Actas de las 4ª Jornadas de Intercambio y Reflexión acerca de la Investigación en Bibliotecología, La Plata, 29-30 de octubre de 2015. La Plata: Facultad de Humanidades y Ciencias de la Educación, Universidad Nacional de La Plata, 2015. ISSN 1853-5631

Poniendo en contexto los datos de descargas de documentos en repositorios digitales. Una aproximación a través del uso de percentiles

Cristian Merlino-Santesteban[1]

[1] Universidad Nacional de Mar del Plata. Facultad de Cs. Económicas y Sociales
E-mail: csantest@mdp.edu.ar

Resumen: Los datos de descargas de documentos son la evidencia objetiva utilizada para demostrar el uso de los contenidos de los repositorios digitales. Por lo general los rankings de uso de publicaciones, en distintos niveles de agregación, se basan en el recuento del número de descargas. Dadas las especificidades propias de cada campo disciplinar y la naturaleza asimetría de las distribuciones de datos de uso, la instrumentación de rankings basados únicamente en conteos sin normalizar puede dar una imagen distorsionada del desempeño de los documentos. Se propone enriquecer los rankings con medidas de posición relativa (percentiles) a fin de permitir la comparación normalizada del desempeño de una publicación determinada con el desempeño de su grupo de referencia.

Palabras clave: Distribución de descargas, Marco de referencia, Medidas de posición, Percentiles

1. Introducción

Los repositorios digitales son una de las estrategias fundamentales definidas en la Declaración de Budapest de Acceso Abierto (BOAI, 2002) para alcanzar la libre distribución en la Web de la producción académica y científica mundial. Sin duda, hoy en día es incuestionable el valor de estas plataformas tecnológicas abiertas en el proceso de democratización y socialización de los resultados de investigación generados por universidades, centros de ciencia y tecnología y otras organizaciones de I+D. Por supuesto, dicha valoración positiva no sería posible si no estuviera apoyada, además de en juicios valorativos subjetivos, en evidencia empírica que demostrase, con cierta confiabilidad, el acceso y uso de los contenidos depositados.

A la evidencia objetiva que hacemos referencia son los registros de transacciones de archivos (*web log files*) guardados en los servidores web que alojan a los repositorios. Más precisamente a los datos de acceso a los objetos digitales dispuestos en abierto. A partir de los cuales, previa depuración y filtrado de las solicitudes redundantes, se puede extraer y calcular el número de veces que un documento cualesquiera fue descargado durante un período determinado (Merk, Scholze y Windisch, 2009). Estas cuantificaciones, en distintos niveles de agregación, suelen emplearse para dar cuenta del grado de atención, difusión, visibilidad o uso que han alcanzado los contenidos accedidos y utilizarse, asimismo, como insumo para la toma de decisiones en temas asociados a la evaluación de la investigación.

Por lo general los datos de acceso a los objetos digitales, a los cuales nos referiremos en adelante genéricamente como descargas de documentos, son presentados en forma de rankings (e.g., DASH, 2015; LogEc, 2015a, 2015b; SciELO, 2015; SSRN, 2015). Estos rankings listan en orden descendente las publicaciones más descargadas o agregados significativos de éstas en función del número absoluto de descargas recibidas durante un espacio temporal determinado, que no necesariamente es el mismo para todos los documentos analizados[1]. Si bien los conteos brutos sin normalizar provistos por estos ordenamientos nos permiten discernir fácilmente entre los documentos que captaron mayor atención y aquellos que no, poco nos informan sobre el contexto de análisis y menos aún sobre la posición relativa de cada publicación en el conjunto de datos estudiados. Incluso si para comparar agregados de publicaciones se presentan rankings de uso basados en promedios de bajadas recibidas, estas métricas no son robustas ya que su cálculo se ve afectado por el comportamiento asimétrico de las distribuciones de datos de descargas. Por tal motivo, se hace necesaria la utilización de medidas estadísticas más fiables que caractericen de mejor manera este tipo de observaciones con valores anormales (valores distantes del resto). En este sentido el uso de percentiles, que se viene sugiriendo para los datos bibliométricos (Bornmann y Marx, 2013; Bornmann, Leydesdorff y Mutz, 2013), es una elección adecuada para realizar comparaciones normalizadas de desempeño.

2. Aplicación de percentiles a los datos de descargas

Las distribuciones de datos de descargas se caracterizan por ser muy asimétricas (Brody, 2006, p. 131; Doemeland y Trevino, 2014) y seguir un comportamiento similar a las distribuciones de datos de citaciones a documentos, aunque no tan acusado (Moed y Halevi, 2015). Se dice que una distribución es asimétrica cuando hay presencia de valores atípicos (valores muy altos o muy bajos), es decir, hay observaciones muy alejadas de las demás las que ocasionan una mayor dispersión en los datos de la distribución. Este desequilibrio repercute negativamente en el cálculo de métricas basadas en promedios puesto que si bien éstas sintetizan de manera global los datos de la distribución, son más sensibles a la variabilidad de las observaciones.

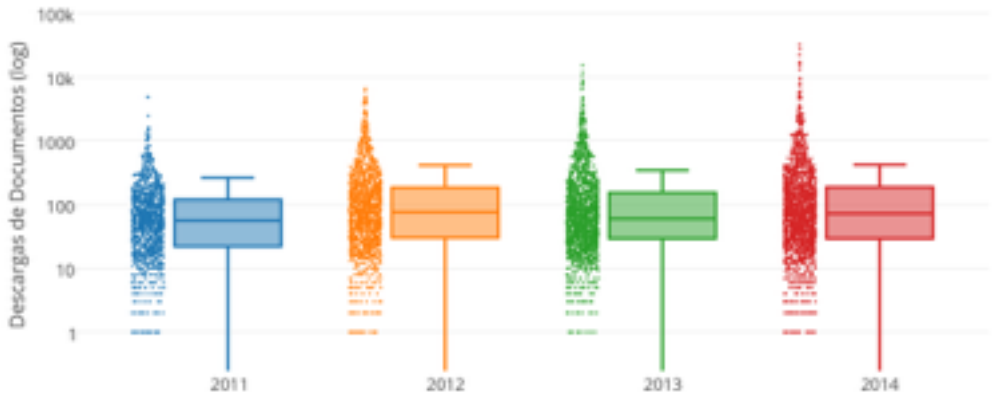
Como por lo general los indicadores presentes en los rankings de uso se basan en simples recuentos de descargas o, en un mejor escenario, en el número medio de descargas por documento, éstos padecen de las mismas debilidades que se les atribuyen a los indicadores bibliométricos básicos basados en conteos de citaciones. Por un lado, los recuentos absolutos de descargas son poco informativos por sí mismo si no se consideran las especificidades propias de cada campo disciplinar y se establece un estándar de referencia con el que confrontar tales conteos; por otro, los valores medios de descargas se ven seriamente afectados por algunos

[1] Si no se define una ventana de descarga común para todos los documentos analizados, los documentos que llevan más tiempo depositados tendrán una mayor ventaja potencial de descarga sobre aquellos documentos que llevan menos tiempo disponibles.

valores extremos, alejándolos del patrón general (una fracción muy pequeña de publicaciones altamente descargas).

A fin de mostrar con datos reales el comportamiento asimétrico de las descargas y con ello evidenciar el sesgo que pueden introducir las observaciones atípicas en el cálculo de indicadores relativos basados en promedios, se exhiben las distribuciones de uso de los documentos correspondientes a un repositorio digital institucional para los años 2011, 2012, 2013 y 2014 (Figura 1) y, seguidamente, la correspondiente estadística descriptiva de cada año. Como se observa notoriamente en los conjuntos de datos presentados al lado de los diagramas de cajas, en todos los años hay presencia de valores muy extremos (documentos altamente descargados) -ubicados por encima de los brazos- que sesgan las distribuciones. Las medidas resumen plasman con claridad ese comportamiento. Alcanza con ver simplemente los valores calculados de la media y la desviación estándar. Ambos son ampliamente influenciados por los datos atípicos, convirtiéndolas en medidas completamente no representativas de los conjuntos de datos observados. La mediana, en cambio, al no contemplar para su cálculo todos los datos observados es casi insensible a esta situación, tornando también este comportamiento extremo como indeseable.

Figura 1. Distribución de descargas de documentos por año



Año 2011	
Media	116,26
Moda	1
Mediana	56
Primer cuartil	22
Tercer cuartil	119
Desviación estándar	269,35
Mínimo	0
Máximo	4.839
Suma	129.402
Conteo	1.113

Año 2012	
Media	227,03
Moda	0
Mediana	76
Primer cuartil	30
Tercer cuartil	185
Desviación estándar	550,30
Mínimo	0
Máximo	6.529
Suma	301.499
Conteo	1.328

Año 2013	
Media	257,65
Moda	20
Mediana	60
Primer cuartil	29
Tercer cuartil	155
Desviación estándar	871,21
Mínimo	0
Máximo	15.179
Suma	394.979
Conteo	1.533

Año 2014	
Media	357,81
Moda	0
Mediana	72
Primer cuartil	29
Tercer cuartil	187
Desviación estándar	1.616,42
Mínimo	0
Máximo	32.012
Suma	622.945
Conteo	1.741

En consecuencia, se necesitan otras medidas -más finas- que caractericen de mejor manera cada publicación en relación a su población de referencia. Una de ellas son los percentiles. A través de ellos podemos indicar la posición relativa que ocupa un determinado documento en relación a todas las publicaciones que conforman el grupo de referencia o control. Como ventajas, podemos señalar que los percentiles no son tan influenciables a los valores anómalos y que se pueden aplicar a distribuciones no normales. En definitiva, se puede sintetizar que estas medidas de localización no son tan sensibles como la media aritmética ni tan insensibles como la mediana ante a presencia de colas (Devore, 2008).

Los percentiles (o centiles) son los cuantiles que dividen la serie de datos en 100 partes iguales, cada una de ellas con el 1% de los datos. El percentil 100 es el mayor valor del set y el percentil 0 es el menor valor. De esta forma podemos decir genéricamente que un determinado valor se sitúa en el percentil 80 (P_{80}) cuando deja por debajo de sí el 80% de las observaciones de la población y simultáneamente es superado por el 20% restante. En otras palabras, cuando estemos expresando que un documento se sitúa en el percentil 90 (P_{90}) con respecto a su nivel de descarga, estaremos diciendo que supera al 90% de las publicaciones del conjunto de referencia. El percentil 50 (P_{50}) es la mediana, por lo tanto es el umbral de descargas que identifica la repercusión promedio.

Ahora bien, para que estas contextualizaciones sean verdaderamente significativas los documentos que conforman el grupo de referencia deben ser “similares” a la publicación a comparar. Se ha advertido que características como la tipología documental y la temática afectan el patrón de descarga de los documentos, introduciendo una alta variabilidad (Moed y Halevi, 2015).

3. Uso de percentiles en la práctica. Un ejemplo sencillo

A modo de demostración y a fin de ejemplificar sencillamente el uso de percentiles, se presenta seguidamente una distribución ficticia de bajadas de documentos para un espacio temporal X (Tabla 1). Las publicaciones se encuentran ordenadas ascendentemente por el número de descargas. A los documentos con el mismo número de bajadas se les asignó el rango promedio. Cada publicación fue asignada a un percentil según su rango usando la formula: (100 * (i-0,5) / n). Los documentos con cero descargas fueron asignados al percentil 0. Cuanto más alto es el percentil de ubicación de una publicación, mayor ha sido el número de descargas que ha recibido en comparación con el resto.

Tabla 1. Cálculo de percentiles para una muestra de 50 documentos

n	i	Descargas	P
50	50	10.508	99,00
49	49	7.233	97,00
48	48	4.567	95,00
47	47	4.323	93,00
46	46	3.122	91,00
45	45	804	89,00
44	43,5	702	86,00
43	43,5	702	86,00
42	42	601	83,00
41	41	589	81,00
40	40	521	79,00
39	38	405	75,00
38	38	405	75,00
37	38	405	75,00
36	36	322	71,00
35	35	313	69,00
34	32	278	63,00
33	32	278	63,00
32	32	278	63,00
31	32	278	63,00
30	32	278	63,00
29	29	134	57,00
28	28	132	55,00

Tabla 1. Cálculo de percentiles para una muestra de 50 documentos (continuación)

n	i	Descargas	P
27	27	121	53,00
26	23	109	45,00
25	23	109	45,00
24	23	109	45,00
23	23	109	45,00
22	23	109	45,00
21	23	109	45,00
20	23	109	45,00
19	19	86	37,00
18	18	76	35,00
17	17	45	33,00
16	16	43	31,00
15	15	37	29,00
14	14	36	27,00
13	12	35	23,00
12	12	35	23,00
11	12	35	23,00
10	10	24	19,00
9	9	15	17,00
8	8	4	15,00
7	7	0	0
6	6	0	0
5	5	0	0
4	4	0	0
3	3	0	0
2	2	0	0
1	1	0	0

n: número de publicación; i: rango promedio de la publicación
descargas: número de descargas recibidas por la publicación; p: percentil

En este sencillo ejemplo podemos apreciar que una publicación que recibió 4.567 descargas se ubica en el percentil 95 (P_{95}). Esto nos indica que el 95% de las publicaciones estudiadas en esta muestra tiene mediciones menores y el 5% registra mediciones superiores. Otra forma de expresarlo, más simple o intuitiva, es manifestar que el documento señalado se encuentra en el Top 5% de las publicaciones más descargadas.

4. Conclusión

Si bien los rankings de uso basados en indicadores básicos de conteos de descargas son un propuesta generalizada para dar cuenta de la repercusión alcanzada por las publicaciones en los repositorios digitales, estos indicadores padecen de importantes debilidades atribuidas a la naturaleza asimétrica de las distribuciones de datos de descargas y a las especificidades propias de cada campo disciplinar. Por dicho motivo se requiere introducir métricas estadísticamente más fiables que permitan, dentro de un marco comparativo común, normalizar los datos de uso a fin de obtener mejores comparaciones. En este sentido, los indicadores basados en percentiles son una solución posible, ya que nos permiten comparar de mejor manera el desempeño de un documento con respecto a un set de referencia. De esta manera podremos determinar, de una manera rápida y sencilla, si una publicación recibió un volumen de descarga "típico" en comparación con publicaciones similares o si ésta excedió los niveles esperados.

5. Referencias bibliográficas

BOAI. (2002). *Budapest Open Access Initiative*. Recuperado de http://www.budapestopenaccessinitiative.org/read

Bornmann, L., Leydesdorff, L., y Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: opportunities and limits. *Journal of Informetrics*, 7(1), 158-165.

Bornmann, L., y Marx, W. (2013). How good is research really?. *EMBO reports*, 14(3), 226-230.

Brody, T. D. (2006). *Evaluating research impact through open access to scholarly communication* (Tesis doctoral, University of Southampton). Recuperado de http://eprints.soton.ac.uk/263313/

DASH. (2015). *DASH Stats*. Recuperado de https://osc.hul.harvard.edu/dash/mydash

Devore, J. L. (2008). *Probabilidad y estadística para ingeniería y ciencias* (7a ed.). México: Cengage Learning.

Doemeland, D., y Trevino, J. (2014). *Which World Bank reports are widely read?* (Policy Research Working Paper No. 6851). Recuperado de https://www.openknowledge.worldbank.org/handle/10986/18346

LogEc. (2015a). *Top 1000 Working Papers by Total File Downloads*. Recuperado de http://logec.repec.org/scripts/itemstat.pf?topnum=1000&type=redif-paper&sortby=td

LogEc. (2015b). *Top 1000 Journal Articles by Total File Downloads*. Recuperado de http://logec.repec.org/scripts/itemstat.pf?topnum=1000;type=redif-article;sortby=td

Merk, C., Scholze, F., y Windisch, N. (2009). Item-level usage statistics: A review of current practices and recommendations for normalization and exchange. *Library Hi Tech*, 27(1), 151-162.

Moed, H. F., y Halevi, G. (2015). On full text download and citation distributions in scientific-scholarly journals. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23405

SciELO. (2015). *Memórias do Instituto Oswaldo Cruz. Article requests*. Recuperado de http://scielo-log.scielo.br//scielolog/scielolog.php?script=sci_statart&lng=en&pid=0074-0276&app=scielo&server=www.scielo.br&dti=20040101

SSRN. (2015). *SSRN Top 10,000 Papers*. Recuperado de http://hq.ssrn.com/rankings/Ranking_display.cfm?TRN_gID=10