UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DISCA

Máster en Ingeniería de Computadores y Redes

Trabajo Fin de Máster

Estudio y caracterización del espacio web de Bolivia

Autor: OMAR EDUARDO RÍOS ESCALIER

Directores: ANA PONT SAN JUAN
JOSÉ ANTONIO GIL SALINAS

Valencia - España
2015-2016

Índice de Contenido

Índice de Gráficos

Índice de Tablas

Índice de Figuras

Índice de Ecuaciones

Resumen

El presente trabajo de investigación muestra las características del espacio web boliviano a partir de la toma de muestras realizadas entre el mes de junio y julio del 2016, un poco más de 604215 Mil páginas web fueron extraídas de un poco más de 1200 Sitios. Esta muestra nos sirvió para realizar un estudio y análisis de contenidos, de enlaces y de tecnologías que son utilizados para la construcción de los sitios.

En la actualidad los sitios web se caracterizan por ser repositorios importantes, los cuales nos permiten compartir información, estos son utilizados para diferentes aspectos como pueden ser: el comercio, publicidad, educación, entretenimiento, redes sociales, multimedia entre otros. Todos los contenidos señalados anteriormente se encuentran en constante crecimiento, el estudio de las características, tendencias de crecimiento y evolución en el presente trabajo, nos permitiría obtener una valiosa información sobre la estructura de las páginas, tecnologías utilizadas y algunos aspectos de los tipos de contenidos de la web.

Abstract

This research shows the characteristic of the Bolivian web space from sampling conducted between May and June 2016, approximately how 604215 thousand of web pages were drawn from a little over 1200 sites. This sample served us for a study and analysis of contents, links and technologies that are used for construction sites.

Today websites are characterized by important repositories, which allow us to share information, they are used for different aspects such as: trade, advertising, education, entertainment, social networking, multimedia and others. All contents listed above are in constant growth, the study of the characteristics, growth trends and developments in this work, allow us to gain valuable information about the structure of the pages, technologies used and some aspects of the content types the web.

1. Introducción

En la actualidad la World Wide Web es considerado como un espacio público el cual es utilizado por una diversidad de personas o usuarios, los cuales persiguen una variedad de objetivos. Al ser un repositorio distribuido[1], este nos permite compartir información de tal menara que nos facilita la publicación en los ambientes de trabajos, tales como el comercio, publicidad, contactos sociales, entretenimiento, educación y otros. Si bien somos conscientes de que la web es dinámica en cuanto a su constante evolución, el estudio de características y tendencias nos entrega una valiosa información, tanto para entendimiento de su estructura como para el desarrollo de herramientas que faciliten la utilización de sus recursos.

Quizá para muchos parezca que el estudio de las características del espacio de una web pueda ser bastante tediosa y compleja, ya que requiere el uso de recursos computacionales a gran escala debido a su tamaño virtualmente infinito y distribución geográfica de la red. Por esta razón muchas veces se ha optado por tomar muestras pequeñas para su estudio, específicamente los dominios nacionales, utilizando diferentes estrategias de recolección de datos. Las muestras tomadas representan un buen balance entre la diversidad y la complejidad, por tal sentido estos constituyen un conjunto de información de alto interés.

En este trabajo de fin de master se presenta un estudio de caracterización del espacio web de Bolivia, en el cual podemos encontrar algunas particularidades, las cuales se diferencian de otras. El presente trabajo estudia características principales que se reportan y se registran para ser utilizadas en futuros trabajos, éstos se comparan de igual manera con otros trabajos similares realizados de otros países. Hasta donde pudimos indagar, este sería el primer trabajo de estudio realizado sobre el espacio web de Bolivia.

1.1. Motivación

Lo que motivó a realizar el estudio de investigación de la web en Bolivia, surge de la necesidad de poder obtener un registro de información general de la estructura existente en dicha web, y al no encontrarse antecedentes algunos a la fecha sobre algún tipo de estudio similar al de este trabajo, nació el compromiso de poder aportar con la información recolectada, y comparar los datos con los de otros países vecinos de la región, o de otros continentes, quienes cuentan con registros (inclusive históricos) del crecimiento de cada una de sus web con el paso del tiempo.

[1] Se entiende por Repositorio Distribuido, al almacén de recursos digitales y sus metadatos a los que se puede acceder sin que sea necesario un conocimiento previo de la organización o estructura de dicho almacén, el cual nos permite la interacción con humanos o con otros sistemas de software.

1.2. Objetivo General.-

El objetivo general de este trabajo se podría redactar de la siguiente manera:

"Determinar un mapa de caracterización del espacio web, utilizando diferentes estrategias de recolección de datos particulares y de importancia, que nos permitan entender las nuevas tendencias de crecimiento y evolución del mismo".

1.3. Objetivos específicos.-

Para poder concretar el objetivo general se plantearon los siguientes objetivos específicos:

- Estudiar y analizar contenidos de enlaces y de tecnologías utilizadas para construir sitios.
- Caracterizar los sitios, páginas, tecnologías utilizadas y aspectos descriptivos de su topología.
- Construir un estado comparativo de resultados con otros espacios webs nacionales existentes

2. Características generales de la Web

La web de manera general puede ser modelada como un grafo dirigido *(webgraph)* donde los nodos corresponden a páginas HTML y los enlaces entre éstas son las aristas. Formalmente este grafo consiste en un conjunto de nodos, denotado como P y un conjunto A de aristas. Cada arista (denotada como q->p) es un par ordenado (q, p) que presenta un enlace o vínculo entre las paginas (nodos) q y p, situación que se da sólo con algunos pares. En este caso, q es un entrante de p y éste un saliente de q. (Tolosa G. , A. Bordignon, Baeza-Yates, & Castillo, 2007).

Se vio conveniente estudiar la topología del grafo web, donde una de sus características principales es la de formar una red libre de escala, es decir, que proporciones menores a éste, mantienen propiedades de un grafo completo. De igual manera otra de las características de la *red libre de escala* es que tienen una distribución dispareja de nodos y enlaces, esto significa que se pueden encontrarse nodos con muy pocos enlaces como también otros con muchos enlaces.

Algunos autores plantean que en la topología de grafo de la web corresponde a una red libre de escala, en el cual la distribución de los enlaces sigue una ley de potencia de la forma de la ecuación 1, la cual expresa la probabilidad de que la página X posea K enlaces.

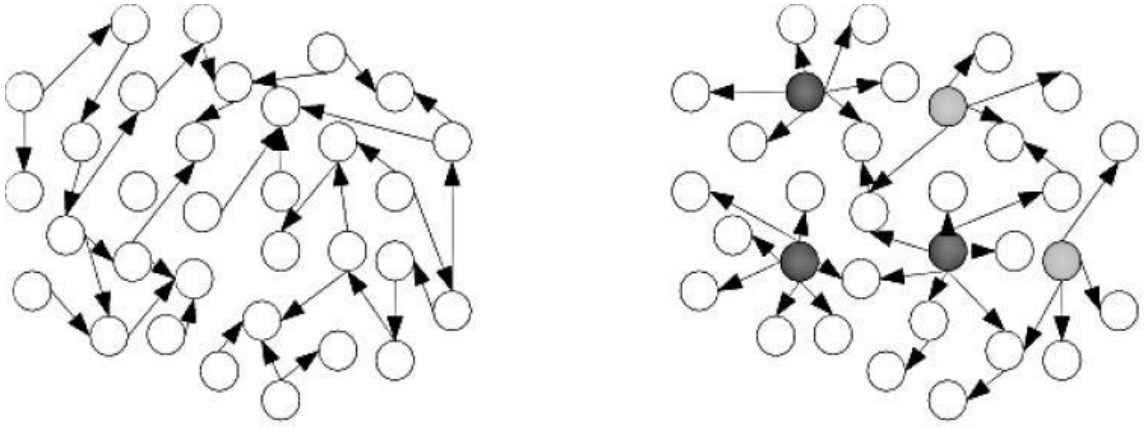
$$P(x = k) \approx k^{-\beta} \qquad \text{para } \beta > 0,$$

Ecuación 1: Probabilidad que la página X posea K enlaces *(Baeza-Yates & Graells, 2008)*.

El exponente β de la ley de potencias describe que tan rápido disminuye el valor de la frecuencia de x. Los ejemplos clásicos de estas distribuciones corresponden a Zipf y Pareto. Esta situación fue luego observada por Broder en un muestreo de la web de gran escala, encontrando como propiedad básica del grafo web que la distribución del grado entrante de los vértices sigue una ley de potencias con exponente β = 2.1. Por otro lado, la distribución del grado saliente sigue una ley de potencias imperfecta con β = 2.72. (Tolosa G. , A. Bordignon, Baeza-Yates, & Castillo, 2007)

2.1. ¿Cómo es la Web?

Muchas personas coinciden en que la Web no más que un simple conjunto de documentos almacenados en distintos servidores, en los cuales generalmente existe información de documentos mediante enlaces establecidos entre ellos. Esta relación existente entre enlaces presenta muchas ventajas para los usuarios, tanto como la búsqueda de información como para los programas que corren en la web al momento de ejecutar motores de búsqueda de contenido y recolección. Es así, que podemos decir e indicar que la Web sigue un modelo de grafo dirigido, donde cada página es un nodo y el arco representara un enlace entre dos páginas.



(a) Red aleatoria (b) Red libre de escala

Figura 1. Ejemplo ilustrativo de una red aleatoria y una red libre de escala de 32 nodos y 32 enlaces.[2]

Cuando una página enlaza a otra, por lo general es posible reconocer o enlazar a páginas más completas o mejores que las anteriores, siendo que pueden recibir un número mayor de enlaces que el normal o promedio.

Como bien ya habíamos mencionado la Web se puede clasificar como una *red libre de escala*, contraria a lo que son las *redes aleatorias,* las primeras se caracterizan por una distribución dispareja de enlaces, en la que los nodos altamente enlazados actúan como

[2] Fuente: Google images

centros que conectan muchos de los otros nodos a la red, tal y como se muestra en la figura 1. (Baeza-Yates & Graells, 2008).

2.1. Arquitectura de un sitio Web

Debemos de considerar tres componentes fundamentales para entender la arquitectura de la Web, estos componentes principales: Un servidor Web, una conexión de red, y uno o más clientes (navegadores).

En la figura 2 se observa una descripción grafica del servidor Web el cual distribuye páginas de información a los clientes o navegadores que los solicitan. Estos requerimientos son hechos a través de una conexión de red, para lo cual se utiliza el protocolo HTTP.

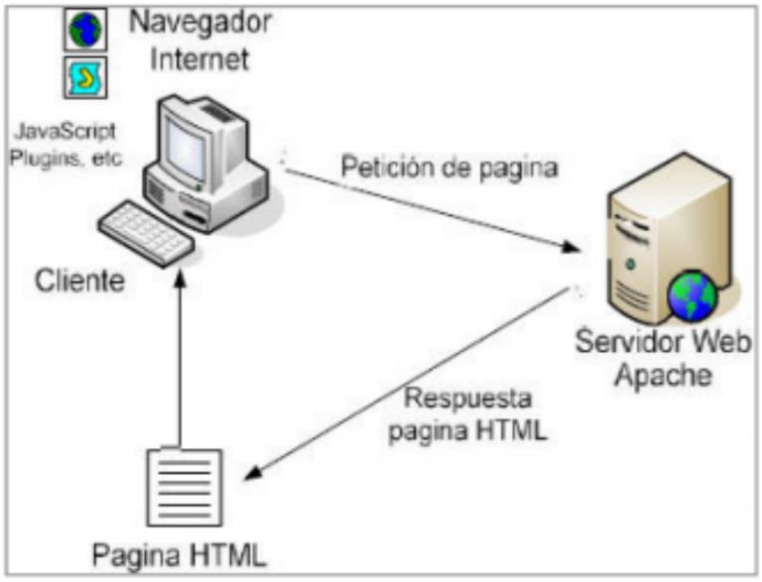


Figura 2. Esquema de la arquitectura básica de un sistema Web. *(Malvezzi, 2010)*

Las páginas estáticas están enfocadas principalmente a mostrar información permanente, donde el usuario se limita a obtener dicha información, sin que pueda interactuar con la página web visitada, por lo general se crean mediante el lenguaje HTML. Sin embargo, con frecuencia esta información esta almacenada en una base de datos y las páginas son creadas dinámicamente. Los sitios Web que usualmente manejan este tipo de esquema son comúnmente llamados sitios Web dinámicos, en la figura 3 se observa el esquema con contenidos almacenados en base de datos. Las páginas se generan dinámicamente en el momento que se hace la petición al servidor.

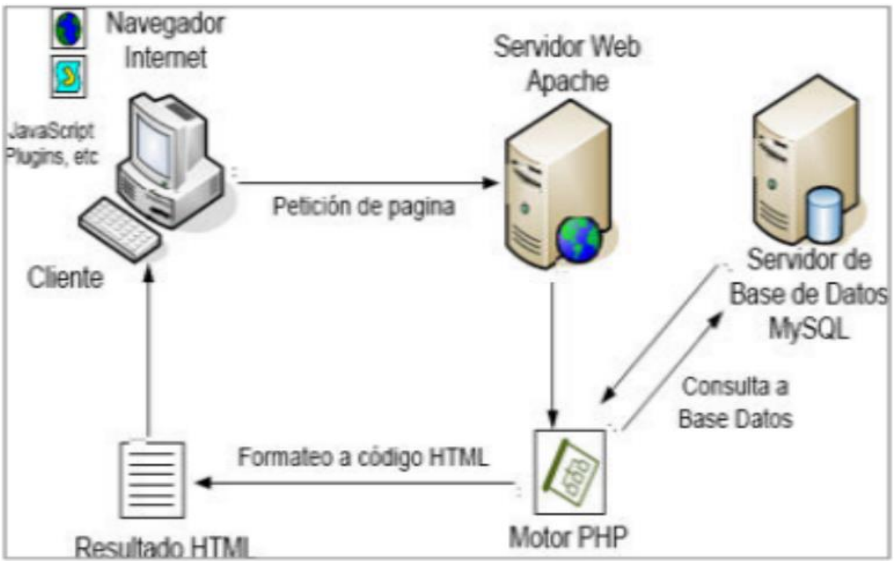


Figura 3. Esquema de la arquitectura de páginas que se generan dinámicamente. *(Malvezzi, 2010)*

En los últimos años las plataformas tecnológicas más estandarizadas sobre los cuales se desarrollan casi todos los sitios Web son los siguientes:

Arquitectura "Open Source":

- Servidor Web Apache
- Lenguajes de programación de aplicaciones: PHP
- Base de datos: MySQL
- Google Sites: Google Apps (sites.google.com)

Arquitectura Microsoft:

- Servidor Microsoft IIS
- Lenguaje de programación de aplicaciones: ASP, NET.
- Base de datos: Microsoft SQL/MySQL

Arquitectura Java:

- Servidor Web Tomcat
- Lenguaje de programación aplicaciones: JSP
- Bases de datos: soporta sistemas de varios fabricantes: Oracle, Microsoft SQL, etc.
- IBM Web Experience Factory

Las características y arquitectura del sistema dependerán de los requerimientos y alcances de cada proyecto y empresa que requiera solicitar un determinado tipo de servicio y configuración de hardware o software.

2.2. Elementos básicos de una página Web

Otra de las características importantes que debe de contener un sitio Web, son los considerados elementos de la página, pudiendo ser estos los contenidos de tipo textos, imágenes, videos, audios, etc.

Como se puede apreciar en la figura 4, los elementos o componentes de una página Web usualmente suelen ser imágenes, textos y otros contenidos multimedia las cuales están enlazadas de manera que el usuario puede navegar de una página a otra utilizando hipervínculos, el cual está asociado al concepto de interactividad surgido con el fenómeno de internet.

Tal y como se hizo referencia en los párrafos anteriores podemos resaltar como las principales características que constituyen una página de internet a los siguientes:

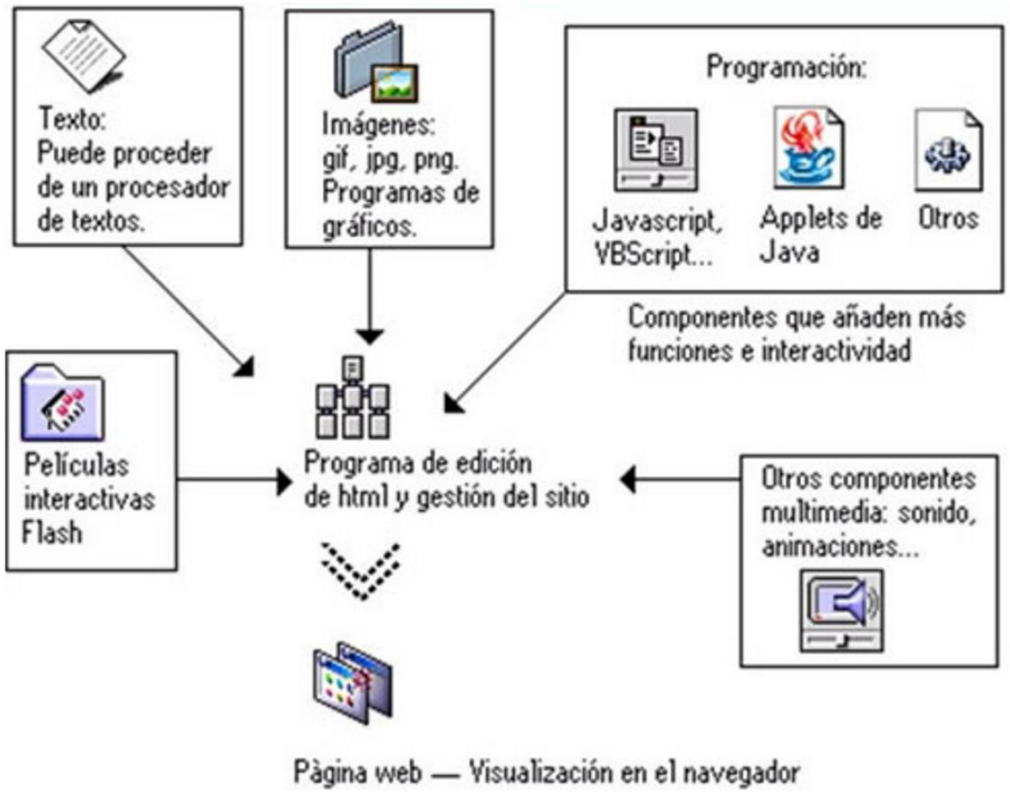


Figura 4. Esquema de elementos que contienen la Web. *(Delgado Rodriguez, 2015)*

El texto, es el elemento más significativo de cualquier sitio Web por que los usuarios navegan por la Web fundamentalmente en busca de información expresada en texto.

Imágenes, aunque hoy en día no puede faltar una imagen en una página, se recomienda no abusar de su uso, porque se corre el riesgo de aumentar el peso de las páginas, por

tanto, los tiempos de descarga. Las imágenes por lo general constituyen un elemento esencial para ofrecer una información visual del contenido además de mostrar un diseño atractivo y personal.

El espacio Web, para que nuestro sitio Web sea visible ante todos, necesitamos alojarlo en un servidor.

Un nombre de dominio, obtener el dominio correcto es un de las primeras cosas que se debe de considerar el cual identificara al servidor que almacena y sirve el sitio.

Hipervínculos, se podría decir que la magia del sitio comienza cuando relaciona varias páginas mediante enlaces internos y externos.

Video, mientras internet se hace más accesible, confiable y veloz, los programas de edición de video están más cerca a los usuarios, siendo más fácil realizar un video desde una cámara digital o un teléfono móvil con cámara incorporada y que este se termine subiendo a la Web para colocarlo al alcance de todos.

Animaciones de Flash, Una imagen animada, fácil de crear y sobre todo que ocupa poca memoria. Este puede tener diferentes usos, desde la creación de dibujos animados hasta el despliegue de dibujos interactivos.

Sonidos, la comercialización de nuevos dispositivos digitales móviles como iPhone ha potencializado este elemento, además de incorporar sonidos a las páginas Web se puede descargar de ellas archivos de audio para sus dispositivos móviles. El formato MP3 es el usado y conocido por su calidad y nivel de compresión.

2.3. Estudio de la web en un país

Las redes libres de escala son auto-similares es decir que: una pequeña muestra mantiene características de la red completa (las características trascienden la escala con que se mire la red). En nuestro estudio realizado se muestra que la Web Boliviana presenta características muy similares a la de la red mundial y redes de otros países, a pesar de haber considerado solo una muestra de la recolección de datos.

Técnicamente se hace difícil poder distinguir si un página está asociada al país que se está estudiando, en el presente estudio se trata de utilizar la mayor cantidad de sitios web con dominios .bo, y dominios públicos genéricos de primer grado conocidos (.com - .tv -. edu y otros), que a la hora de realizar el estudio se encuentran hospedadas en direcciones IP extranjeras a las asignadas a Bolivia.

Para poder completar el estudio del presente trabajo, se quiso indagar, si en años pasados la Web Boliviana fue objeto de algún estudio anterior, lamentablemente no se pudo encontrar registro sobre algún tipo de estudio de sus características. Asimismo, se pudo

evidenciar el estudio sobre otras webs nacionales tales como: (Baeza-Yates & Graells, 2008)

- África (9 países)
- Argentina
- Austria
- Brasil
- Chile
- China
- España
- Grecia
- Hungría
- Corea del sur
- Perú
- Polonia
- Portugal
- Reino Unido, Nueva Zelanda y Australia (sólo universidades)
- Tailandia

Se ha podido establecer y comprobar de igual manera que, si bien la Web crece a través del tiempo a una velocidad enorme, la estructura y sus propiedades se mantienen dentro de un rango de similitud considerable.

2.4. La recolección de Datos.

La recolección de la información fue realizada en el mes de mayo y junio de 2016, utilizando script en PHP y otros Bash Shell ejecutados en consola de la plataforma Linux, de igual manera se hizo un análisis del uso del crawler JWIRE v1.0 el cual es un entorno gráfico basado en el crawler WIRE creado por el departamento de investigación de la universidad de Chile.

Para la obtención de los datos se utilizó una máquina virtual con una memoria RAM de 4 GB, un sistema operativo Ubuntu de 64 bits y un HD 60 GB, procesador QEMU virtual CPU versión 2.0.0x4 de 4 núcleos funcionando en el clúster del laboratorio del Grupo de Informática Industrial de la Universidad Politécnica de Valencia del departamento de investigación de Arquitectura de la Web, y una consola piloto para la centralización de los registros encontrados, el cual contiene tanto una partición Linux como Windows, procesador Intel Core i7, 2.60 GHz., RAM de 8 GB y un sistema operativo de 64 bits.

El proceso del inicio de la recolección de datos fue de la siguiente manera:

En primera instancia se hizo la búsqueda de conjunto de rangos asignados de IPs a Bolivia, a través de diferentes servicios web de información en los cuales se tienen registrados los

rangos, en dicha lista se encontraron un total de 89 rangos iniciales y finales, inicialmente se lograron obtener un poco más de 1.141.028 IPs asignados a Bolivia, dicha recolección se realizó a través de un Script en PHP que se muestra en el Anexo M.

2.5. Características y dificultades de la Web

Una vez generado los números de IPs asignados entre los rangos, se pudo determinar la existencia de un poco más de 8.500 IPs en funcionamiento. De esta manera, teniendo constancia de la existencia de los sitios Web se procedió a realizar el análisis y estudio correspondiente del total y cada uno de ellos en función a los objetivos establecidos.

Se verificaron los resultados del total de IPs existentes, si estos contenían una dirección de dominio o URL para ser considerados en el estudio. En ocasiones la asignación de una IP no siempre direcciona a una página web, sino que también puede estar asociado a un acceso de servicio webmail, servidores o routers; posterior a eso se filtraron los datos utilizando una herramienta de hoja electrónica. La obtención de las direcciones se logró ejecutando el script que se muestra en el Anexo N el cual hace referencia a un recurso del sitio whoishosting.

Si bien la cantidad de los sitios Web encontrados están en menor proporción comparados con otros países, estos se encuentran dentro de lo esperado y registros que se tiene; según datos proporcionados por NIC Bolivia, la cual está bajo la administración de la Agencia para el Desarrollo de la Sociedad de la Información en Bolivia (ADSIB) dependiente de la Vicepresidencia del Estado Plurinacional de Bolivia el año 2013, el crecimiento de los sitios web en Bolivia fue de un 34 % en los últimos 3 años. El dominio ".com.bo" es el que cuenta con mayor demanda: había 5.764 sitios que llevaban esta extensión; seguido por el ".bo", con 2.198; el ".org.bo", con 512; el ".edu.bo", con 280; el ".gob.bo", con 237; el ".net.bo", con 82; el ".tv.bo", con 18; el ".mil.bo", con seis y el ".int.bo", con uno, respectivamente. (Razon, 2013)

De igual manera se consideraron para el presente estudio algunos sitios Web con información y contenido que hacen referencia al país, estos tienen dominios .com, .org, .tv, .info, .net, entre otros, quienes están registrados en servidores fuera del país, los cuales fueron obtenidos de portales Bolivianos que contienen registros de los mismos (mirabolivia.com, eju.tv), haciendo un total aproximado de unos 945 sitios obtenidos; 4 con ".tv", 239 con ".org"; 36 con ".net"; 11 con ".info"; 7 con ".edu"; 617 con ".com".

Los registros de sitios existentes con proveedores fuera del país, pueden ser bastante ventajosos como perjudicial, desde un punto de vista comercial, para muchas empresas ven como ventaja que lo dominios de primer nivel es más fácil de acordarse, también resulta más rápido su registro fuera del país que dentro, como desventaja principal podemos indicar que la burocracia existe, hace más difícil el pronto registro de un domino .bo; el hecho de que distintos autores determinen que tipo de contenido existirá en su

sitio Web sin ningún tipo instancia de control, también puede llegar a ser perjudicial. Todas estas aportan a las dificultades que se presentan para la búsqueda de información como para la caracterización de las mismas, entre las cuales además podemos mencionar a los parámetros *URL y URL Rewriting, réplicas de contenido y spam en general.*

Para complementar la obtención de la muestra, se recurrió de igual manera a sitios web de Alexa para la comprobación de los sitios web bolivianos más visitados.

En la tabla 1 se detalla la cantidad de dominios Web que se consideraron para tomar la muestra necesaria en el estudio del presente trabajo.

DOMINIO	CANTIDAD
.BO	103
.COM.BO	495
.COM	670
.EDU.BO	177
.EDU	3
.GOB.BO	240
.INFO	10
.INT.BO	1
.MIL.BO	3
.NET.BO	15
.NET	40
.ORG.BO	271
.ORG	244
.TV.BO	3
.TV	4
OTROS	32
TOTAL	**2311**

***Tabla 1**: Dominios de la muestra considerados para el estudio.*

Una vez identificado la cantidad de muestra de dominios para el presente trabajo, se procedió a bajar los contenidos de los sitios web para su posterior análisis y estudio, esto se logró utilizando el script que se detalla en el Anexo O.

2.6. Calculo de la muestra correcta

La obtención del cálculo del tamaño de la muestra es uno de los aspectos que se ha considerado definir para la fase previa de la investigación que determina el grado de credibilidad que consideremos a los resultados obtenidos.

Una de las fórmulas muy extendida que orienta sobre el cálculo del tamaño de la muestra para datos globales es la siguiente:

$$n = \frac{k^2 * p * q * N}{(e^2 * (N-1) + k^2 * p * q}$$

***Ecuación 2**: Cálculo de tamaño de la muestra conociendo el tamaño de los sitios Web.*

En donde:

N: es el tamaño de la población o universo

k: es una constante que depende del nivel de confianza que asignemos. El nivel de confianza indica la probabilidad de que los resultados de nuestra investigación sean ciertos: un 95.5 % de confianza es lo mismo que decir que nos podemos equivocar con una probabilidad del 4.5%.

e: es el error máximo admisible en término de proporción. El error es la diferencia que puede haber entre el resultado que obtenemos preguntado a una muestra de la población.

P: probabilidad de éxito, o proporción esperada. Este dato por lo general es desconocido y se suele suponer que p=q=0.5 que es la opción más segura.

Q: probabilidad de fracaso o de que le evento no ocurra, es decir, es 1-p.

n: es el tamaño necesario de la muestra (número de datos mínimos que se van a considerar).

Para poder obtener el tamaño de la muestra que se ha considerado para el presente estudio, se tomó como referencia el número total de sitios Web proporcionados por NIC Bolivia de acuerdo al artículo de prensa “La razón” en su sitio web de los años 2013, de tal manera que si el número de dominios Web existentes es remplazado en la ecuación 1 descrita anteriormente, además de los datos que se consideraron como aceptables para el estudio, obtendremos los siguiente.

Donde:

N = 9098 k = 2.58 = 99% e = 3% p = 0.5 q = 0.5

Por lo tanto la muestra obtenida será un valor de **n = 1537.**

2.7. Contenidos y Características generales del trabajo

Para la construcción de los diferentes contenidos que se analizan en el presente trabajo se contemplaron hasta tres niveles posibles de dominios nacionales de la Web, textos e imágenes que estos contengan, así como la cantidad de páginas y documentos de la web

global. El contenido del trabajo está basado en el estudio de la Web nacional de Bolivia detallados en los siguientes capítulos.

Capítulo 3, Sobre los Documentos: en esta sección lo que se presentara es el nivel de páginas y documentos, se estudiaran tanto las páginas web como los documentos en sus diferentes formatos a HTML de la Web Boliviana.

Capítulo 4, sobre los sitios: esta sección nos permite estudiar y analizar la Web Boliviana a un nivel de los sitios, se enumera los sitios con más documentos, con mayor tamaño y con mayor cantidad de enlaces.

Capítulo 5, sobre los dominios: el estudio en esta sección de la Web boliviana se la realiza a nivel de dominios, la relación que existe entre dominios con más sitios y con mayor tamaño así como los dominios extranjeros en términos de enlace.

Capítulo 6, sus conclusiones y trabajos futuros: finalmente se presentaran las conclusiones del presente trabajo, además de los trabajos futuros que podrían continuar al presente trabajo.

3. Sobre los Documentos

3.1. Paginas descargadas versus enlaces inválidos

El script evaluado para este objetivo extrae y recolecta información de las páginas que han sido descargadas en función al nombre de dominio guardados en un fichero específico (código 3), al realizar dicha descarga se presume que algunas direcciones que se encuentran en la lista del fichero no existen o posiblemente estén mal escritas. Al momento de solicitar el servicio de conexión al servidor, este retorna el código de estado que indica si la página existe o no, solo las páginas que tienen al menos un contenido web son recolectados caso contrario se omitirán.

En la gráfica 1 y tabla 2 se muestra la cantidad de páginas recolectadas según la muestra establecida.

TOTAL DE WEB CONSIDERADAS	WEB PASIVAS	WEB ACTIVAS
1725	483	1242

Tabla 2*: Cantidad de sitos Web activos y pasivos considerados para el caso de estudio.*

La cantidad de enlaces a páginas que no fueron recolectadas, se puedes asumir a diferentes hipótesis. Una de ellas es la migración de contenidos de sitios, muchas de los

sitios que tienen un contenido extenso o estructura compleja sean administradas mediante los denominados *Administradores de contenidos.* Este tipo de software permite que los administradores de un determinado sitio manejen su contenido sin tener que involucrarse en el desarrollo del mismo (programación). Es así que cuando un sitio migra sus contenidos desde una estructura antigua a una nueva, esa migración tiene consecuencia a favor y en contra, será favorable para el sitio desde un punto de vista administrativo pero desfavorable en términos de los enlaces que el sitio recibía.

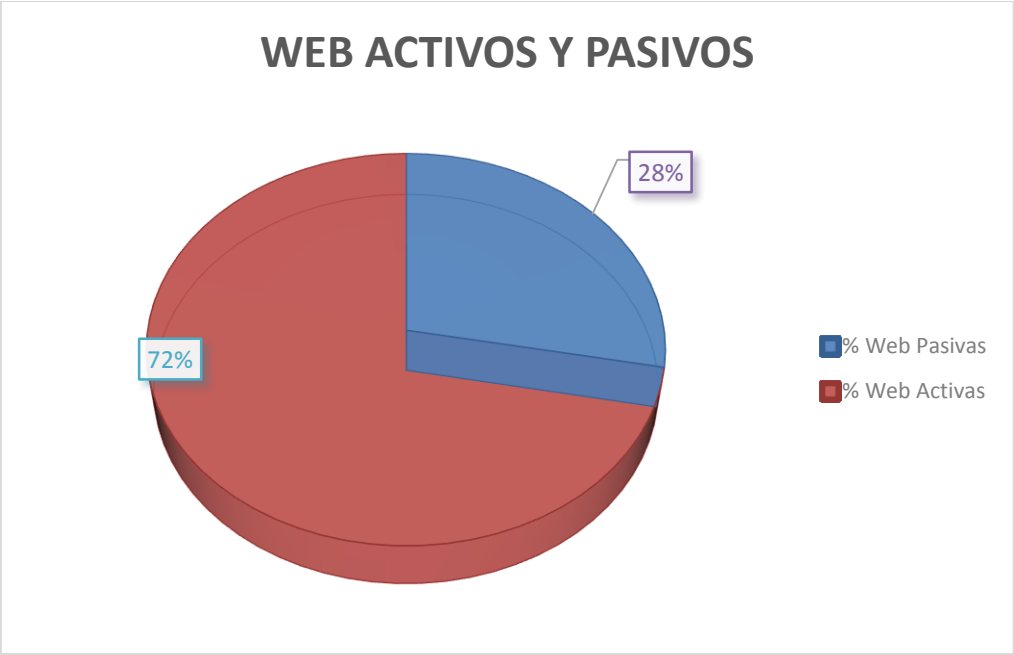


***Grafico 1**: Representación porcentual de sitos Web activos y pasivos.*

3.2. Longitud y profundidad de las URLs

Es bien sabido que la dirección de una página Web se expresa comúnmente mediante una URL (Uniform Resource Locator). Esta sigla y definición tiene un doble propósito, es decir, por un lado identifica un recurso en la Web de manera única, y por otra parte indica cómo es posible acceder a dicho recurso en el servidor.

Por lo general las URLs más usadas en la Web son las que corresponden al protocolo de transferencia de hipertextos (HTTP), éstas usualmente tienen la forma siguiente:

http://sitio/diretorio/subdirectorio/documento

Podemos indicar el siguiente ejemplo, http://www.usfx.info/edif/index.php?id=40, el cual indica que el sitio a contactar es www.usf.info, el documento a que se necesita esta en /edif/ y se llama index.php?id=40, este último nombre dependerá de la configuración del

servidor que se esté utilizando y de la tecnología usada para generar paginas dinámicas en este caso index.php, si no existiese buscara un archivo index.html.

3.2.1. Longitud de las URLs

La longitud promedio de las URLs, incluyendo la especificación del protocolo http://, nombre de servidor, ruta y parámetros se estima que no pase los 57 caracteres de la Web global, la longitud de las URLs Bolivianas llega en promedio a 70, este promedio se puede comparar con los datos ya registrados según (Baeza-Yates & Graells, 2008) por otros países tal y como se muestra en el grafico 2: 74 para España, 69 para Portugal, 75 para argentina. Es posible que la diferencia con el promedio entre la red global esté relacionado con la existencia de las nuevas aplicaciones Web, tanto comerciales como sociales, las que incluyen en gran medida un tamaño considerable de parámetros en sus direcciones. Esto podemos constatar viendo las direcciones más largas que corresponden a las páginas dinámicas.

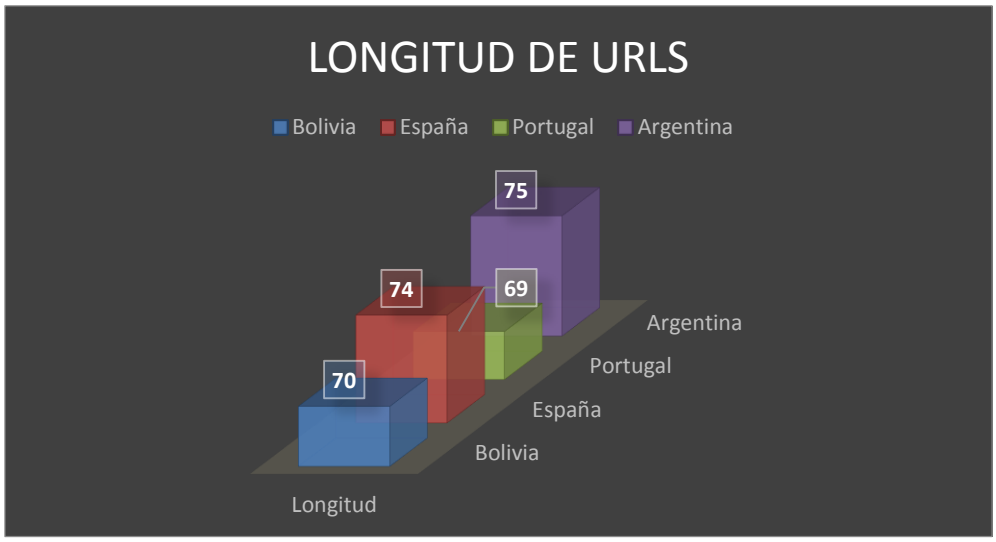


***Grafico 2**: Representación en número de caracteres de longitud de las URLs.*

De igual manera podemos indicar que la longitud más extensa encontrada en nuestro estudio llega a 99 caracteres y el menor a 21 caracteres para ello se hizo uso del script detallado en el Anexo B.

3.2.2. Profundidad de las URLs

Para poder definir la profundidad dentro de un sitio Web, se lo puede realizar de dos maneras:

Profundidad Lógica, generalmente la página inicial de un sitio Web está a una profundidad 1, todas las páginas que son directamente alcanzables desde ella están a una profundidad 2, y así sucesivamente. Entendiendo esta definición podemos señalar que la profundidad lógica mide el número de clics necesarios desde portada de un sitio hasta la página requerida.

Profundidad Física, de igual manera podemos indicar que la pagina inicial de un sitio está a profundidad 1, y las páginas de la forma http://sitio/pag.php o http://sitio/dir/ están a una profundidad 2, y así sucesivamente. En resumen podemos indicar que la profundidad física mide la organización en archivos y directorios de un sitio Web.

Con el uso del script detallado en el Anexo C pudimos obtener los siguientes resultados para la medición de la profundidad física de las web de estudio, donde la profundidad máxima encontrada fue de 9 y la minina de 2, el promedio encontrado fue de 5. En la gráfica 3 se refleja los porcentajes de profundidad, desde la mínima hasta la máxima encontrada en nuestros sitos Web de estudio.

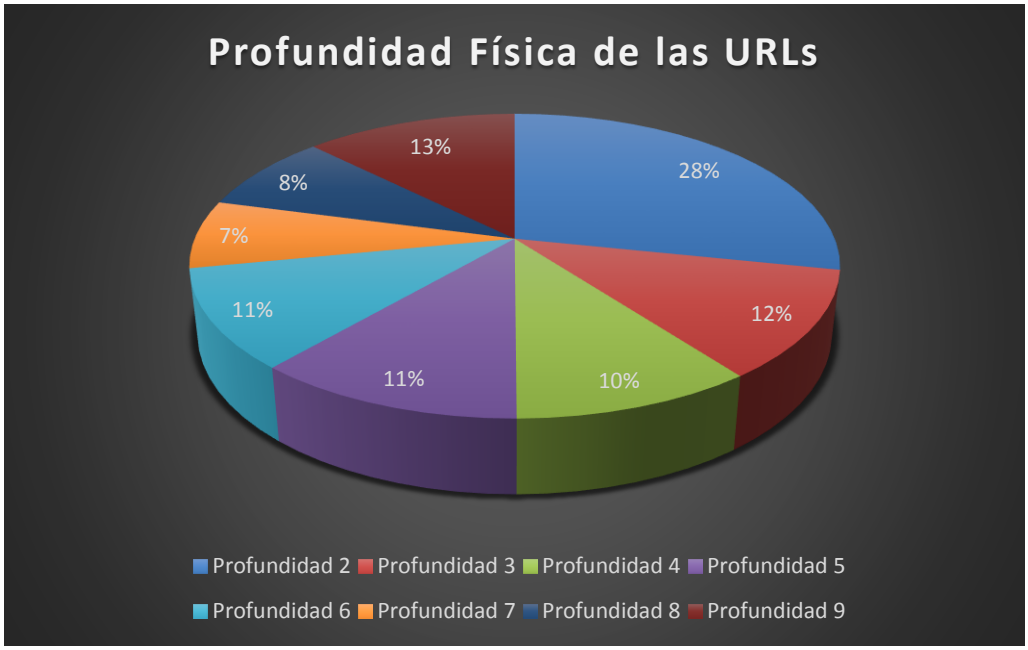


Grafico 3: *Representación gráfica de profundidad en las URLs.*

3.3. Edad de las páginas

La edad de las páginas podemos determinar a partir de la fecha de registro de la URL especificada por el servidor que la contiene. Muchas de las fechas suelen ser erróneas por ser dependientes de la configuración del servidor, que pueden ser fechas demasiado antiguas o fechas del futuro previas a la construcción de la Web.

En nuestro caso la distribución de las edades de las web estará dada en términos de meses para poder agruparlos y ser representados en el gráfico 4. De igual manera se ha podido observar que el dominio más antiguo registrado de la muestra es www.bancosol.com.bo con un valor de 240 meses de antigüedad a la fecha, y el dominio más reciente fue registrado por el www.anapqui.org.bo con solo 1 mes de antigüedad a la fecha de estudio

(fecha de análisis 12 de julio de 2016), esto se logró usando el script que se detalla en el Anexo D.

***Gráfico 4**: Antigüedad de las Webs representada en meses.*

3.4. Título de páginas

Los títulos de la Web Boliviana de acuerdo con lo observado se pude considerar de un largo no adecuado, si tomamos en consideración la referencia de otros países y comparamos con los largos de estos. En España la mayoría de los largos tiene entre 5 y 10 caracteres, en nuestro caso el promedio del largo de título de páginas de la Web en estudio llegan a 29. Este tipo de mediciones es importante porque el título de una página es uno de los atributos más importantes tanto en su usabilidad (es el identificador dentro de los bookmarks de usuario), como en su difusión (lo que más se destaca de una página en un listado de resultados de búsqueda).

3.5. Paginas dinámicas

Para considerar las páginas dinámicas se tomaron en cuenta aquellos sitios las cuales realizan una consulta a una base de datos, y esto involucra al proceso de despliegue de las páginas.

Las páginas Web dinámicas son aquellas en las que la información presentada se genera a partir de una petición del usuario de la página, contrariamente a lo que ocurre con las

páginas estáticas, en las que su contenido se encuentra predeterminado, por lo general suelen estar desarrolladas en XHTML, HTML y CSS.

Los lenguajes utilizados para la generación de los tipos de páginas dinámicas son principalmente Perl CGI, PHP, JSP y ASP, los cuales se consideraron en la ejecución del script.

Por lo general cuando existe una gran cantidad de páginas dinámicas estas no se pueden identificar directamente. Para lograr este objetivo se realizó el script del Anexo E, que determina la detección de la URL de la página web, a través de la detección del archivo, esto nos indica la tecnología con la que se generó a través del signo ?, al igual que los parámetros que recibe la página.

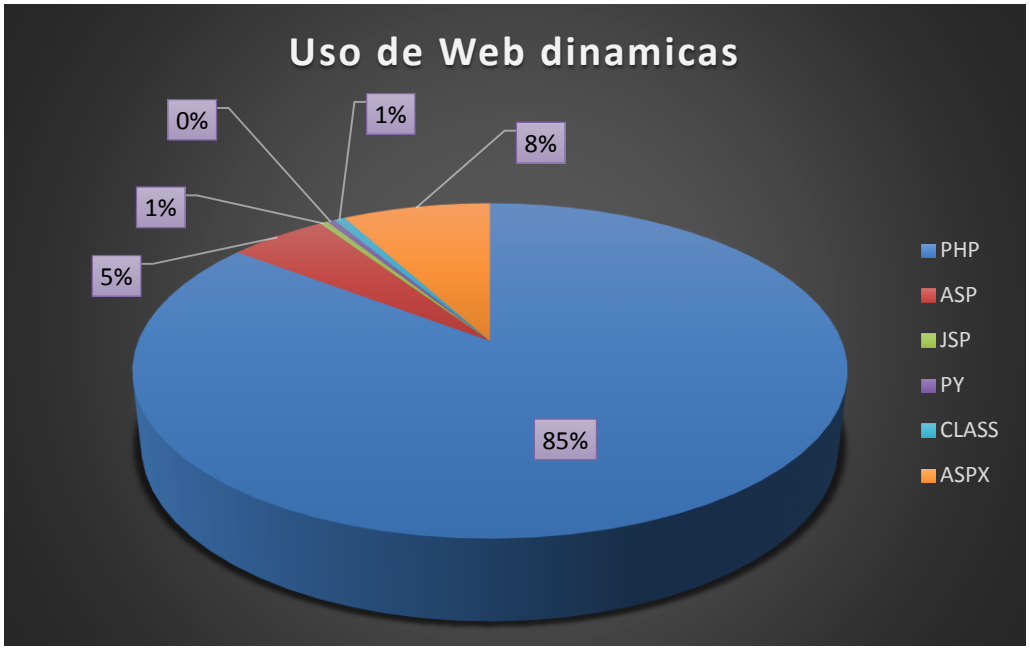


***Gráfico 5**: ficheros Webs dinámicas más usados.*

Sin embargo técnicas como URL rewriting impiden la determinación directa del tipo de página. Esta técnica por un lado, es beneficiosa para los usuarios, y que genera URLs más "amigables", mientras que por otro, usualmente esta técnica genera demasiadas páginas duplicadas que tienen el mismo contenido pero distintas URLs de acceso, por lo que cifras como la cantidad de páginas que tiene un sitio, así como su tamaño, se ven distorsionadas. (Baeza-Yates & Graells, 2008).

En la gráfico 5 vemos que PHP tienen una fuerte presencia y es el formato más usado para las consultas dinámicas seguido de ASPX y ASP, existe una leve presencia de los ficheros de tipo CLASS, JSP, y PY respectivamente.

3.6. Documentos distintos de HTML

Fueron encontrados aproximadamente 60461 enlaces a documentos de textos en formatos distintos a HTML de la muestra de estudio, siendo los más populares PDF y DOC con 57453 para el primero y 2100 para el segundo respectivamente, en el gráfico 6 se puede observar los resultados de los documentos de textos obtenidos con ayuda del script descrito en el Anexo F.

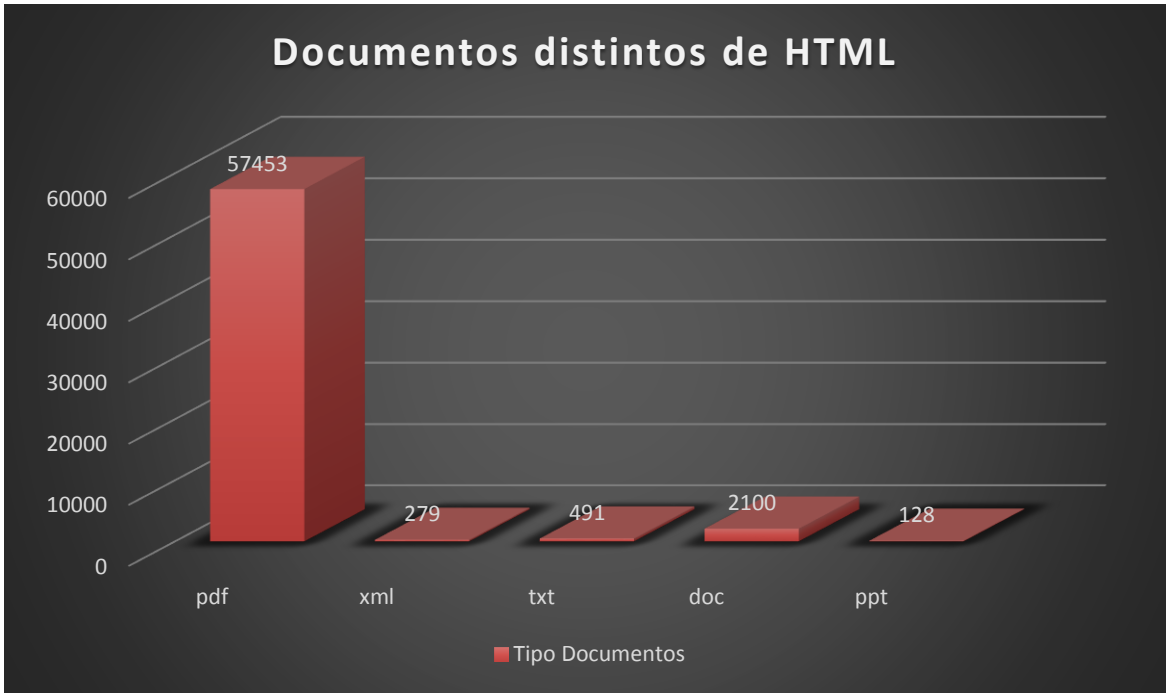


***Gráfico 6**: Documentos de texto distintos de HTML más usados.*

También podemos realizar un comparativo con datos obtenidos por (Baeza-Yates & Graells, 2008) con otros países como Austria, Brasil, Corea del Sur, Grecia, Portugal y Argentina donde los formatos PDF son los más populares.

3.6.1. Imagen, Audio y Vídeo

La tendencia al uso de enlaces de archivos multimedia fue más de lo esperado, con el uso del script descrito en el Anexo G se encontraron 4288882 archivos los cuales se resumen en el gráfico 7.

Las imágenes JPG vienen a ser las más populares en la Web de Bolivia con un 84% de los contenidos. La popularidad del formato JPG se puede deber a que estas son utilizadas como uno de los elementos gráficos principales en el diseño de las páginas, ya que es un formato de compresión con poco pérdida utilizado en contenidos, sean a través de botones u otros elementos gráficos en una página. Los segundos formatos más destacados encontramos son los PNG con 10% de participación, este tipo de formato PNG, que nació

inicialmente como una alternativa de remplazo al formato GIF, al parecer aún mantiene cierta popularidad de participación en los navegadores; y el formato GIF contiene un 6% de participación que a pesar de ser considerado de menor pérdida, generalmente es usado para almacenar fotografías, al parecer aún no tienen una participación significativa en la Web Boliviana. Los demás formatos existentes tienen una presencia mínima en cada uno de los sitios.

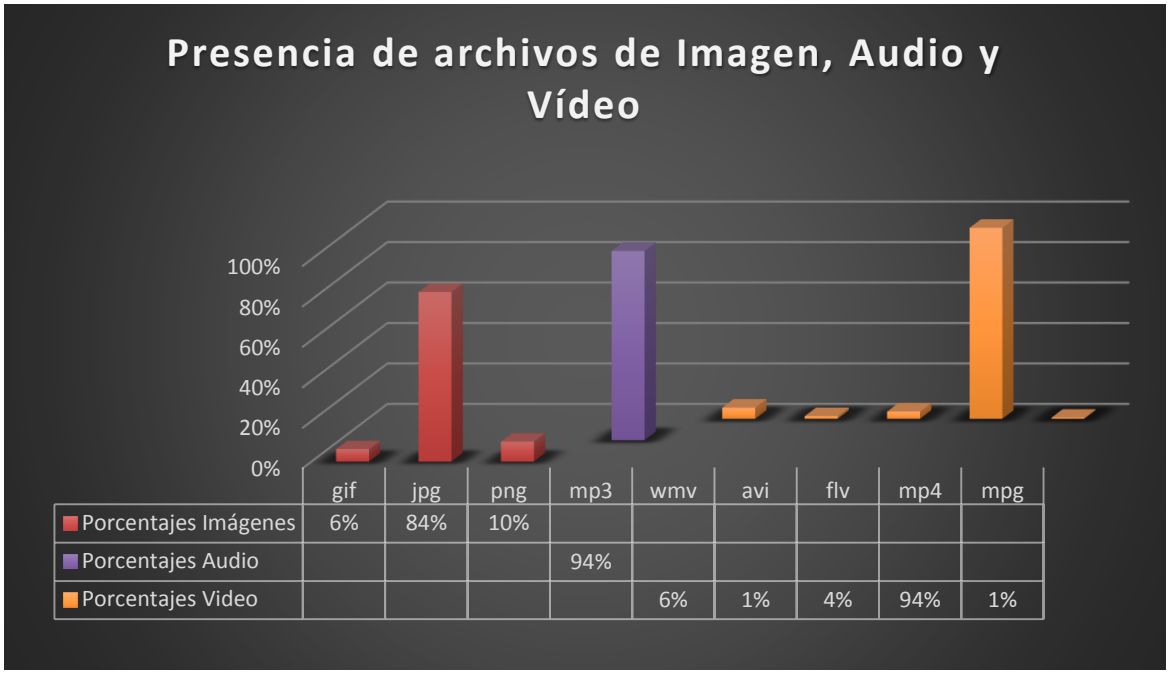


***Gráfico 7**: Relación de Archivos de Imagen. Audio y Vídeo en la Web de Bolivia.*

Entre los formatos de tipo audio, MP3 se mantuvo líder los últimos años desde la década 2010. MP3 llega a tener una participación de casi el 100% en la web boliviana, otros formatos como el WMA y PLS que fueron considerados en nuestro caso de estudio de investigación, no tuvieron valores destacados. Si bien PLS no es un formato de audio por sí mismo, más bien considerado como uno de tipo listas de reproducción fue también incluido para el análisis.

En cuanto se refiere a los formatos de tipo vídeo, se puede apreciar con mayor presencia al formato MP4 con 94% de participación, le sigue el formato WMV con 6% y FLV con 4%, el formato AVI logro una participación de 1% y MPG con 1% respectivamente. Si consideramos el formato con mayor presencia en la Web Global, probablemente encontremos que FLV sea el más utilizado debido a su uso en las redes sociales en la publicación de videos.

3.6.2. Archivos Comprimidos, Software y código fuente

Para el presente estudio fueron analizados alrededor de 9270 enlaces a archivos compresos, no más de 16 a archivos de programas, y por últimos 8190 de enlaces a archivos de código fuente en diversos lenguajes, para este objetivo se hizo algunas modificaciones al script del Anexo G los cuales se reflejan en el Anexo H y gráfico 8.

EL formato de archivos comprimidos que tiene un mayor dominio dentro del espacio Boliviano son: los formatos ZIP con 51% y GZ con 1%. Con frecuencia el formato ZIP es usado para distribuir software para Windows, al contrario del formato GZ que es usado para la distribución de software para Linux. Y por último los archivos RAR tienen una participación del 49%, estos son utilizados para la distribución de contenido arbitrario. Los formatos de tipo TAR (más que un formato de compresión es un contendedor que posteriormente se comprime en GZ) no tuvo registro de valor alguno en nuestro estudio. Otros tipos de formatos son considerados de presencia despreciable las cuales no fueron consideradas.

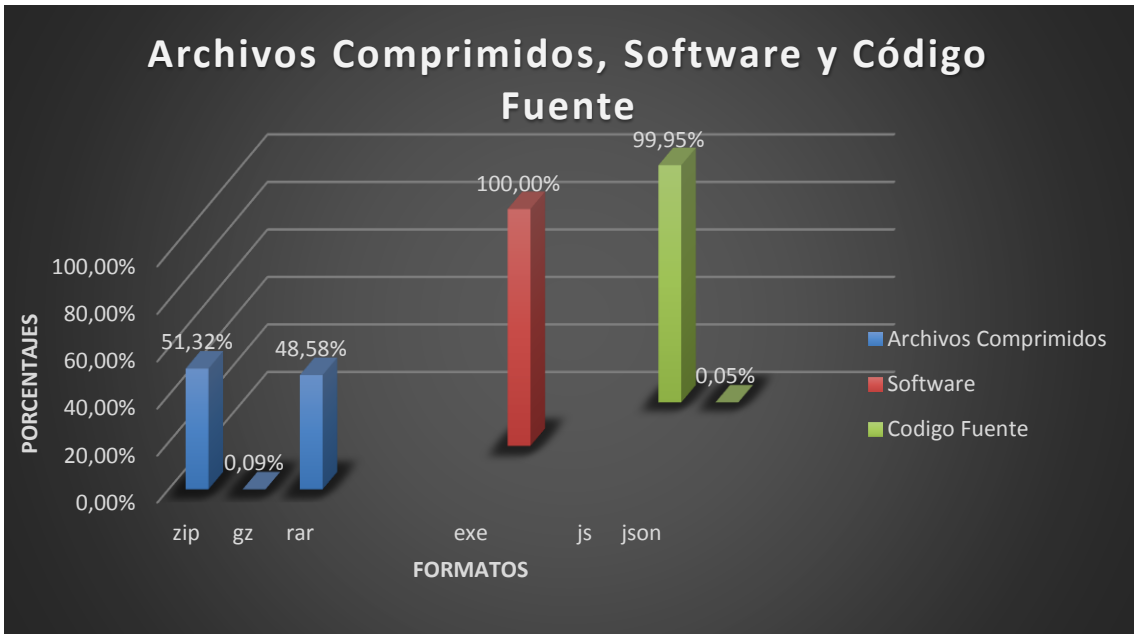


***Gráfico 8**: Relación de Archivos Comprimidos, software y código fuente en la Web Boliviana*

En lo que corresponde al software, no se consideraron los formatos comprimidos las cuales suelen distribuirse en gran cantidad. En nuestro caso los formatos que se tomaron en cuenta fueron el RPM, EXE y DEB respectivamente. Es así que el formato con una presencia mayoritaria por no decir única fueron los EXE con el 100% de participación y de los formatos RPM y DEB no se pudieron obtener registro alguno del grupo de muestra establecida para nuestro estudio, los restantes formatos son considerados como despreciables así que no fueron tomados en cuenta, de tal manera que podemos determinar que el software en formatos para Windows tiene la participación total de presencia que los de formatos para Linux, lo cual ratifica el hecho de uso de Windows por

parte de los usuarios, quizás se deba a que una aplicación de Windows se distribuye como un único ejecutable, mientras que las aplicaciones de Linux son distribuidas mediante una gran cantidad de paquetes.

En las distribuciones de código fuente vemos el gran crecimiento que tuvo y sigue teniendo Javascript con 99.95% de participación considerado como un lenguaje para la construcción de páginas web, las cuales reacciones dinámicamente ante las acciones de los usuarios, más que todo este es usado en sitios que utilizan AJAX para crear aplicaciones basadas en Web. Sin embargo observamos que la presencia como código para aplicaciones C y C++ no tiene prácticamente ninguna presencia, esto puede deberse a que un sitio ya utilizan muchos archivos JS con toda su funcionalidad, de igual manera se pudo observar que existe un cierto porcentaje muy bajo de archivos JSON con 0.05% de participación. De esta manera se vio que enlaces a código fuente C y C++ si bien en algún momento han ido en aumento notoriamente los últimos años en la Web Global, no tienen mucha trascendencia en la estructura web boliviana.

3.7. Enlaces entre páginas Web

Para poder determinar los enlaces entre páginas Web se consideraron los enlaces tanto de grado interno como de grado externo. Los de "*grado interno"* son los enlaces que tiene una página Web dentro de sí misma, y el número de enlaces que sale de una página se denomina "*grado externo"*.

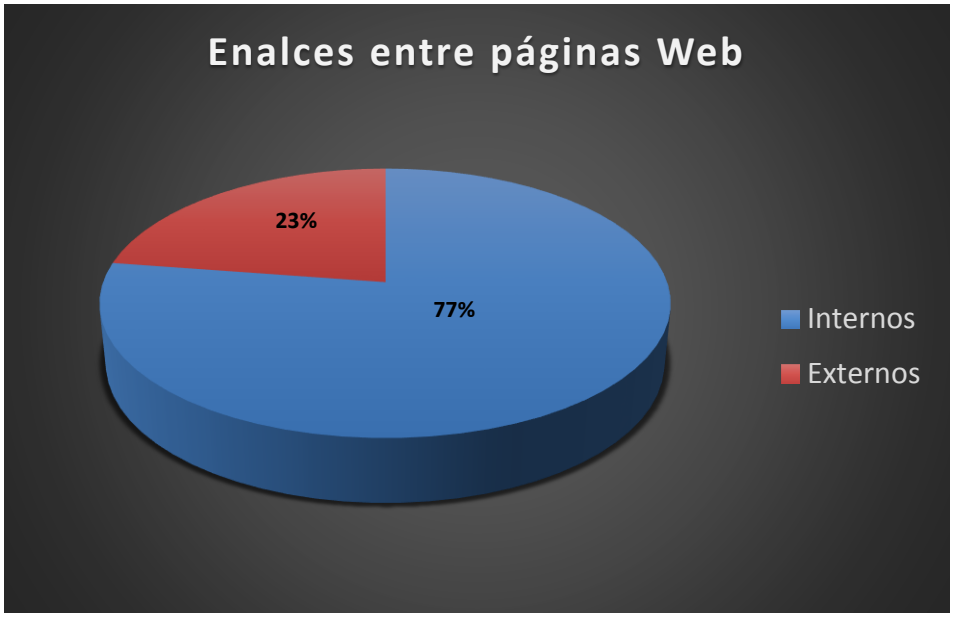


***Gráfico 9**: Relación porcentual de enlaces internos y externos de una página.*

Se hizo uso del script descrito en el anexo I, el script mencionado hace una búsqueda recursiva en todos los archivos de texto de la página web, buscando las etiquetas href y luego diferenciándolas por su tipo, es decir internas o externas, el gráfico 9 muestra la relación porcentual de los resultados.

4. Sobre los sitios

4.1. Número de Páginas

Se pudo observar un promedio de 101 enlaces de la muestra obtenida por sitio, a páginas que se consideraron desde su página de presentación principal, de igual manera se puedo analizar la distribución de páginas a través de los sitos es algo sesgada, ya que el 33% de los sitios tienen un 90% de los documentos, si ajustamos dicha distribución a una ley de potencias de parámetros de 1.29 podemos comparar con datos de (Baeza-Yates & Graells, 2008) España 1.14, Brasil 1.6, Argentina 1.45 y 1.84 de Chile.

4.2. Sitios que contiene solamente una página

Hay un total de 35 sitos que fueron recolectados de la muestra en los que se han encontrado sólo una página, los cuales representan 3% de todos los sitios estudiados, sin embargo si uno comienza a verificar los sitios manualmente, puede que encuentre entre ellos sitios completos con más de una página. El hecho de que solo se encuentre una sola página en un determinado sito se deba a los siguientes aspectos:

- El script no suele interpretar el lenguaje de programación de JavaScript, ya que la navegación de la página está basada en este lenguaje, es necesario interpretar el código JavaScript.
- El sitio sea sólo una redirección a otro sitio. La tecnología para realizar la redirección puede variar entre una etiqueta refresh en los meta-datos de la página, un enlace de redirección manual al que debe acceder el visitante, o una redirección mediante JavaScript. (Baeza-Yates & Graells, 2008) .
- El sito al que se desea acceder requiera el plung-in de Adobe Flash para que se pueda visualizar.
- El sitio al que se desea acceder utiliza Applets Java para su navegación.
- La página del sitio puede contener un contenido normal o en su caso presentar un número considerado de enlaces.
- El sitio al que se accede no es público, por tal sentido se requiere un acceso para su ingreso, es estos casos será normal encontrar una sola página.

SITIO	PAGINAS	COMENTARIO
WWW.FMBOLIVIA.COM.BO	189874	Medio de Comunicación radial
WWW.SICOES.COM.BO	77341	Medio de información gubernamental
WWW.ELDEBER.COM.BO	33097	Medio de Comunicación Escrita
WWW.CLUBSANJOSE.NET	27676	Catálogo de productos
WWW.UNITEL.TV	18985	Medio de Comunicación Visual
WWW.GNB.COM.BO	13881	Directorio de Sitos
WWW.AMARILLAS-BOLIVIA.COM	12152	Directorio de Sitios
WWW.BOLIVIAMALL.COM	11005	Catálogo de productos
WWW.FLORES.COM.BO	10971	Catálogo de productos
WWW.BOLIVIAEXTERIOR.COM	10382	Medio de comunicación Web
WWW.EVISOS.COM.BO	9279	Catálogo de productos y servicios
WWW.SOCIALES.COM.BO	7877	Medio de comunicación Escrita
WWW.BOLIVIAENTUSMANOS.COM	6520	Medio de Comunicación Web
WWW.FACETASDEPORTIVASTV.COM	5812	Medio de Comunicación Visual
WWW.EABOLIVIA.COM	4135	Medio de comunicación Web
WWW.HIDROCARBUROSBOLIVIA.COM	3627	Medio de Comunicación Web
WWW.UREAL.EDU.BO	3121	Educación y Capacitación
WWW.HANSAINDUSTRIA.COM.BO	3057	Catálogo de productos
WWW.EVISEX.COM.BO	3017	Catálogo de productos y servicios
WWW.BECASBOLIVIA.COM	2524	Comunidad de becas
WWW.UAJMS.EDU.BO	2437	Educación y Capacitación
WWW.RADIOPACHAMAMA.COM	2154	Medio de Comunicación, CMS con parámetros en URL
WWW.SANTOTOMAS.EDU.BO	2132	Educación y Capacitación
WWW.BIBLIOTECABICENTENARIO.GOB.BO	1432	Comunidad de Libros
WWW.SENASAG.GOB.BO	1420	Medio de Comunicación Gubernamental
WWW.FPS.GOB.BO	1238	Medio de Comunicación Gubernamental
WWW.FUPUAGRM.ORG.BO	1024	Comunidad de Estudiantes Universitarios
WWW.ENDEANDINA.BO	1018	Medio de Comunicación Web
WWW.FUNDACION-MILENIO.ORG	897	Medio de Comunicación Web
WWW.GOBERNACIONLAPAZ.GOB.BO	862	Medio de Comunicación Gubernamental

***Tabla 3**: Litado de sitos con más cantidad de documentos HTML, y posibles anomalías encontradas de los sitios web de estudio.*

4.3. Los sitios que contienen muchas páginas

En cuanto a los sitios que tienen muchas páginas, podemos indicar que la existencia de algunas tipo de anomalía; en la tabla 3 se muestra un listado de aproximadamente 30 sitos con más contenidos de páginas, en la cual también se muestra el total de páginas encontradas, la dirección del sitio y un comentario en la que se indica la característica de la web o tipo de anomalía que pudiera afectar al sitio en caso de ser identificado.

Una de las anomalías que se podría identificar es a consecuencia de utilizar un administrador de contenidos (CMS) para la gestión de las páginas, ya que estos administradores ofrecen una gran cantidad de formas para poder acceder al mismo contenido, generando una gran cantidad de páginas duplicadas, si bien todas son válidas pero con las direcciones distintas. Otra de las segundas anomalías se podría decir que se debe al uso de parámetros en la URL Rewriting, la cual acentúa la primera anomalía ya que se sitúa un documento en una estructura física que no existe.

Por otro lado, podemos interpretar que los sitios que llegan a tener una gran cantidad de páginas sean catálogos de productos, en especial en los sitios que se refieren a remates y sitios comunitarios, el cual un determinado usuario tiene asignado una gran cantidad de páginas, estas pueden ser: fotos, mensajes, publicaciones y otros, dependerá del tipo de sitio. Cuando se detecta un sitio que contenga catálogos de productos muchos de ellos duplican el contenido de otros sitios, lo que significa que cuando en el sitio se agrega un producto, otros sitios también lo agregan, por lo cual el número de páginas irá creciendo. También se pudo identificar que otro de los sitios que contienen mayor cantidad de páginas son los medios de comunicación, ya que estos guardan información de fechas pasadas las cuales están direccionadas a los enlaces de los artículos que correspondan al mes y año.

4.4. Los títulos de las páginas en un sitio

Uno de los script implementados (Anexo J) nos permite determinar la cantidad de títulos que contiene el sitio. Lo ideal en los sitios construidos es que cada página pueda tener un título en cada documento distinto, los resultados obtenidos que ser reflejan en el gráfico 10 podemos observar que más del 50% de las paginas contienen más de un título en la web respectiva, de igual manera se pudo determinar que los resultados con un valor de 0 son dominios que están fuera de servicio o puestos a la venta.

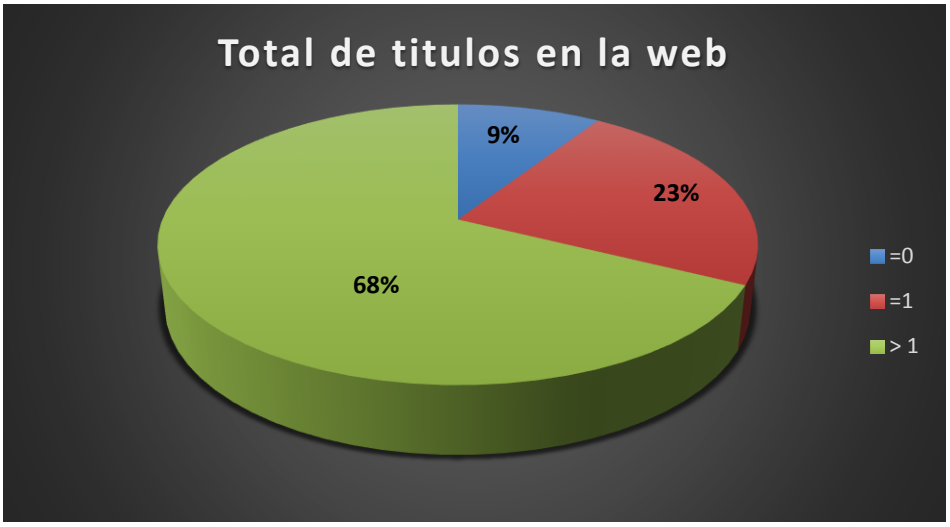


***Gráfico 10**: Cantidad de títulos que contienen los sitios Webs*

4.5. Tamaño de los sitios

Para determinar el tamaño de un sitio Web, se tendría que considerar la suma de todos los tamaños de las páginas que la componen. Para nuestro estudio se consideraron todos los contenidos existentes en el sitio web, los considerados textos de las páginas, incluyendo los contenidos de imágenes y otros documentos o archivos. El script utilizado en esta oportunidad se detalla en el Anexo K.

En el gráfico 11 se muestra la relación porcentual de tamaños de los sitios web de estudio encontrados, llegaron a un total de 207393.13 MiB[3], el gráfico muestra un rango de valores menores a 0 MiB, grupos de 10 MiB, 2 grupos de 100 a 1000 MiB y por último sitios mayores a 1000 MiB. Así mismo en tabla 4 se muestra las Urls que contienen tamaños mayores a los 1000 MiB.

4.6. Enlaces internos del sitio

Podemos entender por los enlaces internos de una página, a los sitios que apuntan sus enlaces a otra página dentro del mismo sitio. En promedio un sitio contiene 101 Enlaces internos. De igual manera podemos indicar que un promedio de los enlaces interno por página es 78, no se consideran los sitios vacíos o con una sola página, ya que no se encuentra la existencia de enlaces internos.

Por lo general los sitios que contienen pocas páginas usualmente no pueden tener demasiados enlaces internos.

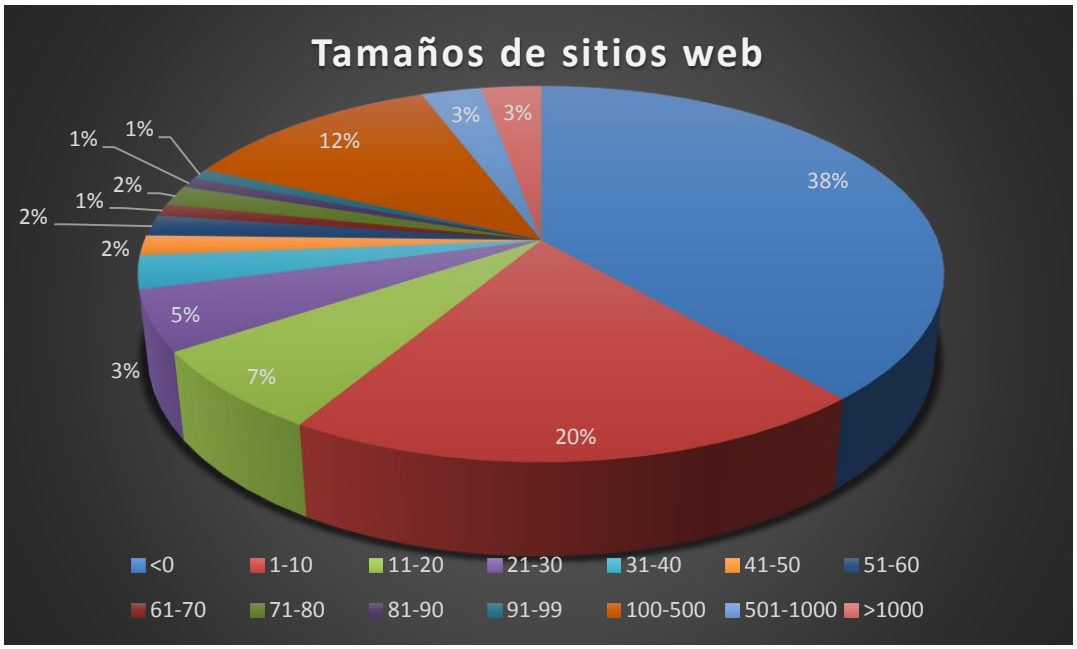


***Gráfico 11**: Relación porcentual de tamaño de sitios Web en MiB.*

[3] Sistema Internacional de Unidades con la palabra binario. Así pues, es denominado un mebibyte (MiB) contracción de megabyte binario.

DESCRIPCIÓN DOMINIO	MIB
WWW.COMUNICACION.GOB.BO	20480
WWW.ANH.GOB.BO	16384
WWW.FISCALIA.GOB.BO	9114
WWW.DEFENSORIA.GOB.BO	4813
WWW.FMBOLIVIA.COM.BO	4813
WWW.CLUBSANJOSE.NET	4301
WWW.FIUMSA.EDU.BO	4301
WWW.SICOES.COM.BO	4010
WWW.CRDASCZ.COM	3277
WWW.ELDIARIO.NET	3277
WWW.ELINTRANSIGENTE.COM	2867
WWW.COURSERA.ORG	2560
WWW.EVISOS.COM.BO	2459
WWW.ELDEBER.COM.BO	2458
WWW.SCIELO.ORG.BO	2458
WWW.GNB.COM.BO	2355
WWW.SANTACRUZ.GOB.BO	2253
WWW.CONCEJOMUNICIPALSCZ.GOB.BO	2150
WWW.ERBOL.COM.BO	2048
WWW.HANSAINDUSTRIA.COM.BO	1962
WWW.AMARILLAS-BOLIVIA.COM	1946
WWW.SENASAG.GOB.BO	1843
WWW.FACETASDEPORTIVASTV.COM	1515
WWW.ADUANA.GOB.BO	1434
WWW.PROMUEVE.GOB.BO	1434
WWW.APS.GOB.BO	1331
WWW.CAMBIO.BO	1331
WWW.ELDIA.COM.BO	1331
WWW.ABC.GOB.BO	1229
WWW.UAJMS.EDU.BO	1229
WWW.UREAL.EDU.BO	1126

***Tabla 4**: Litado Sitios Web que contienen objetos de tamaño mayores a 1000 MiB.*

4.7. Los sitios con más enlaces

En la tabla 5 se muestra la lista de los 30 sitios que contienen más enlaces, entre los primeros lugares podemos observar que se encuentran buscadores y directorios, aunque también podremos encontrar sitios de instituciones de educación y sitios comunitarios.

DOMINIOS	TOTAL ENLACES
WWW.CLUBSANJOSE.NET	153793763
WWW.FMBOLIVIA.COM.BO	68088142
WWW.EVISOS.COM.BO	19813736
WWW.GNB.COM.BO	17962342
WWW.SICOES.COM.BO	16989901
WWW.AMARILLAS-BOLIVIA.COM	15731409
WWW.ELDEBER.COM.BO	8004309
WWW.CEDLA.ORG	6676108
WWW.ELDIARIO.NET	5936952
WWW.EVISEX.COM.BO	5555916
WWW.BOLPRESS.COM	5034352
WWW.ERBOL.COM.BO	4734062
WWW.BOLIVIAENTUSMANOS.COM	4636859
WWW.BOLIVIAEXTERIOR.COM	4300546
WWW.BUENO.COM.BO	4060550
WWW.TIBO.BO	4041283
WWW.ADUANA.GOB.BO	2601566
WWW.USABOL.COM	2111899
WWW.ENBOLIVIA.COM	1983902
WWW.DMC.BO	1852623
WWW.HANSAINDUSTRIA.COM.BO	1807239
WWW.BOLIVIAEMPRESA.COM	1803297
WWW.SOCIALES.COM.BO	1511425
WWW.UNITEL.TV	1421750
WWW.AMARILLAS.BO	1339975
WWW.BOLIVIAMALL.COM	1324061
WWW.ELDIA.COM.BO	1218184
WWW.FACETASDEPORTIVASTV.COM	1148369
WWW.HIDROCARBUROSBOLIVIA.COM	1103932
WWW.AUDITORES.ORG.BO	1060411

***Tabla 5**: Sitios Web con más enlaces en sus páginas.*

5. Sobre los dominios

5.1. Direcciones IP y Software utilizado como servidor

Como bien se dijo al inicio del presente documento la búsqueda de la información para el estudio de los sitios Web se realizaron primeramente con la búsqueda de las direcciones IP asignadas al espacio Boliviano de las más de 8000, existen un aproximado de un poco más de 4000 IPs direccionadas de forma directa y las restantes contienen más de dos dominios distintos, aunque no todas ellas entregan información.

5.2. Tamaño de los dominios

Podemos observar que el tamaño de los dominios .bo de primer nivel en promedio es de 145 MiB, en la tabla 6 se muestra una lista de 30 dominios con más contenido, y en el gráfico 12 se encuentran los valores porcentuales que corresponden al tipo de sitio, si analizamos dicha lista podremos ver que una mayoría de los dominios hacen referencia sobre todo a sitos de tipo comercial, seguidos de los sitios administrados por el gobierno.

DOMINIO	MIB	TIPO
CAMBIO.BO	1331	M
TIBO.BO	809	C
FAB.BO	739	G
BOLIVIATV.BO	663	G
ENDE.BO	567	G
CTLP.BO	537	C
AUDI.BO	438	C
COCHABAMBA.BO	331	G
CNDC.BO	287	G
AMARILLAS.BO	276	C
TCPBOLIVIA.BO	263	G
CBN.BO	256	C
UMSA.BO	248	E
DMC.BO	222	C
CONCEJOMUNICIPAL.BO	147	G
BOLNET.BO	145	G
TIGOMUSIC.BO	126	C
TRABAJOPOLIS.BO	123	C
ABE.BO	62	G
ASOBAN.BO	61	C
PRODEM.BO	49	C
CONTRATA2.BO	44	C
DIPUTADOS.BO	44	G
CEPOS.BO	35	E
ENDEANDINA.BO	32	C
CNC.BO	30	C
COTEL.BO	23	C
DELAPAZ.BO	21	C
FRUTTE.BO	17	C
CLASIFICADOS.BO	10	C

***Tabla 6**: Listado de Dominios de primer nivel (C=comercio, G=gobierno, E=educación, M=medio de comunicación).*

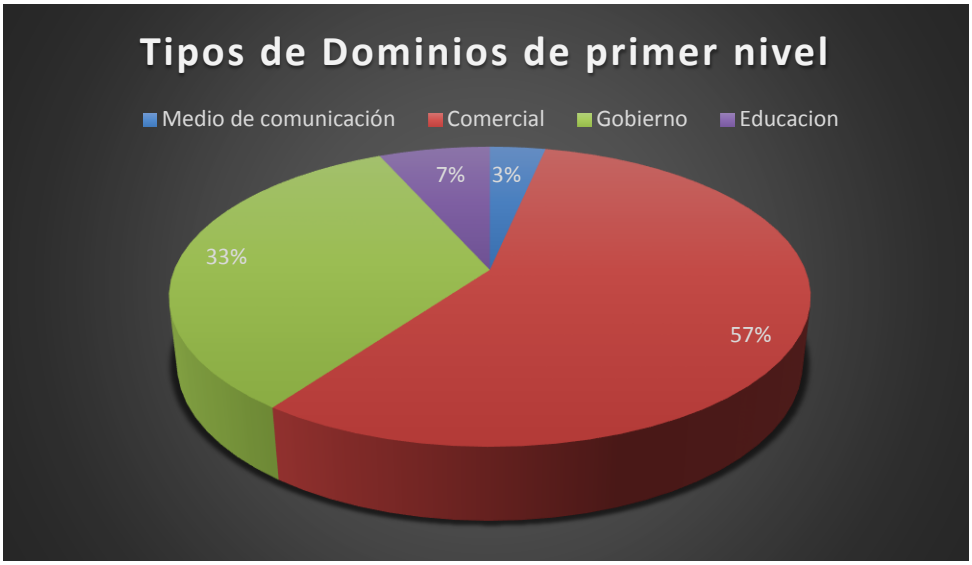


***Gráfico 12**: Muestra la relación porcentual de los dominios .bo más grandes en función de su tamaño en MiB.*

5.3. Los enlaces entre los dominios

En la tabla 7 se detalla los sitios que más enlaces tienen hacia otros dominios, de igual manera encontramos registrado en la lista dominios de sitios comerciales, gubernamentales, entidades educativas, medios de comunicación y otros. El gráfico 13 muestra una relación de los tipos de dominios externos más relevantes.

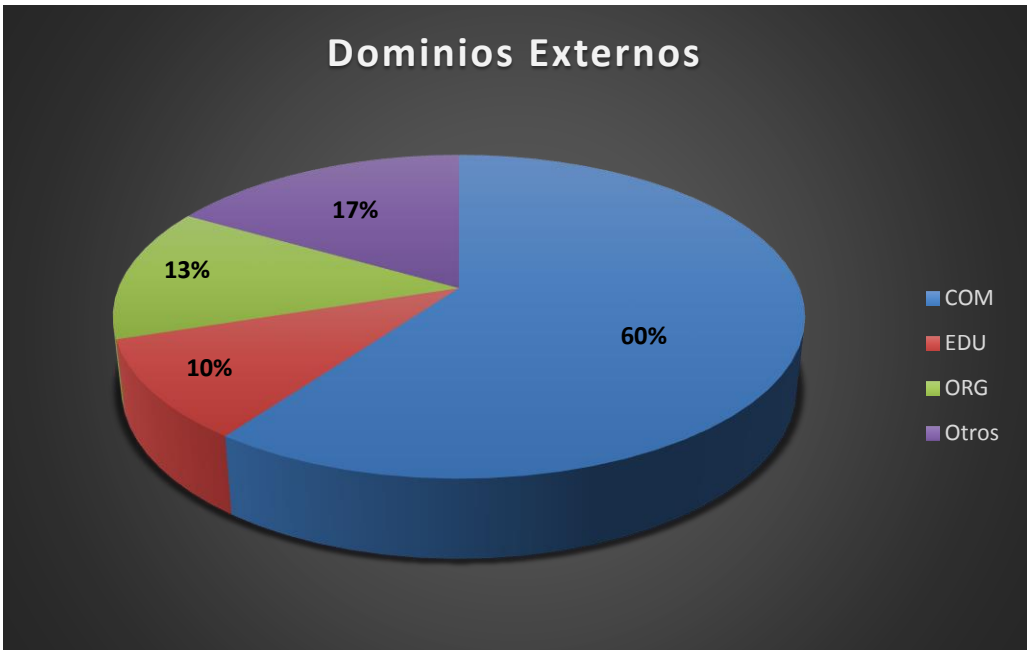


***Gráfico 13**: Relación de tipo de dominios con enlaces externos*

DOMINIOS	ENLACE EXTERNO
WWW.EVISOS.COM.BO	19287663
WWW.CNET.COM	12491551
WWW.FMBOLIVIA.COM.BO	9104952
WWW.EVISEX.COM.BO	6743004
WWW.CEDLA.ORG	6071069
WWW.BOLIVIAENTUSMANOS.COM	5270710
WWW.TIBO.BO	3853791
WWW.SICOES.COM.BO	2592199
WWW.DMC.BO	1936512
WWW.ELDEBER.COM.BO	1707084
WWW.BOLPRESS.COM	1678121
WWW.BOLIVIAEXTERIOR.COM	1565518
WWW.ERBOL.COM.BO	1416416
WWW.FACETASDEPORTIVASTV.COM	1316758
WWW.UNITEL.TV	1271036
WWW.BOLIVIAMALL.COM	1182926
WWW.GNB.COM.BO	1059195
WWW.AMARILLAS-BOLIVIA.COM	1034491
WWW.FLORES.COM.BO	824442
WWW.SCIELO.ORG.BO	801886
WWW.REVISTASBOLIVIANAS.ORG.BO	749039
WWW.CLUBSANJOSE.NET	745888
WWW.FRANCOBOLIVIEN.EDU.BO	664901
WWW.UAJMS.EDU.BO	591564
WWW.ELDIA.COM.BO	588057
WWW.AMARILLAS.BO	540865
WWW.SANTOTOMAS.EDU.BO	530433
WWW.AUDITORES.ORG.BO	472834
WWW.QUEBARATO.COM.BO	462646
WWW.ENBOLIVIA.COM	462646

***Tabla 7**: Listado de Sitios Web con más enlaces externos a otros sitios*

5.4. Dominios de primer nivel

De igual manera se pudo identificar aquellos sitios con dominios de primer nivel que tienen contenido Boliviano pero que no necesariamente están hospedados en IPs nacionales (gráfico 14). Es posible que se tengan otros sitios con dominios externos que estén hospedados en Ips nacionales que no se tenga conocimiento.

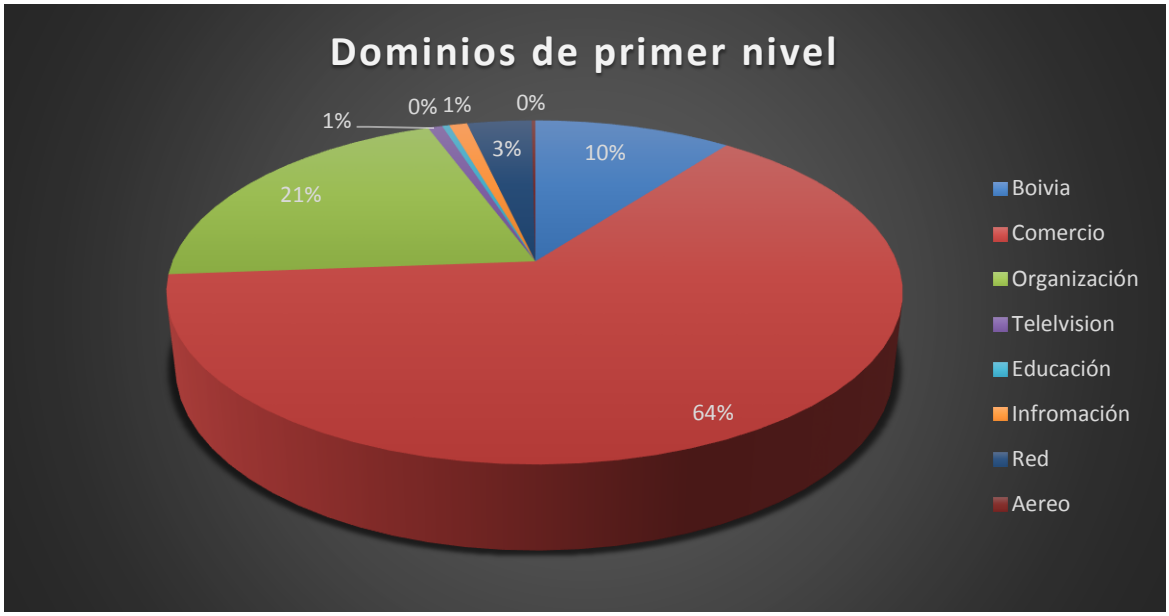


Gráfico 14: *Se observa la relación porcentual de los dominios de primer nivel de la web boliviana. (Bolivia= .bo).*

5.5. Popularidad de los dominios

Se dice que es realmente casi imposible conocer la cantidad de visitantes que recibe cualquier sitio de internet, a no ser que el propietario o administrador de la Web brinde esa información. Los datos que ofrecen todos los programas de analítica y estadística web como Google Analytics y otros similares, son exactos pero privados. Existen servicios de internet que ofrecen información y varias métricas de los sitios, tales como WooRank, sobreintenet.com y otros. Cuando se trata de verificar la cantidad de visitantes, ninguno de estos servicios es confiable, simplemente porque no tienen forma de saber ni de obtener esta información.

A pesar de lo anteriormente descrito puede ser posible calcular de forma aproximada el número de visitas de un sitio web, usando los datos sobre su tráfico que brinda Alexa, este calcula la popularidad de cada uno y se le asigna un índice o Rank. La ecuación utilizada para este objetivo será:

$$\text{Visitantes mensuales} = 104\,943\,144\,672 \times \text{Índice de Alexa}^{-1{,}008}$$

Ecuación 3: *Calculo de índice aproximado de visitas a una web. (norfiPC, 2016)*

En la tabla 8 se muestra la popularidad de los 30 dominios más visitados de nuestra muestra de las web bolivianas obtenidas con el script del Anexo L, las cuales fueron comparadas con datos de la página oficial de Alexa.

N°	SCRIPT ANÁLISIS	N° DE VISITAS REGISTRADAS	N°	RANK ALEXA
1	www.eldeber.com.bo	4466674	1	www.eldeber.com.bo
2	www.eju.tv	3225738	2	www.eju.tv
3	www.trabajopolis.bo	2355484	3	www.lostiempos.com
4	www.boliviaentusmanos.com	1757045	4	www.trabajopolis.bo
5	www.bnb.com.bo	1522361	5	www.la-razon.com
6	www.eldiario.net	1272395	6	www.boliviaentusmanos.com
7	www.erbol.com.bo	1245677	7	www.bnb.com.bo
8	www.bancounion.com.bo	1238623	8	www.erbol.com.bo
9	www.udabol.edu.bo	1035318	9	www.eldiario.net
10	www.diez.bo	1003185	10	www.olx.com.bo
11	www.unitel.tv	1000638	11	www.bancounion.com.bo
12	www.tibo.bo	853089	12	www.ibce.org.bo
13	www.eldia.com.bo	820472	13	www.diez.bo
14	www.correodelsur.com	794913	14	www.opinion.com.bo
15	www.hoybolivia.com	763557	15	www.tigo.com.bo
16	www.sociales.com.bo	712005	16	www.sociales.com.bo
17	www.sociales.com.bo	712005	17	www.hoybolivia.com
18	www.bisa.com	655122	18	www.entel.bo
19	www.ucb.edu.bo	580517	19	www.tibo.bo
20	www.bmsc.com.bo	576178	20	www.udabol.edu.bo
21	www.bcp.com.bo	525393	21	www.unitel.tv
22	www.abi.bo	502569	22	www.bisa.com
23	www.umss.edu.bo	499378	23	www.correodelsur.com
24	www.scielo.org.bo	420319	24	www.eldia.com.bo
25	www.boa.bo	370265	25	www.bmsc.com.bo
26	www.boliviaimpuestos.com	361739	26	www.locanto.com.bo
27	www.cambio.bo	331782	27	www.abi.bo
28	www.comteco.com.bo	330157	28	www.bcp.com.bo
29	www.anunico.com.bo	328946	29	www.viva.com.bo
30	www.fcpn.edu.bo	323693	30	www.umss.edu.bo

***Tabla 8**: Lista comparativa de popularidad de dominios web bolivianas en el mes de junio*

6. Conclusiones y trabajos futuros

Cuando se inició la investigación y recolección de datos en el mes de mayo, no se contaba con ningún tipo de información relevante sobre las características de Web Boliviana, los resultados obtenidos a pesar de no contar con datos de referencia de estudios anteriores, no deja de ser sorprendentes si los comparamos con datos de estudio de otros países con similares contenidos. A pesar de estar la Web en constante cambio, es posible que en futuros estudios se pueda determinar que la estructura sea la similar o se siga manteniendo a la encontrada en el presente estudio.

Después de realzar el estudio y análisis de los datos obtenidos, podemos determinar las características de la web boliviana, las cuales se detallan en los siguientes párrafos:

En la primera parte de la sección del estudio, ***sobre los documentos,*** se lograron analizar diversas características de los documentos, a pesar de tener algunos problemas en la obtención de los dominios registrado por NIC Bolivia se pudo determinar la cantidad de muestra para la colecta y estudio de la cantidad de documentos creados en los últimos años de los sitios Web de Bolivia.

Se logró estudiar diversas propiedades de los documentos, donde el idioma oficial de Bolivia que es el castellano, mantiene su presencia en los sitios, y el idioma inglés, de igual manera mantiene cierta presencia en algunos sitios, se pudo determinar la cantidad de enlaces funcionando actualmente además de los enlaces que ya no existen.

En la secciones relacionado a los ***sitios*** se consideraron un poco más de 1500 establecidos como muestra para la colecta, de los que se pudieron recolectar cerca de 604215 páginas, de estos 32 se lograron identificar como sitios de una sola página, aunque alguno de ellos tenía más documentos. También se pudo determinar la distribución de documentos por sitio, la edad de los meses, la distribución de los enlaces internos, enlaces entre sitios y algunas otras propiedades. Se observó que los sitios con más documentos y con más contenidos suelen ser con frecuencia los sitos del gobierno, de instituciones educativas y medios de comunicación.

Ya en el contenido que respecta a los ***dominios*** se pudo observar que la proporción entre sitios y dominios es prácticamente de uno a uno sin considerar los sitios vacíos, a pesar de haber realizado una pequeña colecta manual, no se pudieron detectar dominios con más de un sito. También se pudo determinar el estudio de la tecnología que utiliza el servidor, en las que se puede observar que tanto en el servidor utilizado como en el sistema operativo, las tecnologías que proporciona Windows tienen una mayor presencia.

Y por último, todos los análisis presentados a los largo de cada título de contenido del presente documento, nos permiten no solo establecer un modelamiento de la Web en términos analíticos o matemáticos, sino nos permites también obtener datos concretos que puedan servir como base para estudios de usabilidad, de mercado y minería de datos, entre otros.

Como trabajo futuro se propone lo siguiente:

- Crear un crawler con una interfaz gráfica y amigable.
- Ampliar los criterios de estudio para un mejor análisis de la estructura de la Web Boliviana.
- Generar futuros trabajos de análisis para evaluar el grado de evolución de la Web Boliviana.

- Proponer la construcción de un Cluster específico para minimizar el tiempo de estudios de estructuras de Web nacionales.
- Solucionar el problema de velocidad de conexión en la recolección de datos utilizando tecnologías alternativas (LIFI).

Lo que se pretende mostrar en el presente documento es una captura de un instante particular de la existencia de la Web, cuya representatividad se pueda afirmar con más certeza con estudios futuros que se puedan realizar en los siguientes años, que sin embargo la constancia de estudios con otras Web nacionales y sus respectivas comparaciones, nos muestra que el presente estudio goza de una total representatividad.

Bibliografía

Baeza-Yates, R., & Graells, E. (2008). *Características de la Web Chilena 2007.*

Clompetch. (s.f.). *www.clomputech.com*. Obtenido de http://www.clomputech.com/paginas-estaticas-vs-dinamicas.html

Delgado Rodriguez, H. A. (30 de Julio de 2015). *www.disenowebakus.net.* Obtenido de http://disenowebakus.net/elementos-de-un-sitio-web.php

editafacil. (s.f.). *http://blog.editafacil.es*. Obtenido de http://blog.editafacil.es/nuevas-extensiones-de-dominios-ventajas-o-inconvenientes/

Google. (s.f.). *www.support.google.com*. Obtenido de https://support.google.com/a/topic/14075?es&ref_topic=9197

Granada, U. d. (s.f.). *www.ugr.es/*. Obtenido de http://www.ugr.es/~focana/dfar/aplica/ajpotencial/ajupoten.pdf

IBM. (s.f.). *www-03.ibm.com*. Obtenido de http://www-03.ibm.com/software/products/es/ibmwebexpefact

Malvezzi, M. D. (Abril de 2010). *www.webasesor.es.* Obtenido de www.linkedin.com/in/manueldocavo/es

Netberry. (s.f.). *www.netberry.co.uk*. Obtenido de http://netberry.co.uk/alexa-rank-explained.htm

norfiPC. (2 de julio de 2016). *norfipc.com*. Obtenido de https://norfipc.com/herramientas/como-saber-visitantes-diarios-sitio-web-internet.php

pendientedemigracion. (s.f.). *www.pendientedemigracion.ucm*. Obtenido de http://pendientedemigracion.ucm.es/info/sevipres/P4/01/ANEXOS01.php

Piura, U. d. (s.f.). *www4.ujaen.es/*. Obtenido de http://www4.ujaen.es/~emilioml/doctorado/guia_rapida_de_citas_apa.pdf

Razon, L. (2013). *www.la-razon.com*.

Tolosa, G., A. Bordignon, F., & J. Lavallén, P. (2007). Caracterización del Espacio Web de Perú. 10.

Tolosa, G., A. Bordignon, F., Baeza-Yates, R., & Castillo, C. (2007). Caracterización del Espacio Web de Argentina. 12.

Webconfs. (s.f.). *www.webconfs.com*. Obtenido de http://webconfs.com/domain-age.php

Wikipedia. (s.f.). *www.wikipedia.org*. Obtenido de https://es.wikipedia.org/wiki/Megabyte

WordReference. (s.f.). *http://www.wordreference.com/*. Obtenido de http://www.wordreference.com/sinonimos/positivo