 10KB. Los resultados del lenguaje refieren a un total de 7 idiomas prevaleciendo el ingles con un 74.68%.

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

- **Chile. Análisis de tendencia.**

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

- **Chile 2000.**

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

Colección: cifras globales y estudio de vocabulario.

Página: tamaño, tipo de documento e idioma.

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

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

La colección descargada contaba con 730 673 páginas distribuidas en 10 352 sitios pertenecientes a 9 102 dominios, la misma utilizo 2.3 GB. Se observó que la mayoría de las páginas tienen poco texto siendo el promedio de texto de 3.4 KB mientras para la página en su totalidad es de 15.3 KB. Además que el tipo de documento con mayor empleo es el HTML con más del 95%.

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

- **Chile 2001-2002.**

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

Una porción importante de los dominios inscritos no se utilizaban, y de aquellos utilizados más de la mitad tenían sólo una página. La de presencia en la Web, el 56% de los dominios y el 54% de los sitios tienen sólo una página. En comparación con el 2000, en que un 45% de los sitios tenía sólo una página, produciéndose un aumento porcentual y absoluto en el número de sitios con una sola página. El tamaño promedio de una página era de 11.562 bytes, considerando sólo el texto y tags HTML. Sólo el 4% de las páginas contenían más de 40 KB de texto. Además del HTML en la Web existen contenidos de diversos tipos, los que también son interesantes de indexar y recuperar. Estos documentos de tipo distinto a HTML se dividieron en:

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

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

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

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

- **Chile 2004.**

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

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

En el estudio se limitó al recolector para que descargue solamente 5 niveles para páginas dinámicas, y solo 15 niveles para páginas estáticas, apreciándose que la cantidad de páginas dinámicas crece exponencialmente en cada nivel. Cerca del 38% de las páginas descargadas eran páginas dinámicas. La aplicación más usada para generarlas es PHP, seguida por ASP y páginas generadas usando Java. De los 370.000 enlaces a archivos que no eran HTML, pero que tenían extensiones que son comúnmente usadas para documentos, el formato Adobe PDF es el más ampliamente usado y el estándar de facto, seguido de texto plano y Microsoft Word.

Hay muchos enlaces a archivos multimedia, incluyendo más de 80 millones de enlaces a imágenes, 50.000 enlaces a archivos de audio, y 8.000 enlaces a archivos de video. Se encontraron enlaces a 30.000 archivos con extensiones usadas para código fuente, y 600.000 archivos con extensiones usadas para programas.

La Web presentó un promedio de 57 páginas por sitio, 17 sitios con 50 mil o más páginas estáticas y sólo otros dos sitios sobrepasaron las 4 mil páginas. El tamaño promedio de un sitio web completo, considerando solamente las páginas HTML, es de aproximadamente 0,8 MB. Cerca del 55% de los sitios web fueron creados en el 2004. De acuerdo con el robot de búsqueda la aplicación para servidor web más usado es Apache con un 70% de participación de mercado, y la segunda aplicación más usada es Microsoft IIS (Internet Information Server) con un 20 %. La distribución de sistemas operativos, en la que Unix y Linux tienen un 65% de participación, se puede inferir que al menos 1/5 de los servidores basados en Windows usaban Apache. Se encontraron más de 700.000 enlaces hacia páginas en otros países.

- **Chile 2006.**

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

Se observó un promedio de 43 páginas por sitio. Donde el 10% de los sitios de mayor cantidad de páginas contienen el 90% de los documentos. Por lo que existen muchos sitios que tienen muy pocas páginas, lo cual puede ser una señal de poco desarrollo de la Web. Además el idioma que más predomina es el español con alrededor del 80% de las páginas seguido del inglés con un 17%, Otros idiomas tienen una presencia muy leve.

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

disminución de los enlaces rotos puede significar que existe mayor conciencia respecto a verificar la validez de los enlaces.

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

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

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

- **Corea.**

El primer estudio sobre el espacio Web coreano fue efectuado en diciembre del 2004, en el cual se utilizo el WIRE rastreador, para obtener la información necesaria para la investigación. Donde fueron analizadas varias de sus características a partir de una muestra de más de 50 000 sitios los que poseen más de 8 millones de páginas, donde el 10% de los mayores sitios web contenían más del 85% del coreano web, lo que sugirió que la distribución de calidad era muy desigual, ya que solo algunas páginas poseían cierta relevancia. Los resultados también mostraron un predominio de formatos estándar

como XML o PDF, y herramientas de uso común como ASP y ZIP, en su mayoría pertenecientes al software propietario, lo que se explico debido al bajo apoyo de Asia al uso del código abierto.

- **Cuba.**

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

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

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

Dentro de los resultados se obtuvo, total de Acceso 33381340, Total de Archivos: 24 129194, Total de Páginas: 4723076, Total de Visitas: 1347647, Total Clientes: 804600, Total de URLs 274161, Total Pág. Entradas 42770, Total Pág. Salidas 44440.

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

- **España.**

La colección fue obtenida entre los meses de septiembre y octubre del 2004, utilizando un programa para recolectar páginas web. El cual comienza descargando un conjunto de

direcciones iniciales, que en su caso fueron obtenidas a partir de las referencias incluidas en el buscador Buscopio.

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

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

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

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

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

Se encontró aproximadamente 200 000 enlaces a ficheros que no eran HTML, lo que si bien es un número grande de documentos representa sólo un 1% de las páginas totales en la Web. Dentro de las características de los sitios tenemos que el número promedio de páginas por sitios es 52. Además los dos servidores dominantes eran Apache y Microsoft IIS (*Internet Information Server*), con ventaja para Apache. El sistema operativo más

usado para servidores era Windows (43%), seguido muy de cerca por sistemas operativos basados en Unix (41%); esto significa que al menos el 15% de los servidores basados en Windows prefieren Apache. En promedio se encontró 2.55 sitios por dominio. Hay un promedio de 133 páginas por dominio. El tamaño promedio de un dominio web completo considerando solamente el texto, es de aproximadamente 373 Kilobytes, sólo un 16% de los títulos son únicos.

- **Hungría.**

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

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

- **Perú.**

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

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

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

Se descargaron 1.629.745 páginas desde 8.908 sitios, que corresponden a 7.945 dominios de tercer nivel. Existen más del 55% de los sitios con un máximo de 10 páginas, Se halló 5.688 sitios (64%) sin enlaces entrantes lo que implica claramente posibles problemas de visibilidad. y 5.948 (68%) sin enlaces salientes lo que provoca una baja conectividad. Por lo que se podía apreciar que el espacio web de Perú se hallaba débilmente interconectado ya que la componente MAIN es pequeña y existe un 53% de sitios en la región islas.

- **Portugal.**

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

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

En este capítulo se realizó el estudio que da paso a la realización de la tesis, brindando información que de una forma u otra fueron de gran utilidad para el desarrollo de la investigación.

1.6 Conclusión.

El mismo arrojó datos importantes como la gran actividad que presenta la Web, y la abundancia de contenido, por no tener un punto de parada (stop) para detener el incremento de sitios en la misma, es que existen páginas que no están reconocidas y más bien pertenecen a las llamadas lagunas de la información. De aquí que surja la ciencia que estudie la web, la Webmetría, que a través de un robot de búsqueda realiza la recolección de páginas por sitios Web, y llevarlos así a un análisis cuantitativo a través de indicadores que ayuden a tener un punto de partida llamado caracterización o estudios de la Web. Este estudio puede llegar a ser engorroso para los que lo quieran desarrollar,

aunque existen países que lo han realizados, dentro de los que se encuentran los antes mencionados.

2 CAPÍTULO

Selección de los elementos cibermétricos para la investigación.

En el presente capítulo se darán ejemplos de los distintos aspectos a tener en cuenta para la realización de la caracterización como es el caso de los indicadores cibermétricos que describen su presencia en la Web. El objetivo perseguido es demostrar la variabilidad que pueden tener los indicadores en la evaluación de la Web de la UCI, después de que hayan sido seleccionados, y documentados de porque han sido escogidos. De igual manera se mostrará una representación de los principales Robot de búsqueda y así la elección del mismo para la recolección de datos para el estudio, el que sea capaz de dar una búsqueda profunda, que no deje URLs sin visitar, y brinde toda la información requerida para que sea analizada posteriormente.

2.1 Análisis de indicadores.

Agrupación de los indicadores según las variables o categorías más generales que permiten evaluar. Utilizando las siglas C para las categorías y la I para los indicadores.

C: Infraestructura:

I: Número de host, de servidores web, de usuarios, de dominios, de sitios, de sitios institucionales.

C: Tamaño:

I: Número de páginas, de objetos, de objetos multimedia, de archivos ejecutables, tamaño de los archivos, distribución por lenguajes, evolución temporal, número de niveles, de enlaces por página.

C: Calidad:

I: Porcentaje de enlaces válidos, de errores de enlace, apariencia.

C: Conectividad:

I: Total de enlaces, de enlaces por página, número de enlaces internos, de enlaces externos.

C: Visibilidad:

I: Número de enlaces recibidos o externos, enlaces nacionales externos, enlaces internacionales externos.

C: Impacto:

I: Factor de impacto.

C: Popularidad:

I: Número de visitas.

2.1.1 Indicadores más frecuentes empleados para el análisis métrico.

Los estudios realizados por diferentes autores para desarrollar sus análisis sobre La Web han tenido un incremento notable en los últimos años. El objetivo de estas investigaciones es obtener una serie de parámetros que permitan conocer la presencia de cualquier entorno en Internet, posibilitar el desarrollo de estudios de tendencias que abarquen campos como: dominios geográficos (.es, .fr, .cu.) y funcionales (.net, .org, .edu.), instituciones públicas y privadas, documentos electrónicos, espacios web científicos y comerciales, contextos sociales, regiones geográficas. Con los resultados obtenidos en estas investigaciones, es posible conocer la situación particular de una institución, documento, región o dominio determinado.

No abundan en la bibliografía sobre el tema, documentos que aborden la aplicación de los indicadores en el entorno digital, desde un punto de vista teórico o conceptual. La mayoría de los trabajos que se encuentran constituyen análisis de algunos indicadores que permitan identificar el comportamiento de determinados aspectos que sirven de base para tomar decisiones. Por esta razón, no se puede mencionar propuestas metodológicas concretas, que incluyan las propuestas de indicadores cibermétricos utilizados. En este

sentido, cada investigador, ha aplicado aquellos que ha estimado relevantes para cumplir con los objetivos de su investigación.

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

- Indicadores de tipos institucionales como la cantidad de páginas en determinados sectores.
- Indicadores regionales.
- Indicadores idiomáticos.
- Indicadores de tipología de sitios donde se encuentran los sitios académicos, comerciales, de sectores públicos o privado.
- Indicadores de tamaño, en sus dos variantes, tamaño documental como el número total de páginas comprendidas en un dominio o tamaño informático respondiendo al mismo el tamaño en bytes de una sede web.
- Indicadores de densidad, también tiene dos variantes, densidad hipertextual que es la media de enlaces por página y la densidad multimedia que es la media de objetos multimedia por página.
- Indicadores de profundidad siendo el número máximo de niveles de una sede.
- Indicadores de luminosidad comportándose como el total de enlaces emitidos desde una sede.
- Indicadores de visibilidad que no es más que el número total de enlaces externos diferentes recibidos por una sede, existen variantes para calcular la visibilidad nacional con límite a los enlaces recibidos en el mismo país.
- Indicador de navegabilidad siendo el número total de enlaces internos respecto al total de páginas.
- Indicadores de validez hipertextual es el porcentaje de enlaces válidos respecto al total.
- Indicador de diversidad es la distribución de las características de los enlaces recibidos por una página.
- Medidas de popularidad es el número y distribución de las visitas recibidas en un plazo determinado.
- Indicadores de impacto se presenta como el resultado de dividir el número total de

enlaces externos diferentes recibidos por una sede por su tamaño expresado en número de páginas.

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

2.1.2 Selección de los indicadores para la investigación.

El propósito del presente epígrafe es hacer una selección de los indicadores que tendrán como fin realizar un primer acercamiento cuantitativo y cualitativo de la estructura y contenido de la Web universitaria. Los mismos se agruparon en 3 niveles fundamentales.

Los indicadores que pertenecen a cada uno de los niveles tienen un objetivo especifico, a continuación se expondrá cada indicador, recogidos por el nivel que lo agrupa, además de ser acompañados del objetivo a alcanzar que tiene cada uno.

1. Nivel Colección:

    - Cifras Globales.

2. Nivel Página:

    - Edad de las páginas.
    - Profundidad de las páginas (La profundidad lógica de una página es el número de enlaces que es necesario seguir desde la portada de un sitio para alcanzarla).
    - Páginas dinámicas.
    - Documentos que no están en HTML.

        * Extensiones de imagen, video y audio.

        * Extensiones de documentos y documentos no HTML.

        * Extensiones de interfaz de entrada común.

        * Extensiones extras.

- Páginas descargadas versus enlaces inválidos.

3. Nivel Sitio:

- Tamaño total (Espacio en disco, porción que ocupa el texto en bruto).
- Promedio del tamaño del texto plano de los sitios en MB (Descartando etiquetas, imagen, o algún tipo de multimedios y formatos. Solo el texto de las páginas).
- Promedio del números de páginas (cantidad de páginas por sitios).
- Sitios con mayor cantidad de enlaces internos.
- Enlaces a dominios externos.
- Sitios con mayor cantidad de documentos.

Nivel Colección.

El indicador Cifras Globales es utilizado para que muestre una pequeña idea de cuántos sitios se hablará en la caracterización al igual de las páginas bajadas de una forma abreviada, pues los siguientes niveles son para detallarlos, sirve para que en los próximos estudios analicen la tendencia de La Web a incrementar o quedarse con el mismo número de sitios y de páginas.

Nivel Página.

El nivel página es el que le da paso al nivel sitio, al estar los sitios compuestos principalmente por las mismas. En este nivel se analizarán indicadores que describirán a las páginas, y darán una idea de cómo están conformadas hasta el momento, así se sabrá si el trabajo realizado valió la pena o si hay que perfeccionarlas.

Uno de los indicadores a tener en cuenta es la edad de las páginas que está muy ligado al desarrollo de la universidad, puesto que si las páginas no han sido creadas o modificadas en un intervalo de tiempo corto, entonces se puede decir que el desarrollo está bajo, sin embargo si las páginas están en constante cambio quiere decir que la Web esta evolucionando satisfactoriamente.

Los documentos que no están en HTML se analizan con un enfoque de verificar, si los enlaces a archivos que prevalecen, son los que están dirigidos a documentos HTML, o los que se dirigen a archivos como imagen, sonido, video o simplemente documentos. Con esta información se podrá entonces llegar a decir que tipo de extensión es la que prevalece después del HTML, además de la existencia de extensiones no conocidas.

El análisis de la profundidad de una página determinada o la media entre la mayoría es un indicador que se reduce a, si la misma es la portada de un sitio entonces tiene profundidad 1, las que se encuentran directamente alcanzables desde la portada presentan profundidad 2, y así sucesivamente, este indicador brinda la posibilidad de hacer el análisis de cuan engorroso puede ser alcanzar una página determinada o si la media de alcance a las páginas en general es fácil para los que acceden a ellas, con estas estadísticas se pueden hacer arreglos a los sitios, en específico a la manera de acceso a las páginas, se puede dar el caso de que una página de gran relevancia se encuentre con una profundidad muy grande y no sea de conveniencia para el administrador del sitio.

La existencia de las páginas dinámicas permite determinar la interacción existente entre los usuarios y los sitios web, lo que posibilita una mayor funcionalidad de los mismos. El presente indicador ayuda a tener conocimiento de cuanto ha evolucionado la Web con respecto a la Web estática.

Para medir la forma exitosa de llegar a una página se escogió el indicador páginas descargadas versus enlaces inválidos, ya que brinda las estadísticas necesaria para saber si las direcciones de URLs son válidas o realmente existen. En estos tiempos se trabaja con mucha agitación lo que provoca que en ocasiones no se realice el trabajo con calidad o simplemente no este completamente bien, ocasionando problemas a la hora de querer acceder a una página y no se pueda ya sea por que esta mal escrita o porque la misma no existe, provocando molestias para los usuarios. Los otros problemas a surgir a la hora de querer llegar a una página son los errores de servidor, o simplemente que haya que autenticarse porque todos los requerimientos no son permitidos.

La distribución de documentos por sitios, ayuda a tener una panorámica de cuales son los sitios que contienen mayor cantidad de contenido, de este indicador se puede deducir

que es la información que esta disponible mediante vínculos, pero que no son en si páginas web. Generalmente suelen ser aquellos ficheros que se pueden descargar o abrir, independientemente de que estos puedan o no cargar de una página web dinámica. En esos casos, son sitios que ocupan mucho espacio físico, pero que pueden tener pocas páginas web en su composición.

Los enlaces internos de un sito son los proporcionan mantener un nivel de intercomunicación o navegabilidad dentro de cada página de un sitio web, mientras más enlaces internos tengan, será mucho más fácil navegar dentro de cada una de sus páginas, suponiendo que estos no sean falsos, o enlaces rotos.

2.2 Herramientas tecnológicas empleadas para desarrollar estudios cibermétricos.

Hoy en día existe gran variedad de robot de búsquedas, pero no todos están disponibles para el uso abierto, un ejemplo claro es la existencia de un grupo de servicios de búsqueda en Internet, cada uno de los cuales tiene su propio Spyder o robot; pero los que utilicen ese servicio tienen que contentarse solo con los datos que estos facilitan, lo cual muchas veces no resuelve los problemas, puede darse el caso de que se necesite saber el número de páginas de un sitio determinado y que el robot no esté lo suficientemente robusto como para proporcionar este dato o la cifra exacta, esto sería en el caso de este indicador, pero hay otros indicadores de carácter significativo con los que desean realizar un estudio, que no se pueden utilizar para el análisis por el hecho de que no se tiene un dominio sobre el robot.

El comportamiento de un rastreador web o robot de búsqueda es el resultado de una combinación de cuatro políticas las cuales son:

- Política de selección: que indica las páginas a descargar.

- Política de revisión: revisa si ha existido algún tipo de cambio en el estado de las páginas

- Política de cortesía: que establece la manera de evitar la sobrecarga en los sitios web.

• Política de paralelismo: que establece la forma de coordinar los rastreadores web que están distribuidos en ese momento.

Política de Selección.

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

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

Política de revisión.

La Web tiene un carácter dinámico, y una fracción del rastreo de la Web puede llevar mucho tiempo, por lo general medido en semanas o meses si es muy grande la misma. En el momento en que un rastreador web ha acabado su rastreo, muchos acontecimientos podrían haber sucedido. Se caracterizan estos eventos como creaciones o actualizaciones.

Política de cortesía.

Es una norma para administradores para indicar qué partes de sus servidores web no deben ser visitados por los robots. La presente política no incluye una sugerencia para el intervalo de visitas a un mismo servidor, aunque este intervalo es la forma más eficaz de evitar la sobrecarga del servidor.

Política de paralelismo.

Un paralelo es un rastreador que ejecuta múltiples procesos en paralelo. El objetivo es aumentar al máximo la descarga y de reducir al mínimo los gastos generales de paralelización y de evitar una repetición de las descargas de la misma página.

2.2.1 Características de los robots de uso abierto.

Como mismo existen robots que son creados para el uso particular de la institución u organización, están los robots que sus creadores lo tiene para el uso público, los que a continuación se describen pertenecen de forma general a rastreadores con una breve descripción que incluye los nombres dados a los diferentes componentes y características.

WebBot:

El WebBot es un robot buscador en el Internet para obtener los documentos públicos (HTML, PDF) enmillones y millones de documentos de identificación de palabras clave y frases, que luego almacena la información y se ejecuta en contra de un complicado algoritmo que identifica el pensamiento subconsciente.

El proyecto fue desarrollado al finalizar el año 1990, según sus creadores, en un principio, este programa fue creado para realizar diversas predicciones para asuntos del mercado. Esto gracias a una tecnología llamada “sistema de araña”, que consiste en crear una red de búsqueda en Internet para encontrar determinadas palabras claves en diversas páginas web. Cuando la palabra es localizada, el programa toma una pequeña parte del texto de donde procede y ésta es enviada a una central en donde es filtrada y analizada para obtener de ella algunas conclusiones o predicciones básicas.

El WebBot trabajando en una web es muy rápido andador, con soporte para expresiones regulares y teniendo el SQL como registro. Se puede utilizar para comprobar los vínculos, encontrar páginas HTML, el mapa de un sitio en la Web y descargar imágenes.

Dentro de sus características más sobresalientes esta el apoyo a las expresiones regulares donde permite detallar el recorrido de las limitaciones de la Web. Apoyo a la

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. Otras de sus características es que sus posibilidades al hacer peticiones GET y HEAD son limitadas por lo que solo descarga lo estrictamente necesario.

Harvest-NG:

Harvest-NG es un conjunto de herramientas compatibles con los estándares del Web crawler. Fue desarrollado en Perl, aprovechando muchas de las actuales herramientas del lenguaje, y ayuda a que se proporcione una completa solución para el trabajo con un sitio web.

La versión actual de Harvest-NG ha recorrido un largo camino hacia el logro de preservar las características de la arquitectura a utilizar para la recolección de páginas teniendo un rápido desarrollo de prototipos además de tener un mejor código estructurado, lo que proporciona una base sólida sobre la construcción de los sistemas de descubrimiento de recursos. Estas son las dos funciones del proyecto, tanto para construir una forma rápida, resistente y eficiente en un sistema de indexación y, a través de este, para producir un conjunto de herramientas altamente extensible para la construcción de otros sistemas para alcanzar sus propósitos.

Harvest-NG está diseñado para ser capaz de trabajar en gran medida con la variedad de tipos de contenido compatible. Harvest-NG puede seguir URLs expresados en cualquiera de los formatos de documento de apoyo, no sólo los que figuran en los archivos HTML. Aunque carece de un plug-and-play (PNP) que es una tecnología que le permite conectarse a un ordenador sin la necesidad de configuración, por lo que se va a utilizar se necesita de un software o drivers específicos para configurar al mismo.

Webvac Spider:

Este robot es un repositorio de más de 110TB 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.

Webvac rastrea hasta una profundidad de 7 niveles tanto para páginas estáticas como dinámicas y obtiene un máximo de 10KB de páginas por sitio. Sólo sigue los vínculos del dominio. En estos momentos los rastreos los hace más estables sobre la lista de los sitios. Se demora de uno a diez segundos entre las páginas.

SocSciBot 3:

Es rastreador diseñado con fines de investigación. Junto con su apoyo a los programas de las Herramientas SocSciBot. Se puede utilizar para llevar a cabo análisis de enlace 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 y 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 de 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.

Ubicrawler:

Es un rastreador distribuido y escrito en Java, y 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 hots. 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 provocara que el otro agente deba volver a rastrear las páginas de la falta de agente. 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.

Wire Crawler:

El Wire como proyecto es un esfuerzo iniciado por el Centro de Investigación de la Web para crear una solicitud de recuperación de información, diseñado para ser utilizado en la Web local aunque posteriormente fueron desarrollándolo y haciendo nuevas versiones que han llegado a servir para estudios internacionales. El formato de datos de la recogida o depósito, está diseñado para escalar a varios millones de documentos Se compone de varios directorios. Cada directorio está apuntado por una variable de configuración.

El subsistema de almacenamiento es un archivo grande y contiene algunas estructuras de datos diseñadas para almacenar, de longitud variable que contiene los registros de páginas descargadas y detectar duplicados. Brinda soporte para archivos sobrecargados.

Actualmente incluye un formato simple para almacenar una colección de documentos web, así como un rastreador web, herramientas para extraer estadísticas de la colección y herramientas para la generación de informes sobre la misma. 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 que lo ayudan con el funcionamiento, normalmente el mismo funciona de forma reiterativa, los programas son Wire-Bot-reset, Wire-Bot- seeder, Wire-Bot-manager, Wire-Bot-harvester, Wire -Bot-Gatherer y Wire -Bot-Run.

Wire-Bot-reset:

Es un módulo que trabaja como un programa para resetear, borrando el repositorio, creando estructuras de datos y prepara todo lo necesario para un nuevo rastreo. Como algunas de las estructuras de datos requieren el espacio en disco que se asignará desde el principio, esto llevará algún tiempo dependiendo de su configuración de máximo de documentos y máximo sitios en el archivo de configuración.

Wire-Bot-seeder:

Recibe la URLs de los recolectores (o desde la URL inicial) y añade los documentos para el repositorio. Esto se utiliza tanto para dar al rastreador el conjunto inicial de direcciones URL como para analizar las URL que se extraen de los programas recolectores de las páginas descargadas.

Wire-Bot-manager:

En este tipo de programas los documentos de la colección son analizados para dar los "resultados" y crea un lote de documentos para el recolector, dentro de los que están las URLs que les suceden, a las ya analizadas. Los resultados se dan por una combinación de factores que se describen en el archivo de configuración.

Score functionScore function (Puntuación función):

La puntuación de cada una de las páginas se calcula tal como se define en el fichero de configuración, esta función incluye actualmente el PageRank, la profundidad y dinámica / estática páginas.

El gerente trata de determinar para cada documento que es la probabilidad de que el documento está obsoleto. Si esta probabilidad es, por ejemplo, 0,7, entonces su puntuación actual se define como 0,7 x Resultado.

Saber que si por cualquier motivo, usted debe cancelar el lote actual utilizando Wire-bot-manager - cancelar, cuando termine tiene que activar de nuevo o generar un conjunto de páginas que tienen la mayor diferencia entre los futuros y actuales resultado, las que serán seleccionadas para la próxima ronda de la cosechadora.

Wire -Bot-Harvester:

Este programa descarga los documentos de la Web. El programa trabaja en su propio directorio con sus propias estructuras de datos, y puede a continuación el lote actual utilizando Wire-bot-manager.

Wire -Bot-Gatherer:

Analiza el proceso de descargar documentos y extraer URLs De este modo, se descargan las páginas de la cosechadora de su directorio, y combina esas páginas en la colección principal.

Wire-Bot-Run:

Ejecutar varios ciclos rastreadores de la forma “seeder-manager- harvester – gatherer”.

2.2.*2* Robot seleccionado para la recolección de datos.

Al culminarse el análisis de los robots y haber analizado, las características que ofrecen por separado de cada uno de ellos para la realización de una recolección de datos, los objetivos que se trazaron sus creadores para su funcionamiento, la forma y sistematicidad en los que cada uno han sido utilizados para las diferentes caracterizaciones, dio como resultado que el más apropiado fuera el Wire Crawler para la realización de el estudio, debido a que proporciona la información necesaria para el análisis cuantitativo de la Web, al arrojar reportes estadísticos que son el pilar fundamental de cualquier estudio de una web.

El Wire Crawler es el robot de búsqueda más utilizado hasta el momento, para los diferentes estudios existentes, éste es uno de los factores que influyo en la selección del robot, aunque sus características se sobresalen antes los demás. La forma de escalar del mismo, es proporcionada por estar diseñado para trabajar con grandes volúmenes de documentos, llegando a millones. Fue desarrollado en C/C++ para un alto rendimiento, lo que permite la prestación del servicio de una forma eficiente. Es altamente configurable pues todos los parámetros para el rastreo y la indexación se puede configurar a través de un archivo XML. Incluye varias herramientas para analizar, extraer estadísticas, y la generación de informes sobre subconjuntos de la Web, por ejemplo: la Web de un país o institución.

El objetivo de utilizar el Wire Crawler es para descargar y hacer análisis de las colecciones, teniendo en cuenta un número de documentos no conocidos hasta el momento, puesto que la Web a la que se le va a hacer el estudio, la Web de la UCI es más grade que la mayoría de los sitios Web, pero más pequeña que una Web completa como la de un país.

El rastreador Wire incluye un módulo para generar informes y estadísticas acerca de la colección, que es una de las características por lo cual se seleccionó.

El procedimiento para la generación de informes tiene los siguientes pasos:

1. El análisis de los meta datos que se mantiene en las estructuras de datos y generación de estadísticas como archivos de texto.

2. Generación de scripts de gnuplot para generar gráficos, y la invocación de gnuplot.

3. Generación del informe utilizando LTEX.

Este procedimiento hace que el mantenimiento de los informes sea más fácil, ya que los datos se separan de la representación grafica. Los informes generados incluyen:

- Un informe sobre las características de las páginas que se descargan.
- Un informe sobre los enlaces que se encuentran en dichas páginas.
- Un informe sobre las lenguas.
- Un informe sobre los sitios Web.
- Un informe sobre los enlaces en el sitio Web.

No se escogió otro robot de búsqueda debido a las inconveniencias que presentan los restantes como por ejemplo el **SocSciBot 3** solo se ejecuta en Windows 95, sistema operativo propietario y la facultad tiene una política de uso de software libre, además no existe una garantía de que el mismo funcione como debe de ser, ni de que los resultados sean los esperados, sobre la recolección de datos; el **Webvac Spider** no se escoge, pues sus posibilidades a la hora de configurar están limitadas, al no permitir cambios necesarios en su configuración que permitan adecuarlo a las necesidades que se requieran para el estudio de cada web; el **Ubicrawler** debido que para su uso es necesario la autorización de sus creadores; el **WebBot** no se escogió pues no hace un buen uso de las peticiones GET y HEAD las cuales arrojan resultados necesarios para la investigación y por ultimo el Harvest-NG porque al no poseer el plug-and-play seria muy trabajoso su configuración.

2.3 Conclusión.

En el presente capítulo se escogieron los indicadores cibermétricos así como el robot de búsqueda que satisface las necesidades y requerimientos para desarrollar la caracterización de la Web universitaria. Teniendo en cuenta de que los indicadores que fueron seleccionados para el posterior estudio permitirán realizar un primer acercamiento de la Web, cumpliendo así con uno de los objetivos propuesto en la investigación.

3 CAPÍTULO

Análisis cualitativo y cuantitativo de la Web.

Como la parte más representativa de esta investigación se presentan los resultados sobre el estudio y trabajo de la caracterización del espacio web de la UCI, a los efectos de obtener una fotografía inicial sobre la cual identificar fortalezas, debilidades y oportunidades de crecimiento y evolución. Al conocer su estado de desarrollo se cuenta con una base para planificar acciones tendientes a expandir su tamaño y mejorar su calidad.

La colección fue obtenida el mes de Mayo de 2008, utilizando el programa para recolectar páginas web, WIRE Crawler. El recolector comenzó descargando un conjunto de direcciones iníciales que fueron escogidas de acuerdo a los resultados de las entrevista realizada al administrador de hosting de la universidad (anexo # 1), por ser los sitios que más enlaces internos y externos presentaba según su respuesta dentro de los 103 que están publicados, además de ser los más representativos para la UCI, siendo los que más visitados son por los usuarios dígase estudiantes, profesores e incluso trabajadores, estas direcciones pertenecientes a los sitos intranet.uci.cu, inter-nos.uci.cu, softwarelibre.uci.cu y wiki.prod.uci.cu las que corresponden al dominio que se posee en la universidad, de las páginas que se van descargando se van extrayendo nuevos enlaces, de los cuales se discriminan los que no apuntan a páginas o documentos pertenecientes al dominio .uci.cu.

3.1 Nivel colección.

En el presente nivel se resumen las principales estadísticas de la colección obtenida para la investigación las cuales se resumen en el indicador **Cifras globales**. Para el estudio de la Web, el Wire Crawler como todos los granes robots de búsqueda realizó una selección de las páginas más relevantes, las cuales no se toman solamente de forma aleatoria si no por la importancia de la misma, además respetó el protocolo de seguridad en el que los administradores deciden que parte de su sitio los robots no podrán entrar, cumpliendo así con el código de ética de un informático. Por lo que fueron descargadas 856 páginas para

el análisis detallado de las mismas en el nivel página, lo que representa un 4.25% de la cantidad total visitada que fueron 20157 páginas.

La tabla #1 muestra las características principales de las páginas descargadas.

La tabla #2 muestra las características principales de los sitios web.

	Cantidad	Por ciento
Páginas Web Descargadas	856	100%
Estáticas Dinámicas	403 453	47.08% 52.92%
Únicas Duplicadas	851 5	99.42% 0.58%

Tabla # 1: Resumen de la colección de páginas descargadas.

	Cantidad	Por ciento
Sitios Web Visitados Promedio de páginas por sitios	76 265.22	100%
Promedio de páginas estáticas por sitios	99.48	37.51%
Promedio de páginas dinámicas por sitios	165.74	62.49%

Tabla # 2: Resumen de la colección de los sitios web.

Las páginas visitadas ya sean descargadas o no corresponden a 76 sitios lo que representa al rededor de un 74%, de los sitios que se encuentran ubicados en los servidores UCI, el restante por ciento de los sitios que no fueron analizados fue porque no estaban referenciados, o se encontraban en mantenimiento.

3.2 Nivel Página.

En esta sección se presenta el análisis de las páginas de forma individual, sin considerar su agrupación en sitios o dominios. Primero se analiza el número de páginas descargadas correctamente. Luego se analizó la edad de las páginas, seguido de la profundidad de las mismas, más tarde el impacto de las páginas dinámicas y finalmente los documentos que no están en HTML.

Páginas descargadas versus enlaces inválidos.

De las páginas descargadas es frecuente que entre sus direcciones aparezcan páginas que ya no existen o que simplemente se escribieron mal. Cada vez que el recolector contacta con un servidor web, este retorna un código de estado que indica si la página existe o no, o si existe un motivo por el cual no se puede entregar el documento pedido.

La tabla # 3 muestra la distribución de páginas de acuerdo a estos códigos de estado, entre los que se encuentran:

- OK: incluye todos los requerimientos exitosos: OK (200).
- NOT FOUNT: el servidor no encuentra el documento pedido: NOT FOUND (404), ERROR CONNECT (97), ERROR TIMEOUT (95).
- MOVED: incluye todos los requerimientos en los cuales el servidor redirige al recolector a una u otra página: MOVED (301), FOUND (302).
- SERVER ERROR: incluye todas las fallas en el lado del servidor: ERROR Dns (98), SERVER ERROR (500), BAT REQUEST (400).
- FORBIDDEN: incluye todos los requerimientos que no son permitidos, principalmente por tratarse de páginas con clave: UNAUTHORIZED (401) y NOT ACCEPTABLE (406).

De acuerdo a los resultados obtenidos se puede observar que las páginas descargadas presentan un buen por ciento de enlaces válidos, lo que implica el buen acceso a las mismas por parte de los usuarios, al igual que el bajo promedio de páginas en las que hay que autenticarse para poder acceder a ellas lo que permite una mayor navegabilidad por parte de los usuarios siendo menos trabajoso y asequible acceder a un sitio. El bajo promedio de los enlaces rotos puede significar que existe mayor conciencia a la hora de la realización de un sitio respecto en la verificación y correctitud de los enlaces en los mismos

Estado Http	Código Http	Documentos	Por ciento
Ok	200	629	73.48%
Found	302	89	10.40%
Moved	301	78	9.11%
Not Found	404	39	4.56%
Unauthorized	401	9	1.05%
Error Dns	98	3	0.35%
Not Acceptable	406	3	0.35%
Error Connect	97	2	0.23%
Internal Error	500	2	0.23%
Error Timeout	95	1	0.12%
Bad Request	400	1	0.12%

Tabla # 3: Distribución del código de estado HTTP.

Edad de las páginas.

Para determinar la edad de las páginas, el Wire se basa en la fecha de última modificación entregada por los servidores web para cada una de ellas.

La distribución de las edades de las páginas en términos de años se muestra en la Tabla # 4.

Partiendo que la Web de la UCI, tiene alrededor de 6 años de existencia, es significativo que todas las páginas escogidas para el estudio han sido modificadas o creadas en los últimos 3 años, y de forma más especifica el 90.48% del total fue en menos de un año, lo que indica que la Web universitaria está en constante transformación y crecimiento, lo

que posibilita que exista disponible para la comunidad de la UCI un mayor caudal de información cada ves más actualizada.

Edad en años	Por ciento
0	90.48%
1	8.66%
2	0.43%
3	0.43%

Tabla # 4: Por ciento de edad de las páginas en años de las páginas descargadas.

Profundidad de las páginas.

El navegador se configuró para que descargara hasta cinco niveles de páginas estáticas, y 15 niveles de páginas dinámicas. Basada en el número de enlaces que hay que pulsar para llegar al contenido. En general los usuarios prefieren sitios poco profundos. Una buena regla a seguir es que el usuario no tenga que pulsar más de 3 enlaces para encontrar lo que busca.

La distribución de páginas por su profundidad se muestra en la Tabla # 5.

Profundidad de las páginas.	Páginas	Por ciento
1	65	7.59%
2	156	18.23%
3	635	74.18%

Tabla #5: Distribución de las páginas por su profundidad.

Se observó que aproximadamente las tres cuartas partes de las páginas descargadas tienen tres niveles de profundidad lo que, significa que el contenido de la misma es de gran importancia para los usuarios o simplemente para el administrador del sitio, debido a que la principal razón por la cual las páginas se colocan a un pequeño valor de profundidad, es para que las mismas puedan ser a accedidas de una forma más rápida y sencilla. Pues esta demostrado que es más fácil memorizar la dirección de URL de un sitio que de una página en particular, por lo que si de la portada de un sitio hay que realizar un proceso de navegación muy profundo para alcanzar la información que el usuario solicita, el mismo puede desistir o en el mejor de los casos que concrete la búsqueda pero siempre se ira insatisfecho o con una mala impresión de lo que al sitio se refiere.

Páginas Dinámicas.

De las páginas descargadas 453 fueron páginas dinámicas, lo que representa un 52.92%, es decir, páginas generadas en el momento de ser solicitadas sin que existieran previamente. Esto es lo normal cuando hay una consulta a una base de datos, involucrada en el proceso de desplegar las páginas.

Con las páginas dinámicas se mejora la interacción entre el usuario y el sitio web, lo que permite que exista un incremento en la aceptación por parte de los usuarios hacia las mismas, al incluir novedosas funcionalidades para las que son necesarias la utilización de otros lenguajes de programación aparte del simple HTML, como la programación especial en PHP y ASP. Manifestándose el incremento de los conocimientos obtenidos, por parte de los desarrolladores de los sitios web de la universidad, demostrando cuanto ha evolucionado la UCI en cuanto a las herramientas y técnicas para la realización de sus sitios web, proporcionando así un mayor desarrollo tecnológico.

Se estima que el número de páginas dinámicas seguirá creciendo, debido a la tendencia actual de tener sitios cuyos contenidos se puedan administrar en línea y que sean independiente del diseño y de la estructura de los documentos, al ser más fácil y práctico tener el contenido de un sitio en una base de datos que en archivos HTML que resultan toscos a la hora de modificarlos para ingresar o modificar información. También se debe

considerar que existen páginas estáticas, con extensiones HTML, que son generadas por procesos en lote en los servidores que se ejecutan constante y automáticamente.

Documentos que no están en HTML.

El robot encontró en la colección descargada 13957 de enlaces a documentos en formatos distintos a HTML (Ver anexo 1). Las extensiones o archivos más populares se recogieron por grupos que serán explicados por separados.

Imagen, video y audio.

De los 7883 enlaces existentes a archivos de multimedia, se encontraron 27 enlaces a archivos de audio, 126 enlaces a archivos de video, y 7730 enlaces a imágenes. La distribución de formatos de archivo se muestra en la Tabla # 6.

En audio, el formato MP3 es el más representativo en la Web, con un 96.30% (Anexo # 2) probablemente debido al auge de los reproductores portátiles, el formato MP3 es el estándar utilizado para oír un archivo directamente en una página web, sin necesidad de descargarlo antes al ordenador y poseer una compresión de audio de alta calidad gracias a la posibilidad de ajustar la calidad de la compresión, y por tanto el tamaño final del archivo, que podía llegar a ocupar 12 e incluso 15 veces menos que el archivo original sin comprimir.

En vídeo, el formato WMV fue el único reconocido dentro de las páginas descargadas con 126 referencias, debido a que este formato es un excelente reproductor de video, con mínimo uso del CPU a fin de dejarlo libre para otras tareas, sin perder frames ni calidad de imagen, lo que permite que la Web de la UCI proporcione un buen servicio audiovisual en la red.

El formato de imagen es el más usado en la Web UCI, con más del 98% del total de archivos de multimedia. Entre los más representativos se encuentran el GIF con un 63.08% de los enlaces, seguido del PNG con 29.84%, con respecto a los formatos de imagen.

Este alto por ciento se debe a que GIF es muy usado a la hora de diseñar una página, debido a que presentan compresión sin pérdida y que pueden incluir animación. Esta extensión es independiente de plataforma, lo que posibilita a la universidad utilizarlos en una simple PC siempre y cuando la misma posea un visor de imágenes. Es un formato de imágenes diseñado para minimizar el tiempo de transferencia de archivos sobre la red.

Por otro lado la presencia de la extensión PNG se debe a que es el formato de archivo nativo de Macromedia Fireworks. Los archivos PNG conservan la información original de capa, vector, color y efectos (como por ejemplo las sombras), y todos los elementos pueden editarse siempre que se desee. Los archivos se deben guardar con la extensión .PNG para que Dreamweaver pueda reconocerlos como tales.

Nombre de archive	Archivos	Por ciento
Gif	4876	61.85%
Png	2307	29.27%
Jpg	472	5.99%
Wmv	126	1.60%
Ico	75	0.95%
Mp3	26	0.33%
Wma	1	0.01%

Tabla # 6: Distribución de enlaces a distintos archivos multimedia.

Extensiones de documentos y documentos no HTML.

De los 1199 enlaces existentes a documentos, se encontraron 464 enlaces a documentos y 735 enlaces a documentos no HTML. La distribución de formatos de archivo se muestra en la Tabla # 7.

La presencia de extensiones de documentos es de gran importancia para la UCI, pues son éstos los tipos de archivos que brindan mayor información, debido que están conformados mayormente por texto.

Entre los documentos el más utilizado fue el PDF con un 60.55%, seguido del XHTML con un 34.78%. La superioridad del PDF sobre los restantes tipos de documentos se debe a que el mismo permite realizar cualquier combinación de texto, gráficos, imágenes e incluso música. Por el hecho de ser multiplataforma, el administrador del un sitio puede publicar cualquier información sin el temor a que la misma no pueda ser consultada o simplemente se modifique el aspecto y la estructura del documento. Además de permitir proteger el contenido mediante un método de cifrado e incluso firmarla digitalmente.

Mientras que el XHTML debe su presencia a que le proporciona a la Web universitaria compatibilidad con navegadores, además de poder adoptar diseños distintos en diferentes dispositivos y de brindar facilidad de mantenimientos a los administradores de los sitios en el momento de hacer algún cambio a la información que este contiene.

Nombre de archive	Archivos	Por ciento
Pdf	726	60.55%
Xhtml	417	34.78%
Txt	47	3.92%
Doc	8	0.67%
Ppt	1	0.08%

Tabla #7: Distribución de enlaces a documentos, excluyendo enlaces a páginas HTML.

Extensiones de interfaz de entrada común.

Se encontró dentro de las páginas descargadas un total de 4548 extensiones de interfaz de entrada común (CGI), donde el que prevalece es el PHP con total de 3397 enlaces para un 74.69%. La tabla # 8 muestra la distribución de enlaces por cada una de las interfaces de entrada común.

Partiendo de que prevalece en la Web de la UCI las páginas dinámicas, para lo cual el factor principal para su elaboración es la utilización del PHP, pues el mismo permite la

combinación con el motor de base de datos MySQL, aunque cuenta con soporte para otros motores, lo que amplía en gran medida sus posibilidades de conexión. Es un lenguaje libre, lo que permite la reutilización del código, representando una alternativa de fácil acceso para todos, además de proporcionar a los desarrolladores las ventajosas técnicas de Programación Orientada a Objetos.

Nombre de archive	Archivos	Por ciento
Php	3397	74.69%
Asp	1142	25.11%
Shtml	4	0.09%
Cgi	2	0.05%
Cfm	1	0.02%
Jsp	1	0.02%
Pl	1	0.02%

Tabla # 8: Distribución de los enlaces de interfaz de entrada común.

Extensiones extras.

Por encontrase dentro de las 10 extensiones con mayor número de enlaces encontrados con un total de 252 enlaces es importante determinar su impacto sobre la Web. La tabla #9 muestra la distribución de enlaces a extensiones extras.

Nombre de archive	Archivos	Por ciento
Css	129	51.19%
Swf	123	48.81%

Tabla # 9: Distribución de extensiones extras.

La presencia de los CSS en la Web de la UCI est`´a dada por proporcionarle un control centralizado de la presentación de un sitio web completo con lo que se agiliza de forma considerable la actualización del mismo, permitiendo a los usuarios a través del navegador especificar su propia hoja de estilo local que será aplicada a un sitio web, con lo que aumenta considerablemente la accesibilidad. Además se utiliza SWF para crear archivos pequeños, pero que a su ve permite la interactividad y que funciones en cualquier plataforma, el mismo proporciona poder ser transmitir gráficos sobre un ancho de banda reducido.

3.3 Nivel sitio

Una vez culminado el estudio de las páginas, elemento principal de la composición de los sitios web, se adentrará la investigación al nivel sitio, profundizándose en los temas como la porción que ocupa el texto en bruto, al igual que el promedio del texto plano, así como la cantidad promedio de páginas por sitios y la profundidad de los mismos. Se abordarán los enlaces internos como externos y los sitios con mayor cantidad de documentos.

Tamaño total

Se considera que el tamaño total de un sitio es la suma de todos los elementos que lo componen, dígase texto imagen, videos, audio y gráficos. En la Tabla # 10 se listan una relación de los 10 sitos por su tamaño en bruto de los cuales el Wire recolectó sus páginas. Donde se aprecia una marcada presencia de sitios en los que su principal función es el apoyo a la docencia, dando una panorámica de que en función de ella es que se desarrolla la Web de la universidad.

Sitios por contenido en bruto	Contenido en bruto
softwarelibre.uci.cu	16,167,252
inter-nos.uci.cu	1,954,451
wiki.prod.uci.cu	1,544,302
intranet.uci.cu	1,326,106
laboratorios.uci.cu	31,162
softwarelibre.hab.uci.cu	25,329
debian.prod.uci.cu	22,439
seriecientifica.uci.cu	22,423
investigaciones.uci.cu	22,344
pase.uci.cu	17,310

Tabla # 10: Distribución de los 10 sitios con mayor contenido en bruto.

Promedio del tamaño del texto plano de los sitios en MB.

En esta sección se analizará solamente el texto de las páginas que fueron recolectadas: para determinar el tamaño de un sitio sólo se considera el tamaño de los documentos HTML, no el de sus imágenes u otros documentos o archivos multimedia.

El tamaño promedio del texto plano en la Web UCI es muy pobre, siendo el mismo 0.78MB, de lo que se puede interpretar que los sitios se tienen a sobrecargar de multimedios y formatos que aunque puede brindar una información nunca puede ser comparada con la claridad que la información en forma de texto, llega al usuario, además muchas de las páginas que componen a los sitios están solo conformadas por etiquetas HTML, que no brindan ninguna información.

Promedio del números de páginas.

Después del recorrido del Wire por la Web universitaria, se obtuvo que el promedio de páginas por sitios se puedan dividir en las estáticas y las dinámicas, el promedio de páginas estáticas fue 99.48, mientras que el de las dinámicas 165.74, para un total de 265.22.

Si se parte del análisis del por ciento total de páginas se puede decir que es un porciento alto, lo que demuestra el nivel de avance que se tiene en el desarrollo de la Web, pues supera la Web de países como Chile, que contaba con una distribución de 43 páginas por sitios en el año 2006, evidenciándose que la universidad no sigue la política de libertad de publicación, sino que antes de ser publicados los sitios son pasados por especialistas del tema, lo que permite que no se encuentren sitios solo con una página o sin información alguna.

Sitios con mayor cantidad de enlaces internos.

Un enlace se considera interno si apunta a otra página dentro del mismo sitio web. Los sitios de la universidad presentan un promedio de 594 enlaces internos. Un sitio web tiene como promedio aproximadamente 2.24 enlaces internos por página. Además existen muchos sitios con un gran número de enlaces internos

La tabla # 11 muestra la distribución de los 10 sitios con mayor cantidad de enlaces internos de los que el Wire descargo páginas.

Sitios por enlaces internos	Cantidad de enlaces internos
softwarelibre.uci.cu	21,484
wiki.prod.uci.cu	2,340
intranet.uci.cu	970
inter-nos.uci.cu	707
softwarelibre.hab.uci.cu	82
www.uci.cu	64
isos-linux.prod.uci.cu	39
intranet.hab.uci.cu	36
feu.uci.cu	36
facultad1.uci.cu	36

Tabla #11: Distribución de los 10 sitios con mayor cantidad de enlaces internos.

La distribución del número de enlaces internos está relacionada con la distribución de páginas por sitio, por lo que el valor obtenido por el Wire se encuentra en una buena proporción con respecto a la cantidad promedio de páginas pudiéndose decir que las

mismas presentan en su estructura los enlaces necesarios que le permiten a los usuarios navegar por el contenido de las páginas con mayor facilidad.

Enlaces a dominios externos.

Se encontró un total de 30531 enlaces hacia páginas que no pertenecen al dominio de la universidad. Los 25 dominios referenciados desde la Web de la UCI muestran en la Tabla 12.

Los enlaces tienen destinos heterogéneos, al contrario de lo que podría pensarse, lo que permite que los usuarios a través de la Web de la uci puedan acceder a otros sitios no pertenecientes al dominio de la misma, ya sean estos nacionales o internacionales. Los cuales son un elemento fundamental para el apoyo a la docencia y el proceso investigativo.

Dominios de nivel superior	Número de enlaces externos	Por ciento
CU – Cuba	28,472	93.26 %
ORG	1,316	4.31 %
COM	617	2.02 %
NET	46	0.15 %
ES – Spain	29	0.09 %
AR – Argentina	9	0.03 %
MX – México	6	0.02 %
CL – Chile	5	0.02 %
DE – Germany	5	0.02 %
EDU	5	0.02 %
AU – Australia	3	0.01 %
BR – Brazil	2	0.01 %
FI – Finland	2	0.01 %
GOV	2	0.01 %
IT – Italia	2	0 %
CA – Canadá	1	0 %
CO – Colombia	1	0 %
FR – France	1	0 %
INFO	1	0 %
INT	1	0 %
LU – Luxembourg	1	0 %
NL – Netherlands	1	0 %
PE – Perú	1	0 %
UK	1	0 %
VE – Venezuela	1	0 %

Tabla #12: Distribución de enlaces externos a dominios de nivel superior.

Sitios con mayor cantidad de documentos.

Dentro de los resultados arrojados por el Wire, es importante destacar los sitios con mayor cantidad de documentos, puestos que son estos los que presentan el mayor número de información de relevancia en la Web universitaria, para un total de 6256 documentos.

La tabla #13 muestra la distribución de documentos por sitios Web.

Sitios por cantidad de documentos	Cantidad de documentos
softwarelibre.uci.cu	3,482
wiki.prod.uci.cu	1,768
intranet.uci.cu	265
softwarelibre.hab.uci.cu	85
www.uci.cu	66
inter-nos.uci.cu	51
isos-linux.prod.uci.cu	41
feu.uci.cu	39
intranet.hab.uci.cu	38
facultad1.uci.cu	38

Tabla # 13: Distribución de documentos por sitios.

Se evidencia que según la muestra bajada, no hay equilibrio entre los sitios, puestos que unos tienen mucha información y otros no presentan casi ninguna, aunque esto es posible que ocurra debido a que cada sitio tiene dentro de la universidad un fin determinado, sin dejar de destacar que la importancia de los sitios no esta dada por la cantidad de documento que presente, sino por la utilidad y servicios que ofrecen los mismos.

Es significativo destacar que entre los sitios que mayor cantidad de documentos posean, se encuentren los pertenecientes al tema del software libre, lo que demuestra la política de la universidad para una futura migración.

3.4 Conclusiones.

En este capítulo se analizaron los datos obtenidos de la web de la UCI tanto cualitativo como cuantitativo concretándose el objetivo general de la investigación. El estudio demostró no sólo el avance tecnológico sino también el incremento de los conocimientos por parte de los desarrolladores, al observarse una transición de la Web, dejando se ser una Web estática para convertirse en una dinámica.

CONCLUSIONES GENERALES

El objetivo de este trabajo era presentar algunos indicadores de carácter cibermétrico con el fin de mostrar su interés y capacidad para describir la estructura y contenido de la Web, aplicándolos a un caso particular especialmente interesante como la Universidad de las Ciencias Informáticas.

La investigación permitió tomar una fotografía de la Web de la UCI durante el mes de mayo de 2008, utilizándose el Wire Crawler como rastreador web. El presente estudio sirvió para mostrar el desarrollo tecnológico que presenta la Web universitaria.

Una particularidad en la Web UCI es que una cantidad considerable de páginas referenciadas incluyen en sus URLs los requerimientos necesarios de forma exitosa, lo que muestra que se está haciendo un buen trabajo a la hora de enlazarlas por parte de los desarrolladores. Al encontrarse las páginas con un máximo de tres niveles de profundidad así como su constante cambio y transformación son puntos a favor que demuestran su desarrollo, permitiendo a la Web una excelente navegabilidad y el fácil acceso por parte de los usuarios además de brindar nuevos servicios e información actualizada.

Otra de sus particularidades es que la media de páginas por sitios es alta, lo que demuestra que la realización de los mismos no es solo para publicar y tener un espacio en la Web universitaria sino que son realizados con un objetivo en específico.

RECOMENDACIONES

- Se recomienda al Grupo de Cibermetría Aplicada (CIBA) que tomen el presente trabajo como punto de partida para posteriores estudios, con los que podrán realizar análisis de tendencias.
- Deberán profundizar en nuevos indicadores que permitan realizar un análisis más profundo.
- Desarrollar una herramienta propia para el estudio de la Web, que cumplan con los requisitos necesarios para el rastreo.

REFERENCIA BIBLIOGRÁFICA

1. **Woodcock.J.** *Diccionario de Informática e Internet de Microsoft.* Madrid : Mc Graw Hill, 2001. págs. 621-622.

2. —. *Diccionario de Informática e Internet de Microsoft.* Madrid : Mc Graw Hill, 2001. pág. 626.

3. Web and Macros. *Web Estática (Definición, Ejemplos, Blog, Webquest...).* [En línea] 2006. [Citado el: 5 de Diciembre de 2007.] http://www.webandmacros.com/webestatica.htm.

4. Web and Macros. *Web Dinámica (Definición, Ejemplos, Aplicaciones Web...) .* [En línea] 2006. [Citado el: 5 de Diciembre de 2007.] http://www.webandmacros.com/webdinamica.htm.

5. **Van.Ch.** Maestros del Web. *¿Qué es la Web 2.0?* [En línea] 27 de Octubre de 2005. [Citado el: 7 de Diciembre de 2007.] http://www.maestrosdelweb.com/editorial/web2/.

6. **Aguillo.I.** *Indicadores de contenidos para la web académica iberoamericana.* [En línea] Diciembre de 2005. [Citado el: 10 de Diciembre de 2007.] http://www.ub.es/bid/pdf/15aguil2.pdf. ISSN 1575-5886.

7. **Blázquez.M.** Documentación Sin Límites. *Cibermetria y sucedáneos. Primera Parte.* [En línea] 14 de Julio de 2005. [Citado el: 10 de Diciembre de 2007.] http://docunlimited.blogsome.com/2005/07/14/cibermetria-y-sucedaneos-primera-parte.

8. **Arroyo.N, Ortega.J, Pareja.V, Prieto.J, Aguillo.I.** *Cibermetría. Estado en cuestión.* [En línea] 14-15 de Abril de 2005. [Citado el: 13 de diciembre de 2007.] http://eprints.rclis.org/archive/00007206/01/ArroyoEtAl_FESABID2005.pdf.

BIBLIOGRAFÍA

2007. *Web Bot Predictions.* [En línea] 23 de Mayo de 2007. [Citado el: 11 de Febrero de 2008.] http://aphroditeastrology.com/2007/05/web-bot-predictions.html.

2008. *El WebBot y la fecha del fin del mundo.* [En línea] 2008. [Citado el: 13 de Marzo de 2008.] http://laterceraguerramundial.blogspot.com/2008/02/el-web-bot-y-la-fecha-del-fin-del-mundo.html.

2005. *Web Hosting support front page extension WebBot.* [En línea] 2005. [Citado el: 20 de Febrero de 2008.] http://www.hostitwise.com/frontpage/web-bots.html.

2003. *Solución para: ¿Qué significa "WebBot" en el código de los formularios?* [En línea] 2003. [Citado el: 18 de Febrero de 2008.] http://office.microsoft.com/es-es/frontpage/HA011430013082.aspx.

Harvest-NG. [En línea] [Citado el: 20 de Febrero de 2008.] http://webharvest.sourceforge.net/ng/.

2006-2007. *Harvest-NG.* [En línea] 2006-2007. [Citado el: 18 de Febrero de 2008.] http://www.duamu.com/re/script/1122/id/2292/scripts-harvest-ng.html.

Fetching Web Pages from the WebBase Web Page Repository. [En línea] [Citado el: 20 de Febrero de 2008.] http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/webbase-pages.html.

Aguillo.I. 2004. *Indicadores y posicionamiento Web para inventores cientificos.* [En línea] 19 de Abril de 2004. [Citado el: 15 de Diciembre de 2007.] http://www.ibt.unam.mx/biblioteca/presentacion_IBT.ppt.

—. 2004. *Indicadores cibermétricos. Midiendo y evaluando los contenidos de la Sociedad de la Información.* [En línea] 15 de Abril de 2004. [Citado el: Diciembre de 10 de 2007.] http://www.eventos.bvsalud.org/INFO2004/docs/es/Indicadores_SI.ppt.

—. 2006. *Estudios de producción científica a través de la Web.* [En línea] 16-18 de Enero de 2006. [Citado el: 10 de Diciembre de 2007.] http://www.cincel.cl/documentos/Recursos/ISidroAguillo.ppt.

—. 2004. *POSICIONAMIENTO EN EL WEB DEL SECTOR ACADEMICO IBEROAMERICANO.* [En línea] 15-17 de Septiembre de 2004. [Citado el: 11 de Diciembre de 2007.] http://www.ricyt.org/interior/normalizacion/VItaller/S5_produc/aguilloppt.pdf.

—. 2006. *Indicadores Web de actividad científica formal e informal en Latinoamérica.* [En línea] 17-21 de Abril de 2006. [Citado el: 6 de Diciembre de 2007.] http://www.congreso-info.cu/UserFiles/File/Info/Info2006/Ponencias/1.pdf.

—. 2006. *Cibermetría Introducción teórico-práctica.* [En línea] Noviembre de 2006. [Citado el: 10 de Enero de 2008.] http://internetlab.cindoc.csic.es/cursos/cibermetria.pdf.

—. 2000. *Cibermetría: La métrica de la Web.* [En línea] 2000. [Citado el: 10 de Enero de 2008.] http://www.archivovirtual.org/seminario/busqueda/ponencias/p2.htm.

—. 2002. *Indicadores de POPULARIDAD, VISIBILIDAD, IMPACTO en el Web académico y de investigación. .* [En línea] 2002. [Citado el: 12 de Enero de 2008.] http://www.eicstes.org/EICSTES_PDF/PRESENTATIONS/Indicadores%20de%20popularidad,%20visibilidad,%20impacto%20en%20el%20web%20acad%C3%A9mico%20y%20de%20investigaci%C3%B3n%20(Aguillo).PDF.

—. 2006. *Cibermetría. Introducción teórico-práctica.* [En línea] Noviembre de 2006. [Citado el: 13 de Diciembre de 2007.] http://eprints.relis.org/archive/00008231/01/cibermetr%c3%ADa.pdf.

Aguillo.I, Costa.J. 2002. *MEDIDAS DE POPULARIDAD, VISIBILIDAD, IMPACTO Y DIVERSIDAD EN LA DESCRIPCIÓN CUANTITATIVA DEL WEBESPACIO ACADÉMICO Y DE INVESTIGACIÓN.* [En línea] 25 de Abril de 2002. [Citado el: 15 de Enero de 2008.] http://www.congreso-info.cu/UserFiles/File/Info/Info2002/Ponencias/208.pdf.

Aguillo.I, Granadino.B. 2006. *Indicadores web para medir la presencia de las universidades en la Red.* [En línea] 1 de Abril de 2006. [Citado el: 16 de Diciembre de 2007.] http://www.uoc.edu/rusc/3/1/dt/esp/aguillo_granadino.pdf.

Alonso.J. *¿Cómo trabajar en el grafo web?* [En línea] [Citado el: 27 de Enero de 2008.] http://www.fesabid.org/madrid2005/descargas/presentaciones/actividades/alonso_jl.pps.

Arroyo.N, Pareja.V. *Metodología para la obtención de datos con fines cibermétricos.* [En línea] [Citado el: 10 de Diciembre de 2007.] http://internetlab.cindoc.csic.es/varios/Metodolog%EDa%20datos%20ciberm%E9tricos.pdf.

Baeza.R, Castillo.C. 2000. *Caracterizando la Web chilena 2000.* [En línea] 2000. [Citado el: 12 de Noviembre de 2007.] http://www.cwr.cl.

—. 2005. *Caracterizando la Web chilena 2004.* [En línea] 2005. [Citado el: 12 de Noviembre de 2007.] http://www.ciw.cl/webcl2004/.

Baeza.R, Castillo.C, Graells.E. 2007. *Características de la Web chilena 2006.* [En línea] Marzo de 2007. [Citado el: 12 de Noviembre de 2007.] http://www.ciw.cl/material/web_chilena_2006/index.html.

Baeza.R, Castillo.C, Lalanne.F, Dupret.G. 2004. *Comparing the Characteristics of the Korean and the Chilean Web.* [En línea] Diciembre de 2004. [Citado el: 20 de Enero de 2008.] http://www.chato.cl/papers/baeza_04_comparing_chilean_web_korean_web.pdf.

Baeza.R, Poblete.B, Jean.F. 2003. *Evolución de la Web chilena 2001-2002.* [En línea] Enero de 2003. [Citado el: 12 de Noviembre de 2007.] http://www.ciw.cl/recursos/estudio2002/index.html.

Benezúr.A. 2003. *Searching a small national domain--Preliminary report.* [En línea] Mayo de 2003. [Citado el: 21 de Enero de 2008.] http://www.ilab.sztaki.hu/websearch-data/Publications/p184-benczur.html.

Boldi.P, Codenotti.B. 2002. *UbiCrawler: A Scalable Fully Distributed Web Crawler.* [En línea] 2002. [Citado el: 13 de marzo de 2008.] http://vigna.dsi.unimi.it/ftp/papers/UbiCrawler.pdf.

Boldi.P, Codenotti.B, Santine.M, Vigna.S. 2002. *Structural Properties of the African Web.* [En línea] Mayo de 2002. [Citado el: 20 de Enero de 2008.] http://www2002.org/CDROM/poster/164/.

Castillo.C. 2004. *EffectiveWeb Crawling.* [En línea] Noviembre de 2004. [Citado el: 14 de Diciembre de 2007.] http://chato.cl/research/crawling_thesis.

Castillo.C, Baeza.R, López.V. 2005. *Características de la Web de España.* [En línea] Junio de 2005. [Citado el: 20 de Enero de 2008.] http://eprints.rclis.org/archive/00009297/.

2006. CÓMO BUSCAR INFORMACIÓN EN INTERNET. [En línea] 25 de Junio de 2006. [Citado el: 12 de Febrero de 2008.] http://www.ual.es/Universidad/Biblioteca/turcana/Usuario/Investigacion/inf_internet.htm.

Cotin.A, Valdés.M. *Estudio de las estadísticas Web de accesos y visitas del Portal Cuba.cu.* [En línea] [Citado el: 22 de Enero de 2008.]

http://www.congreso-info.cu/UserFiles/File/Info/Info2004/Ponencias/107.pdf.

cwr.cl. *WIRE - Web Information Retrieval Environment.* [En línea] [Citado el: 30 de Noviembre de 2007.] http://www.cwr.cl/projects/WIRE/.

Fernandez.H. *Motores de búsqueda. Tipos de buscadores.* [En línea] [Citado el: 20 de Febrero de 2008.] http://www.buscarportal.com/articulos/motores_busqueda.html#spiders.

Figuerola.C, Alonso.J, Zazo.A, Rodríguez.E. 2006. *Diseño de spiders.* [En línea] Marzo de 2006. [Citado el: 14 de Diciembre de 2007.] http://reina.usal.es/pub/figuerola2006diseno.pdf.

Gómez.D, Silva.M. 2003. *A characterization of the portuguese web.* [En línea] 2003. [Citado el: 21 de Enero de 2008.] http://xldb.fc.ul.pt/daniel/webarchive2003.ppt.

Hopkins.J. *SocSciBot 3 and SocSciBot 4.* [En línea] [Citado el: 13 de marzo de 2008.] http://socscibot.wlv.ac.uk/.

Lopez.M. 2007. *La Cibermetría, una Nueva Alternativa para Evaluar la Visibilidad de la Publicación Académica Electrónica. El caso de la REDIE1.* [En línea] Agosto-Septiembre de 2007. [Citado el: 16 de Febrero de 2008.] http://www.cem.itesm.mx/dacs/publicaciones/logos/anteriores/n58/mlopez.html.

masadelante.com. *¿Cómo funciona un motor de búsqueda?* [En línea] [Citado el: 10 de Febrero de 2008.] http://www.masadelante.com/fac-como-funciona-motores-de-busqueda.htm.

Mendoza.J. 2001. *Definiendo el género de su sitio web.* [En línea] 1 de Marzo de 2001. [Citado el: 26 de Enero de 2008.] http://www.informaticamilenium.com.mx/paginas/mn/articulo33.htm.

Organista.J, Cordero.G. 2001. *Indicadores cibermétricos para el caso de una revista electrónica de investigación educativa.* [En línea] Diciembre de 2001. [Citado el: 16 de Diciembre de 2007.] http://www.dgbiblio.unam.mx/servicios/dgb/publicdgb/bole/fulltext/volIV22001/pgs_67-76.pdf.

Rouber.A, Witvoet.O. 2002. *Uncovering Information Hidden in Web Archives.* [En línea] Diciembre de 2002. [Citado el: 22 de Enero de 2008.] http://www.dlib.org/dlib/december02/rauber/12rauber.html.

Tolosa.G, Bordignon.A. *Caracterización del Espacio Web de Perú.* [En línea] [Citado el: 20 de Enero de 2008.] http://eprints.rclis.org/archive/00007703/01/webpe.pdf.

Tolosa.G, Bordignon.F, Baeza.R, Castillo.C. 2006. *Caracterización del Espacio Web de Argentina.* [En línea] 2006. [Citado el: 21 de Enero de 2008.] http://www.chato.cl/papers/tolosa_2007_web_argentina.pdf.

Anexos

1. Nombre del entrevistado.
2. Cargo que ocupa.
3. Sabiéndose que la Web de forma general está en constante cambio. ¿Cuántos sitios están publicados en los servidores de la universidad?
4. ¿Cuáles usted cree que son los que más enlaces internos y externos presentan?
5. ¿Cuáles presentan mayor información?

Anexo # 1. Guión de la entrevista.

Nombre de las extenciones	Link encontrados	Por ciento
gif	4 876	34.94%
php	3 394	24.32%
png	2 307	16.53%
asp	1 139	8.16%
pdf	726	5.20%
jpg	471	3.37%
html	283	2.03%
htm	134	0.96%
css	129	0.92%
wmv	126	0.90%
swf	123	0.88%
ico	75	0.54%
txt	47	0.34%
x	42	0.30%
mp3	26	0.19%
net	9	0.06%
doc	8	0.06%
mspx	6	0.04%
shtml	4	0.03%
aspx	3	0.02%
cgi	2	0.01%
deb	2	0.01%
net de microsoft	2	0.01%
odg	2	0.01%
odt	2	0.01%
phtml	2	0.01%
asmx	1	0.01%
bz2	1	0.01%
c	1	0.01%
cfm	1	0.01%
iso	1	0.01%
jpeg	1	0.01%
jsp	1	0.01%
loc	1	0.01%
php3	1	0.01%
pl	1	0.01%
ppt	1	0.01%
python	1	0.01%
remote	1	0.01%
wma	1	0.01%
xpi	1	0.01%
zip	1	0.01%

Anexo # 2. Total de documentos que no son HTML.

GLOSARIO DE TÉRMINOS

Acceso directo: Archivo con el cual se puede acceder de forma rápida a un programa o a un fichero.

ActiveX-R: Un conjunto de tecnologías que permiten a los componentes de software interactuar entre sí en un entorno de red independientemente del lenguaje en que los componentes fueron creadas.

Applets: Es otra manera de incluir código a ejecutar en los clientes que visualizan una página web. Se trata de pequeños programas hechos en Java, que se transfieren con las páginas web y que el navegador ejecuta en el espacio de la página.

Archivo: Es la unidad básica de almacenamiento que habilita una computadora para distinguir un conjunto de información de otro.

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

Ciberespacio: Aunque ciberespacio se entiende de forma general como el conjunto de contenidos disponibles en formato electrónico, en la www.

Dominio: nombre base que agrupa a un conjunto de maquinas o dispositivos, este proporciona un nombre.

Driver: Un Driver, o controlador, es un programa que controla un dispositivo.

Etiquetas (tags): es una marca con tipo que delimita una región en los lenguajes basados en XML

Explorador: Una aplicación cliente que le permite a un usuario visualizar HTML en el www, otra red o computadoras del usuario.

Frames: Marco. Área rectangular en una página web que la separa de otra

GET: La Petición GET te devuelve el contenido del elemento pedido, ya sea una página Web, o un elemento multimedia, fichero.

HEAD: Y la head, te devuelve la Cabecera en la cual hay una serie de informaciones del elemento que le haces el pedido, tales como estado, tamaño, algunas configuraciones, e informaciones de utilidad.

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

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

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

Host: En redes locales basados en ordenadores, una computadora que proporciona accesos a otras.

Hostgraph: es un gráfico dirigido con un nodo correspondiente a un anfitrión y un borde dirigido ponderada correspondiente al número de enlaces entre un par de los ejércitos.

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

Indexar: Se usa en las aplicaciones de bases de datos para indicar la operación de ordenar los registros contenidos en ella de manera especial, en función de unos parámetros definidos previamente.

Indicador: Magnitud utilizada para medir o comparar los resultados efectivamente obtenidos, en la ejecución de un proyecto, programa o actividad. Resultado cuantitativo de comparar dos variables. Medida sustitutiva de información que permite calificar un

concepto abstracto. Se mide en porcentajes, tasas y razones para permitir comparaciones.

Infometría: Infometría e Ingeniería del Conocimiento: Exploración de Datos y Análisis de la Información en vista del Descubrimiento de Conocimientos.

Intrínseca: característica, esencial.

Java: Lenguaje de programación orientado a objetos.

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

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

Linux, Windows, UNIX: sistemas operativos.

Megabyte (MB): Una medida utilizada para el almacenamiento de datos. Representa un millón de bytes.

Mercadotecnia: Enfoque de administración de mercadotecnia que sostiene el logro de objetivos organizacionales depende de la determinación de las necesidades y deseos de los mercados objetivos y de la satisfacción de los mismos de manera más eficaz y eficiente que los competidores.

Protocolo: Un protocolo es un método establecido de intercambiar datos en Internet.

PostScript: es un formato de documentos, creado por la empresa Adobe, para describir documentos listos para imprimir.

Servidor: En informática, un servidor es un tipo de software que realiza ciertas tareas en nombre de los usuarios. El término servidor ahora también se utiliza para referirse al ordenador físico en el cual funciona ese software, una máquina cuyo propósito es proveer datos de modo que otras máquinas puedan utilizar esos datos.

Script: Concepto del término script: Grupo de lenguajes de programación que son típicamente interpretados y pueden ser tipeados directamente desde el teclado.

Terabyte (TB): Una medida utilizada para el almacenamiento de datos de alta capacidad. Representa un trillón de bytes.

Topología: La topología hace referencia a la forma de una red.

Usuario: En informática, un usuario es un individuo que utiliza una computadora, sistema operativo, servicio o cualquier dispositivo de interacción.

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

Webespacio: Directorio de páginas web sobre medios de comunicación.