# Una forma diferente de acercamiento al análisis de Internet

Víctor Angel Fernández

## Índice



## 1 Introducción

Durante los últimos años he tenido la oportunidad de estar ligado al análisis de la presencia cubana en Internet y como caso particular el de la entrada, desarrollo y posicionamiento de la prensa cubana en la Red de Redes. En este tiempo he visitado una amplia mayoría de los sitios cubanos y en el caso particular de la prensa a todos, con el objetivo de conocerlos, analizarlos y poder establecer bases de comparación aplicables como modelo para evaluar la calidad y cumplimiento de su objetivo, en tanto vehículo de presentación de la realidad cubana y también como elemento que pueda competir ante la abrumadora cantidad de sitios que ofrecen sus versiones, positivas, negativas, objetivas o no, sobre el muy tratado tema cubano.

En todos los casos, las bases de análisis han partido de las conocidas facilidades que más se evalúan en Internet, o sea, si se tiene acceso a las estadísticas, como son los sitios hospedados en Cubaweb, cuyo servicio estadístico es abierto, compilar las mismas y sacar conclusiones a partir de accesos (hits), cantidad de visitas, páginas vistas, países de los visitantes, tiempos de estancia en las páginas, errores de presentación, principales puntos de envío hacia los sitios (referrers) y otros datos que las mismas ofrecen.

También ha existido el análisis de la composición propia del código fuente de las páginas, donde se incluyen el tratamiento de las diferentes secciones, a saber, Título, Meta Tags o el cuerpo propiamente dicho de la página.

Muchos de estos análisis se basaron en los resultados obtenidos al someter diferentes páginas de esos sitios a los analizadores internacionales y gratuitos que se pueden encontrar en Internet, principalmente Web Site Garage y Hit Box, los cuales ofrecen un estudio muy completo y sobre todo, muy sugerente de las páginas que a ellos se le someten.

Aunque los resultados en cualquiera de los casos, son muy serios, siempre tienen una componente importante, basada en el punto de vista del analista que de hecho puede al-

terar la muestra y las correspondientes conclusiones, no obstante que su repetición periódica, va sugiriendo acercamientos y soluciones al tema que permiten, a falta de otros, implementarlos como norma a seguir para un análisis serio y proyectable de la realidad de los sitios cubanos en Internet, fundamentalmente de los pertenecientes a la prensa, donde se ha ceñido el objetivo principal de las investigaciones.

Buscando nuevos horizontes a estas capacidades de análisis de sitios web, encontré el artículo "Impacto de los sitios web: una comparación entre Australasia y América Latina", presentado al Congreso INFO de La Habana en 1999, por el profesor Alastair G. Smith, en cuyo inicio, él se pregunta cuál es el impacto de los sitios de Internet sobre los recursos generales de información que esta red ofrece y más adelante plantea que su estudio se basa en la webmetría (webmetric o webometric en inglés) como disciplina emergente que aplica las técnicas y modelos matemáticos de estudio a Internet y que para ello utiliza el concepto de Factor de Impacto en la Web, FIW (WIF o Web Impact Factor), una medida que parte del número de links hecho a un sitio cualquiera para determinar la influencia general que tiene el referido sitio sobre la Web.

El conocimiento de este trabajo, así como otro del mismo autor, pero dedicado sólo a estudios de sitios universitarios australianos y neozelandeses, tuvieron la virtud de introducirme en una variante matemática del análisis y hacer regresar los estudios a temas muy conocidos por los especialistas del área de documentación, como son los relacionados con Bibliometría, Informetría, Cienciometría y las leyes en que estas disciplinas se basan, desarrolladas por Zipf, Bradfrod, Solla Price y Lotka, entre otros, para medir la productividad del trabajo de los científicos, la de los autores, así como las leyes de análisis de los artículos y de las revistas científicas.

## 2 Leyes para estudiar

En el año 1973, gracias a la colaboración entre el Instituto de Documentación e Información Científica y Tecnológica (IDICT) y la editorial Nauka de la Unión Soviética, se publicó en La Habana el libro, "Fundamentos de la Informática" de los autores Mijailov, Chiornii y Guiliarevski, (publicado originalmente en idioma ruso en Moscú en el año 1968), una especie de biblia, para todos aquellos que nos iniciábamos en el campo de la información científica, la documentación y la bibliotecología en esas épocas lejanas.

El texto recorre toda la Actividad Científico-Informativa hasta aquellos momentos y en él se definían y explicaban exhaustivamente algunos de los términos y leyes que se retoman en el presente trabajo. Antes de continuar, es importante recordar que en una de las conferencias presentadas al Congreso de IFLA de Jerusalem en el pasado año 2000, se reconoce este libro y esta edición en particular, como muy influyente en los estudios bibliográficos y bibliotecológicos, realizados en América Latina en las dos décadas pasadas.

Los análisis que dan pie a las disciplinas científicas nombradas en la Introducción de este trabajo, se adaptan casi perfectamente a las necesidades de análisis que hoy presenta Internet, pues en la web, al igual que en la literatura científica, cualquiera que sea su nivel de seriedad, también se ha producido el crecimiento incontrolado e incontrolable de

sitios dedicados a todo tipo de temas. También se produce la interrelación entre los sitios que tratan temas iguales o similares, llegándose a hablar ya de la cantidad de clicks necesarios para viajar desde un sitio determinado hasta cualquier otro sitio en la Red, basándose sólo en la relación de enlaces entre los mismos o entre ellos y terceros similares, sin que los dos primeros tengan un enlace directo.

Los citados autores plantean, por ejemplo, que "todas las publicaciones periódicas pueden agruparse por zonas concéntricas en orden descendiente de su productividad", considerándose por productividad, la capacidad de una publicación o de un artículo científico para satisfacer la necesidad de información del lector del mismo. A continuación explican que "aumentará el número de publicaciones periódicas en cada zona y disminuirá proporcionalmente su productividad".

Si se toma esta definición y en los lugares donde dice publicaciones periódicas, se sustituye por sitios-web, casi la definición parece escrita para resolver el problema actual de navegación y búsqueda de información en la red y no para una disciplina explicada hace ya más de treinta años.

Veamos entonces algunas de las leyes de interés, que utilizan métodos de análisis cuantitiativo y estadísticos para describir modelos de investigación dentro de un campo determinado de las ciencias o de la vida en su más amplio concepto.

## 3 Ley de Bradford

Expresada en las primeras décadas del siglo XX, esta ley, sirve como una guía general para los estudiosos de la documentación, que desean determinar el número de publicaciones núcleo en un campo determinado. La misma, parte de dividir un campo de la ciencia en tres partes, cada una de ellas contentiva de la misma cantidad de artículos.

Parte 1.- un núcleo de revistas especializadas en el referido campo, que producen una tercera parte de los artículos

Parte 2 – una zona conteniendo el mismo número de artículos significativos que la primera, pero extraídos de un número mayor de publicaciones

Parte 3 – la última zona que contiene la misma cantidad de artículos de interés de la anterior, pero localizados todavía en una mayor cantidad de publicaciones.

Es obvio que a la primera zona corresponderán las revistas más especializadas en el tema buscado y que en las zonas subsiguientes irán apareciendo las menos especializadas o las de poca relación con el tema en cuestión, que no obstante, cada cierta cantidad de tiempo publican algún artículo de interés.

Su modelo, también conocido como Ley de Dispersión, describe el carácter lineal de la dispersión de los artículos científicos en las publicaciones periódicas y lo representa por

P : P1 : P2 = 1 : N1 : N2

Donde p, p1 y p2 son la cantidad de revistas contenidas en el núcleo y n, n1 y n2, las relaciones entre las mismas y la cantidad de artículos que en ellas aparecen.

B. C. Vickery en 1948, demostró matemáticamente que estas conclusiones podían tener ciertas curvaturas en la propuesta linealidad y desarrolló una variante para la misma que corregía el error.

Esta ley y su aplicación ha sido un elemento fundamental para desarrollar, por ejemplo, planes de compra de publicaciones periódicas. Así, ante la avalancha de posi-

bles títulos para adquirir, usted puede localizar los núcleos por temas y realizar su selección más acorde con la realidad de respuesta en artículos que necesita.

Trasladando esta ley a la situación actual de dispersión de la información en Internet, puede significar grandes ahorros de tiempo el aplicar esta ley para los análisis de sitios en Internet y similarmente irlos ubicando en núcleos y zonas periféricas según su mayor o menor profundidad en el tratamiento de los temas buscados.

## 4 Ley de Zipf

Estudios posteriores de la Ley de la Dispersión de Bradford, demostraron que la misma no era más que un caso particular de la Ley de Zipf, sin que por ello minimizara en ninguna medida el valor de la primera.

Se parte de la base de que un texto, cualquiera que este sea, y que posea cierta longitud, tendrá cierto grupo de palabras, muy pequeño, que se repiten una gran cantidad de veces, por ejemplo las preposiciones o las conjunciones, pero con la características de no tener valor intrínseco por sí mismas. De igual forma irán apareciendo otras cantidades de palabras, muy pocas veces repetidas, pero que dado su valor de significado, definen el contenido del texto escrito. Este análisis permite establecer rangos para las palabras, de acuerdo con su índice de repetición y con su valor en el texto. Entre la frecuencia con que aparece una palabra en el texto y el rango de la misma se establece la siguiente relación:

Pr= 0,1/r

Donde P es la posibilidad de encontrar en el texto la palabra r y r, a su vez, representa el rango.

Esta ley expresa propiedades inherentes a las lenguas universales, que de forma muy común y corriente se escuchan cuando algunas personas dicen saber inglés técnico, francés técnico o alemán técnico y no están diciendo otra cosa que conocen los significados de los “núcleos” de palabras de mayor rango en su rama particular. La Ley de Zipf, puede llevarse incluso, hasta los análisis de las composiciones de las letras en un texto, variantes estas de mucha utilización en el campo de la criptografía y la decodificación.

Según un análisis del Ulises de James Joyce, las primeras 10 palabras en el rango aparecen 2653 veces, mientras que el centenar siguiente sólo se repiten en 256 ocasiones y así sucesivamente.

## 5 Ley de Lotka

Esta ley, describe la frecuencia de publicación por autores en un campo dado, conocido como productividad de los autores y expresa que el número de autores que hacen N contribuciones (o sea que publican N artículos) es aproximadamente $1/N^2$ de aquellos que hacen sólo una contribución. De igual forma, la proporción de todos los contribuidores es aproximadamente el 60% de los que hacen una sola contribución.

Estas tres leyes, expresan de hecho variantes sobre un mismo tema, a saber, la necesidad de conocer las publicaciones o los autores más productivos, que a su vez serán los más importantes en la rama determinada y con ello, lograr un significativo ahorro de tiempo al decidir el horizonte (creciente en forma exponencial) de fuentes a consultar.

Obviamente que la pregunta estará dada en cómo conseguir estas informaciones y estos estudios, pues al parecer a simple vista

no estarán a la mano de cuanta persona lo necesite.

La base fundamental de análisis para estos estudios, ha sido la costumbre surgida en el siglo XIX, de citar en el cuerpo o al final de la obra, los autores o artículos consultados para llevar a cabo un escrito o una investigación en particular. El desarrollo de esta actividad llevó a la creación de índices de citas, donde se presentan no ya los contenidos de los trabajos propiamente dichos, sino el contenido de las citas bibliográficas que hacen los mismos y de ahí llegar a una conclusión de autoridad de la referida obra, del autor o de la publicación que se esté analizando.

El más completo y accesible de estas obras, es el Science Citation Index, producido periódicamente por el Institute for Scientific Information de Fildelfia en los Estados Unidos.

El uso más común del análisis de citas es para determinar el impacto de un autor o una publicación en determinado campo, partiendo de la cantidad de veces que es citado su trabajo por otros autores. Es obvio que a veces estos autores son citados en forma negativa, o sea, algo así como "fulano (ahí está el nombre) no sabe nada sobre el referido tema".

Una variante más compleja de este análisis de citas, el llamado análisis de parejas de citas (cocitation coupling), el cual refiere a la relación de dos autores independientes, debido sólo al hecho de que son citados por un tercer autor. Expresado más o menos por la conocida ley matemática de carácter transitivo, o sea, si el artículo del autor A es citado por el autor B y este también cita al autor C, entonces existe una relación (por carácter transitivo) entre los autores A y C.

Llegados a este punto, retomamos el tema de la aplicación de estos modelos matemáticos en el análisis de sitios-web y los resultados que pueden ofrecer los mismos, tanto en la evaluación de los sitios, como en la orientación de la búsqueda dentro de la gran y creciente madeja de Internet.

Los autores Almind e Ingwersen en su trabajo *Informetric Analysis of the World Wide Web*, introducen el término "webometric", como la forma de aplicar algunos de estos conceptos a la web. También se va popularizando el término "sitation", propuesto por Rousseau en *Sitations: an exploratory study*, para expresar las variantes de citas entre uno y otro sitio, manipulando la utilización de la "c" y la "s", según se adapten a citas bibliográficas o enlaces de sitios.

Todos parten de la base de que la dificultad para la búsqueda de más y mejor información es igual para los llamados soportes tradicionales que para los nuevos medios digitales en la red y que también se incluyen los factores tiempo y dispersión para encontrar la información deseada.

Ellos y otros autores citados por el profesor Smith, también incluyen una complicación intrínseca a la documentación existente en Internet, que es la desaparición de información, debido a factores de tiempo o espacio en servidores, en la misma medida que trabajos de cualquier nivel de interés permanecen intemporalmente localizados en la red.

Los citados autores concuerdan en que esos trabajos informétricos y cienciométricos realizados sobre documentos aparecidos en soportes tradicionales, sea papel, medios magnéticos u otros, tienen las características necesarias para recibir el traslado analítico, no mimético, de estas conocidas técnicas evaluativas.

La base de los documentos y sus relacio-

nes en Internet es el hipertexto con los enlaces de uno a otro sitio o página en cualquier orden de ida o de retorno, lo cual en la realidad no es más que una cita de un documento a otro. Asimismo es posible aplicar las técnicas de emparejamiento de citas, debido a los conocidos listados de enlaces relacionados que muchos sitios ofrecen.

Estos análisis podrán revelar si existe un impacto de un sitio determinado, sobre su tema particular, dentro de toda la red y revelar además, de qué nivel es el referido impacto. Esto se logra con el conocimiento de la cuantía de enlaces existentes a un sitio, la discriminación de los enlaces hechos por partes internas del mismo sitio o por otros sitios externos que conocen y valoran su importancia dentro de una rama del saber o simplemente sobre un posible término de búsqueda.

Definidas las cuestiones anteriores, sólo falta entonces el cómo llevarlas a cabo y para ellos se han analizado las propuestas de los autores citados y de otros consultados, así como las experiencias de trabajo con algunos de los buscadores más conocidos.

A partir de la utilización de la llamada popularidad de enlaces (link popularity) muchos de estos buscadores permiten conocer la cantidad de enlaces (links) que tiene un sitio y con ello establecer un rango para ubicar los más “sitados” (utilizando ya el término con la letra “s”, anteriormente explicada). Por ejemplo, en http://www.marketleap.com es posible obtener un análisis de ranking realizado por los más conocidos buscadores, precisamente a partir de la cantidad de enlaces que existen a las direcciones que se le suministran.

En el caso particular de Google, su ranking lo establece precisamente por ese análisis de enlaces y posee al mismo tiempo una opción de búsqueda que permite conocer directamente la cantidad de links (enlaces) que tiene una determinada dirección en Internet. Para ello se utiliza la sintaxis:

Link: www.sitio-deseado.com

Y se obtienen las cantidades solicitadas.

Estas y otras posibilidades se pueden encontrar analizando las opciones de Search Advanced o Search Help que poseen los buscadores.

No obstante, aunque muchos autores y especialistas se cuestionan la forma, a veces incomprensible, en que Altavista establece su rango de sitios, al mismo tiempo, al analizar sus capacidades de búsqueda, es en estos momentos la herramienta idónea para aplicar algunos de los métodos antes explicados. Al igual que Google posee una opción de link:www.sitio-deseado.com, con resultados iguales en correspondencia con los sitios que posee, pero tiene además las opciones host:www.sitio-deseado.com, para saber cuántos de esos enlaces son internos, o sea realizados desde páginas del mismo sitio analizado. Asimismo ofrece la opción domain:.cu, que refiere la cantidad de sitios que posee con un dominio determinado.

Pero la opción más importante, sólo ofrecida por Altavista entre los buscadores analizados y que ha llevado a que algunos analistas lo señalen como el Science Citation Index de la Red, es que permite a un mismo tiempo establecer búsquedas que combinen las opciones anteriores con la búsqueda de temas. Por ejemplo: cuba AND domain:.cu ofrece como respuesta los sitios que tratan el tema Cuba y que tienen como dominio CU. Y esta combinación booleana se va a las variantes más conocidas incluyendo NOT y las agru-

paciones de términos con paréntesis, lo cual hace casi ilimitada sus posibilidades.

## 6 Conclusiones

A partir de variantes ampliamente conocidas en otras disciplinas es posible y recomendable utilizarlas para obtener datos matemáticamente demostrados sobre los posicionamientos y capacidad de respuesta reales de los sitios cubanos, principalmente en el caso de la prensa, a las solicitudes y temas más importantes que los mismos deben cubrir.

## 7 Anexo

**Descripción de la Hoja de Búsqueda de Altavista**

**And**

Busca documentos que contienen todas las palabras especificadas o las combinaciones de frases. La solicitud agua AND tierra, devolverá los documentos que respondan a los dos términos al mismo tiempo.

**Or**

Encuentra los documentos que contenga al menos uno de los términos especificados. En esta caso, la solicitud agua OR tierra responde con los documentos que contengan los dos o uno cualquiera de los términos.

**And not**

Excluye los documentos que contienen una palabra o una frase en específico. Siguiendo con el ejemplo, agua AND NOT tierra, devolverá las respuestas que se refieran a “agua” y al mismo tiempo no tengan referencia al término “tierra”

**Near**

Encuentra los documentos que contengan las palabras especificadas con una distancia en el término no mayor de 10 palabras entre ellas. La solicitud agua NEAR tierra, responderá, por ejemplo, con un documento de este contenido agua palabra1 palabra2 palabra3 tierra, pero, de igual forma no responderá con un documento de este tipo: agua palabra1 palabra2 ... palabra10 palabra11 tierra.

( )

Los paréntesis agrupan elementos complejos de las búsquedas boolenas. Por ejemplo, (agua AND tierra) OR (barcos AND autos)

**Anchor:texto**

Encuentra las páginas que contienen la palabra especificada, pero sólo cuando está formando parte de un enlace o vínculo. Por ejemplo, anchor:agua responderá con la página que tenga un enlace de este tipo: <a href.:http://www.unsitio.com>agua</a>. Es importante no poner espacios después de los dos puntos (:)

**Applet:class**

Encontrará las páginas que contengan un Applet de java con ese nombre.

**Domain:nombrededominio**

Su respuesta, son las páginas de sitios dentro de ese dominio. Por ejemplo, domain:cu, devuelve todos los sitios o páginas procesados por Altavista que provengan del dominio CU

**Host:nombredehost**

Busca páginas de un hospedaje determinado

**Image:fichero**
En este caso, la respuesta serán las páginas que contengan un nombre de imagen determinado. El nombre se refiere al fichero de las imágenes y no al valor "Alt", que a veces se pone en las mismas

**Like:URLtexto**
Aquí las respuestas son páginas cuya URL tenga alguna relación con la que se ha solicitado.

**Link:URLtexto**
Esta opción busca todas las páginas con enlaces a la URL especificada

**Text:texto**
Busca las páginas que contienen el texto especificado, pero que no esté ubicado en un enlace, un Metatag o una imagen

**Title:texto**
Busca páginas que contiene el texto especificado como parte de la sección TITLE del código htm.

**Url:texto**
En este caso las páginas buscadas son las que tienen en su URL la palabra especificada.

## 8 Bibliografia

Almind e Ingwersen. *Informetric analyses on the Worl Wide Web: methodological approaches to "Webometrics"*. 1996. Revisado en Internet Julio 2001

Boudurides, Sigrist y Alevizos. *Webometrics and the Self-Organization of the European Information Society*. 1999. Revisado en Internet, agosto 2001

http://www.altavista.com (Advanced Search Cheat Sheet)

http://www.alltheweb.com

http://www.cubaweb.cu/stats

http://www.google.com

http://www.hitbox.com

http://www.marketleap.com

http://www.msn.com

http://www.websitegarage.com

http://www.yahoo.com

Mijailov, Chiornii y Guiliarevski. *Fundamentos de la Informática*. Mocú-Habana, Nauka, 1973.

Ríos, Daniel. *La bibliometría: nivel de penetración en la enseñanza bibliotecológica universitaria y su aplicación en el campo bibliotecario en los países del MERCOSUR*. 2000. Revisado en Internet Agosto 2001.

Smith, Alastair. *ANZAC webometrics: exploring Australasian web structures*. Revisado en Internet, septiembre 2001.

Smith, Alastair. *Criteria for evaluation of Internet Information Resources*. 1997. Revisado en Internet, agosto 2001.

Smith, Alastair. *The impact of Web sites: a comparison between Australasia and Latin America*. 1999. Revisado en Internet Julio 2001.