Contando Citas
en Artículos de Revistas y Conferencias

Mauricio Marín
Sociedad Chilena de Ciencia de la Computación. Yahoo! Research Latin America, Universidad de Chile. PhD en Computer Science, University of Oxford, UK.
mmarin@yahoo-inc.com

RESUMEN

En este trabajo se presentan estadísticas que muestran la tendencia observada en las ultimas décadas sobre el tipo de publicaciones realizadas en ciencia de la computación. Utilizando como métrica de calidad la cantidad de citas que reciben los artículos de los medios de publicación más difundidos en la disciplina, se compara el impacto de artículos publicados en revistas y conferencias.

Los resultados muestran que las publicaciones en conferencias pueden ser muy relevantes para determinadas áreas de la ciencia de la computación. También el impacto de estas publicaciones en el desarrollo de la disciplina ha ido creciendo en importancia en los últimos años. En varios casos el impacto es mayor que el de los artículos publicados en buenas revistas del área. En este artículo se describen los experimentos realizados para fundamentar estas tres afirmaciones.

Un artículo publicado en una conferencia o congreso es un documento completo de varias hojas con discusión y desarrollo similar al de un artículo de revista científica, el cual ha sido evaluado y seleccionado por un comité científico internacional, con tasas de aceptación bajo 30% y publicado por una de las casas editoriales reconocidas en la disciplina tales como IEEE-CS, ACM y LNCS de Springer. Gran parte de los artículos en revistas y conferencias en ciencia de la computación son indexados por medios alternativos a la indexación del \Web of Science" tales como DBLP, \ ACM Digital Library Portal" y CiteSeer^{x}. Los experimentos del presente artículo utilizan dichos sistemas de indexación como fuente de información para contabilizar citas por cada medio de publicación.

CITAS BIBLIOGRÁFICAS DEL "ACM COMPUTING SURVEYS"

El ACM Computing Surveys es una revista con volúmenes publicados desde el año 1969, en la cual los artículos tienen la forma de revisiones del estado del arte en tópicos específicos de ciencia de la computación. Generalmente se trata de tópicos que ya han sido investigados profundamente al momento de la publicación. Por lo tanto la lista de referencias bibliográficas de dichos artículos da cuenta de manera exhaustiva del estado del arte en el tema y, por supuesto, dichas referencias incluyen prioritariamente los trabajos que presentan las contribuciones más relevantes.

El sitio Web del ACM Computing Surveys muestra todos los volúmenes y números de esta revista a partir del año 1969. Por cada artículo se muestra la lista de referencias bibliográficas. Utilizando esa información calculamos la división entre el total de conferencias detectadas y el total de referencias bibliográficas de cada artículo, considerando la suma de conferencias y revistas.

También calculamos promedios anuales. Consideramos sólo los artículos del ACM Computing Surveys con más de 15 referencias para filtrar artículos tales como Cartas al Editor. El sitio Web también muestra en la lista de referencias bibliográficas de cada artículo las publicaciones que están indexadas en el Portal de la ACM Digital Library. Esto lo hace mediante un enlace al artículo registrado en el Portal, el cual indexa tanto artículos de conferencias como artículos de revistas e indica el tipo de publicación, el que está claramente diferenciado en el contenido del respectivo enlace. Con esto podemos calcular de manera exacta la proporción entre artículos de conferencias y de revistas citados en las listas de referencias bibliográficas.

En la Figura 1 se muestran resultados que abarcan las listas de referencias bibliográficas de artículos que fueron publicados entre 1969 y 2007. Los resultados indican una clara tendencia al alza en la proporción de artículos de conferencias que son citados. Es interesante ver que la Figura además de mostrar los promedios anuales (curva etiquetada con un círculo negro), también muestra la proporción por cada artículo individual (valores indicados con líneas verticales). Se observa que dependiendo del tópico del artículo, la proporción de estos en conferencias puede ser muy dominante en la lista de referencias (más de un 60%). Para otros tópicos, la contribución de los artículos de conferencias puede ser considerada irrelevante (menos de 30%).

FACTORES DE IMPACTO Y CITAS SEGÚN CITESEER^X

El CiteSeer^X (http://citeseerx.ist.psu.edu) es un indexador y máquina de búsqueda especializado en publicaciones en ciencia de la computación. Contiene información estadística desde 1993 a la fecha sobre datos tales como número de citas y utiliza la misma fórmula del ISI para calcular el factor de impacto de revistas y conferencias. Este factor, para un año dado se calcula mediante la división A/B, donde A es el número total de citas a los artículos publicados por la revista/conferencia en los dos años anteriores y B es el total de artículos publicados por la revista/conferencia en esos dos años. Actualmente CiteSeer^X indexa más de un millón de artículos y registra más de 22 millones de citas a los artículos indexados.

En la Figura 2 mostramos los factores de impacto calculados por CiteSeerX en el gráfico hemos agrupado en años, revistas y conferencias. Para esto, bajamos las páginas Web organizadas por año de la sección (\Venue Impact Ratings" de citeseerx.ist.psu.edu/stats/venues) y a estas páginas les aplicamos scripts para contabilizar los factores de impacto asignados a conferencias y revistas por CiteSeerX. En los archivos HTML se puede detectar sin error cuando se trata de una revista o conferencia. Los resultados muestran que las conferencias tienen factores de impacto comparables a los de las revistas cuando no se hace diferencia entre áreas específicas.

En la Tabla 1 mostramos un ranking según valor de factor de impacto de revistas y conferencias conocidas en distintas áreas de ciencia de la computación. Cada columna representa el ranking para los años 2000, 2002, 2004 y 2006. El valor \1" indica que

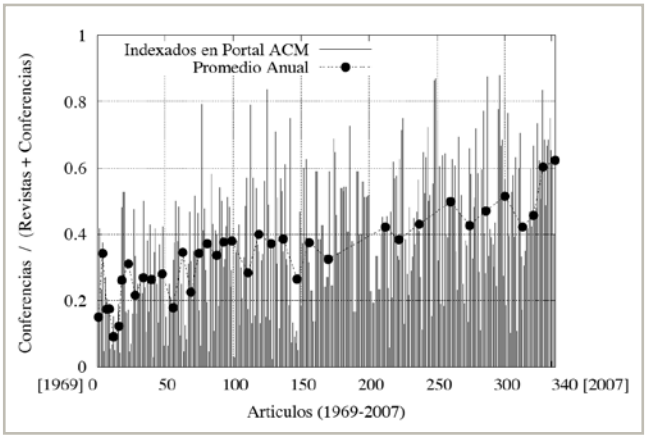

Fig. 1 Proporción de artículos de conferencias citados en cada artículo del ACM Computing Surveys, los cuales están indexados por el portal ACM Digital Library.

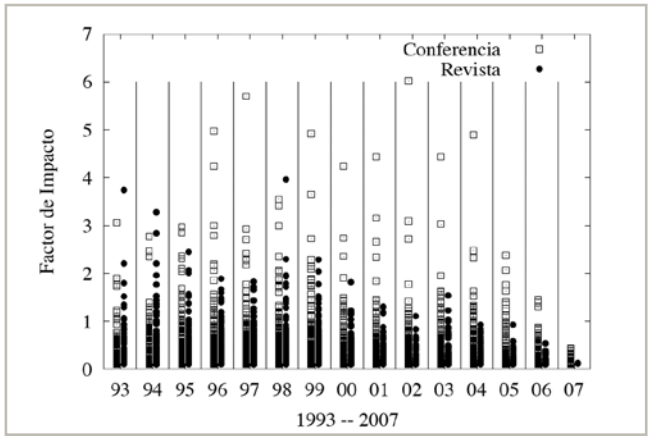

Fig. 2 Factores de impacto tipo ISI para las conferencias y revistas indexadas por CiteSeerX.

Medio	00	02	04	06
Revistas				
TPDS	1	3	2	2
JPDC	3	4	5	5
P.Comp	4	6	6	6
Conferencias				
SPAA	1	3	2	2
IPDPS	3	4	5	5
Euro-Par	4	6	6	6

(a)

Medio	00	02	04	06
Revistas				
TIS	5	2	3	4
TDBS	-	3	4	6
VLDB J.	4	5	6	5
Conferencias				
VLDB	3	1	2	1
SIGIR	1	4	5	3
PODS	2	6	1	2

(b)

Medio	00	02	04	06
Revistas				
JMLR	-	-	3	1
CI	5	5	6	6
ML	2	4	5	5
Conferencias				
IJCAI	4	1	4	4
ICML	1	3	1	2
ACL	3	2	2	3

(c)

Medio	00	02	04	06
Revistas				
TOPLAS	4	4	4	3
IEEE TSE	5	5	5	5
Sci.C.Prog.	6	6	6	6
Conferencias				
PLDI	1	1	3	1
POPL	2	3	1	2
OOPSLA	3	2	2	4

(d)

Tabla 1 Ranking según factor de impacto ISI entre revistas y conferencias en distintas áreas de ciencia de la computación para los años 2000, 2002, 2004 y 2006. (a) Computación Paralela y Distribuida, (b) Bases de Datos y Recuperación de la Información, (c) Inteligencia Artificial (\Machine Learning") y (d) Lenguajes de Programación.

la respectiva revista o conferencia tiene el mayor factor de impacto en su grupo de tres revistas y tres conferencias de la misma área y el valor \6" indica el menor valor de este factor.

VALIDACIÓN DE "CITESEER^{X}" UTILIZANDO "DBLP"

El sistema DBLP Computer Science Bibliography (http://www.informatik.uni-trier.de/ ley/db) mantiene una colección que actualmente contiene sobre un millón de referencias bibliográficas. La base de datos es actualizada con una frecuencia que está prácticamente dentro de la semana en que se publican los nuevos números de las revistas y proceedings de conferencias. En particular es posible bajar desde DBLP un archivo XML que contiene en un formato bien definido los detalles de cada artículo indexado en especial el título, el lugar de publicación escrito de manera consistente para toda la colección y si se trata de un artículo de conferencia o revista.

Para validar los resultados de la Figura 2 bajamos desde CiteSeerX otra sección de este sistema, la cual presenta los diez mil artículos más citados (http://citeseerx.ist.psu.edu/stats/articles). En estos archivos HTML es posible detectar sin error la cantidad de citas que ha recibido el artículo y su título, pero no así el lugar de publicación. Utilizando el archivo XML bajado desde DBLP y los títulos es posible determinar los lugares donde fueron publicados los artículos de CiteSeerX.

Acumulando el total de citas de los diez mil artículos más citados en las respectivas revistas y congresos donde fueron publicados, podemos establecer un ranking de los medios de publicación que consitan el mayor número de citas según CiteSeerX. En la Figura 3 se muestran los resultados, los cuales provienen desde poco más de 200 revistas distintas y casi 400 conferencias. Dichos resultados muestran que un gran porcentaje de las conferencias superan en cantidad acumulada de citas a las revistas. Los artículos de conferencias pueden recibir un número relevante de citas al igual que los artículos de revistas, lo cual se ve reflejado en los factores de impacto ISI mostrados en la Figura 2. De hecho, dentro de los 10 mil artículos más citados obtuvimos el valor 1,04 para la división entre el total de artículos de conferencias y el total de artículos de revistas, y 1,48 al dividir la suma de la cantidad de citas que reciben las conferencias (numerador) y revistas (denominador).

Además el archivo XML de DBLP permite determinar el total de artículos publicados por cada revista o conferencia donde fueron publicados los diez mil artículos más citados en CiteSeerX. Con esto se puede determinar el valor C/P donde C es el total de citas recibidas por los artículos publicados por una conferencia/revista y P es el total de artículos publicados por dicho medio. Es decir, el valor C/P es el promedio de citas que reciben los artículos publicados por

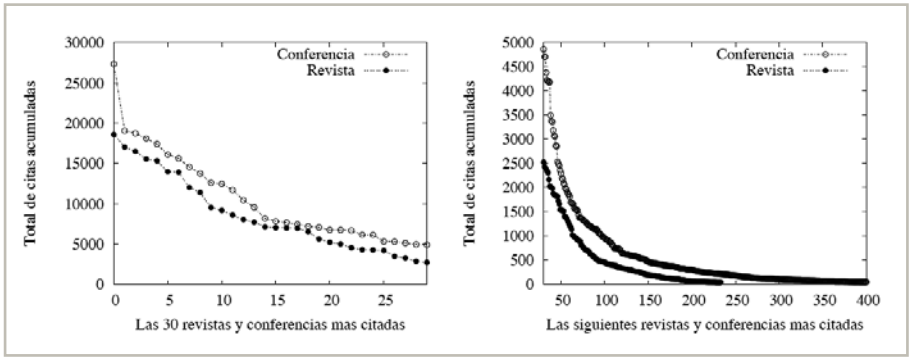

Fig. 3 Total acumulado de citas por conferencia y revista utilizando CiteSeer y DBLP.

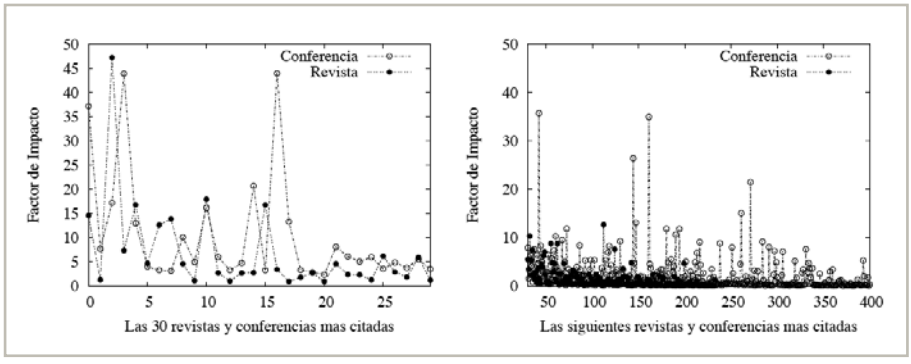

Fig. 4 Número promedio de citas por artículo de conferencia/revista en CiteSeer y DBLP.

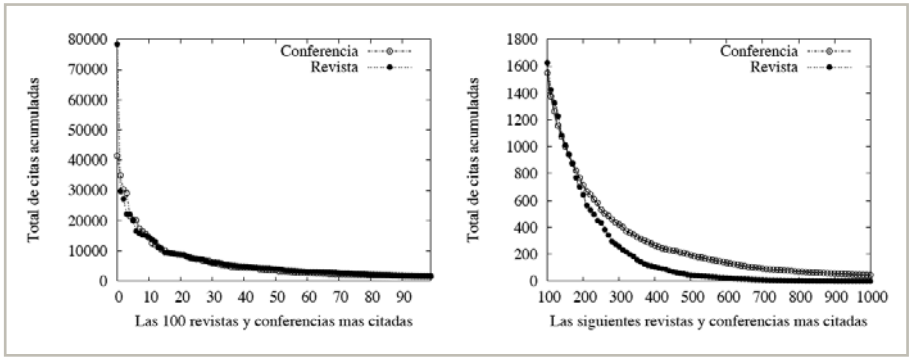

Fig. 5 Total acumulado de citas por conferencia y revista utilizando los artículos indexados en el Portal ACM Digital Library.

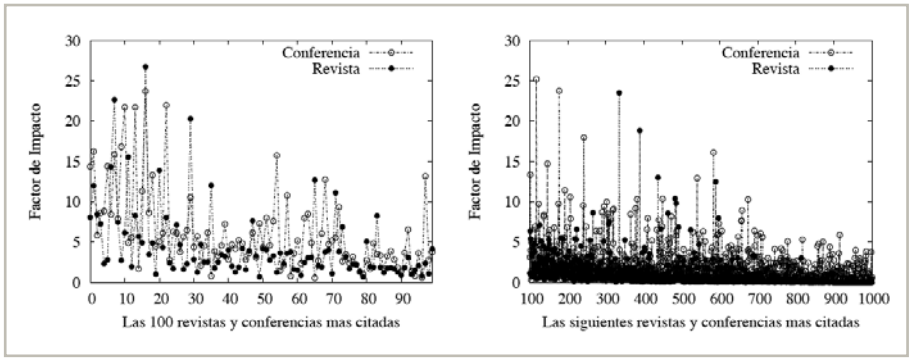

Fig. 6 Citas promedio por conferencia/revista según en el Portal ACM Digital Library.

la revista/conferencia y es similar al índice de impacto ISI, pero considerando todos los años en que el respectivo medio ha publicado artículos. La Figura 4 muestra el factor de impacto C/P que los diez mil artículos más citados le otorgan a la revista o conferencia donde fueron publicados. El orden en que aparecen los datos en el eje X de la Figura 4 es el mismo que el orden dado en la Figura 3.

VALIDACIÓN DESDE EL PORTAL "THE ACM DIGITAL LIBRARY"

El Portal de la ACM Digital Library (portal.acm.org) es un sistema especializado en literatura técnica para ciencia de la computación que contiene una colección de sobre el millón de artículos con sus respectivas listas de referencias bibliográficas e información sobre los artículos que citan a cada artículo. En particular por cada artículo se indica el número total de citas que ha recibido. También es posible diferenciar entre artículos de conferencias y revistas, y la notación de los nombres de cada conferencia/revista es consistente a través de todos los artículos. Utilizamos un cluster de cien procesadores, donde cada uno ejecutó el comando "wget" sobre un URL distinto del Portal ACM para bajar las páginas HTML que son el resultado de ejecutar una búsqueda "vacía" en el Portal. Cada página da acceso a los títulos, autores, "venue" (revista/conferencia) y "citation count" de los 1.233.937 artículos indexados por el Portal al 30 de diciembre de 2008.

Sobre los 62 mil HTML bajados desde el Portal ejecutamos también en paralelo 100 scripts idénticos para obtener por cada revista/conferencia el total de artículos publicados y el total acumulado de citas que recibe cada revista/conferencia a través de sus artículos. Los scripts detectaron un total de 439 mil artículos de conferencia y 456 mil artículos de revistas con más de una cita. El total de citas a los artículos de conferencia es de 912 mil mientras que el total de citas a los artículos de revistas es de 854 mil. Nuestros scripts también detectaron un total de 2.428 conferencias y 1.138 revistas distintas. Los resultados para el total de citas acumuladas por cada revista/conferencia y el número promedio de citas que recibe cada artículo de cada revista/conferencia (*C/P*), para las primeras mil revistas/conferencias se muestran en las Figuras 5 y 6 respectivamente. La tendencia es similar a los resultados obtenidos con las otras muestras de artículos en ciencia de la computación. Es decir, los artículos de conferencia pueden tener una relevancia similar en el avance del estado del arte de la disciplina que los artículos de revistas.

UN ÚLTIMO DATO DESDE "DBLP"

En el XML que bajamos desde DBLP (http://dblp.uni-trier.de/xml) en septiembre de 2008 encontramos que para 28 libros, 6.406 artículos de conferencia y 1.813 artículos de revista, todos publicados antes del año 2005, se incluyen sus respectivas listas de referencias bibliográficas correctamente formateadas con tags XML (actualmente el DBLP no almacena en su base de datos la lista de referencias de los artículos que indexa). Los artículos para los cuales encontramos sus listas de referencias provienen de 22 conferencias y ocho revistas que pertenecen principalmente al área de Bases de Datos. Esto representa una gran oportunidad para analizar lo que sucede en un área clásica y de las más antiguas e importantes de ciencia de la computación. Los libros son

ediciones que están entre los años 1983 y 2004, y las citas en sus listas de referencias abarcan artículos/libros publicados entre los años 1949 y 2001. Los artículos provienen de conferencias anuales que van desde los años 1975 al 2001 y citan artículos/libros publicados desde 1962. Las revistas son números que van desde 1970 al 2001 y citan artículos/libros desde 1945. Las referencias bibliográficas citan artículos publicados en poco más de 150 revistas y 300 conferencias.

En la Figura 7 se muestra el total acumulado de citas por medio de publicación que provienen de los artículos publicados en las conferencias y revistas mencionadas en las listas de referencias bibliográficas de los 28 libros, y todas las ediciones anuales de 22 conferencias y ocho revistas. En la Figura 8 se muestra el promedio de citas por artículo (C=P) que reciben las conferencias y revistas a través de esos artículos. En general, estos resultados muestran la misma tendencia observada en los resultados presentados en las secciones anteriores.

COMENTARIOS FINALES

Tal vez es verdad que gran parte de los avances en ciencia de la computación provienen efectivamente desde artículos presentados en las conferencias más exigentes de cada área de la disciplina. En otras ciencias este tipo de publicaciones no tienen mayor impacto, incluso entre los mismos investigadores de ciencia de la computación no existe consenso al respecto. Sin embargo, los resultados presentados en este artículo indican que para al menos nuestra disciplina este tipo de publicaciones sí tiene importancia. Validamos los resultados utilizando distintas muestras de la literatura indexada por sistemas ampliamente conocidos por la comunidad. En particular, estos sistemas indexan revistas y conferencias que a nuestro juicio están claramente ubicadas en lo que constituye el core de la disciplina tal como se entiende en la iniciativa impulsada en http://www.core.edu.au/.

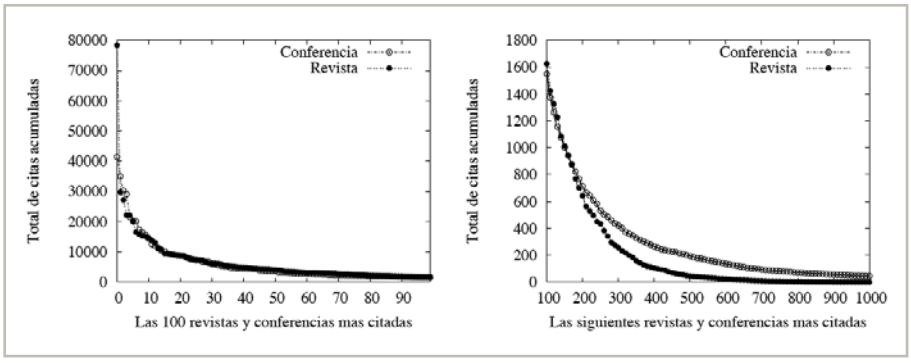

Fig. 7 Total de citas acumuladas por artículos de cada conferencia y revista.

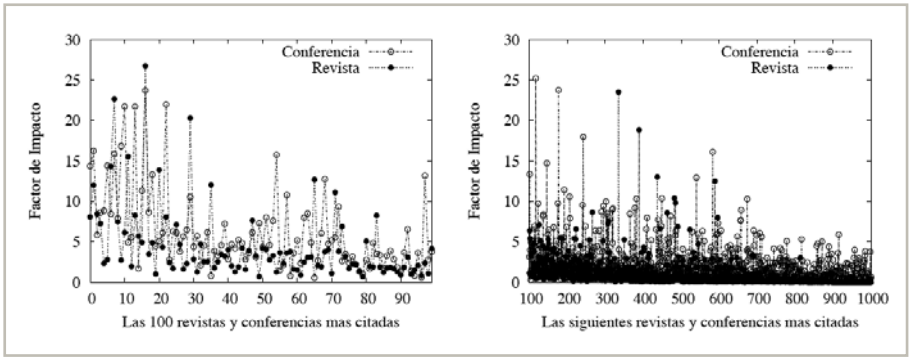

Fig. 8 Factores de impacto de cada conferencia y revista.

AGRADECIMIENTOS

Las revistas y conferencias mencionadas en la Tabla 1 fueron seleccionadas por John Atkinson de la Universidad de Concepción, y Pablo Barceló y Éric Tanter de la Universidad de Chile. Andrea Rodríguez, de la Universidad de Concepción, colaboró con varias sugerencias en distintos puntos de este artículo y ha aportado nuevas estadísticas y vistas de la información para una versión más completa del presente artículo. Senen González, estudiante de Doctorado del DCC, colaboró con los scripts y organización de las máquinas utilizadas para bajar el Portal de la ACM y el procesamiento de los datos. Silvia Menichetti, de la Universidad de Magallanes, colaboró con los scripts utilizados en el ACM Computing Surveys.BITS

Este trabajo ha sido parcialmente financiado por la Sociedad Chilena de Ciencia de la Computación.