Caracterización de las Publicaciones en Ciencia de la Computación

Mauricio Marín
Yahoo! Research Latin America, y
Universidad de Santiago de Chile
mmarin@diinf.usach.cl

Andrea Rodríguez
Centro de Investigación de la Web, y
Universidad de Concepción, Chile
andrea@inf.udec.cl

Resumen Extendido

En este artículo se presentan estadísticas que muestran la tendencia observada en las últimas décadas sobre el tipo de publicaciones realizadas en Ciencia de la Computación. Se utiliza como métrica de calidad la cantidad de citas que reciben los artículos de los medios de publicación más difundidos en la disciplina. En base a este indicador se estudia el impacto en el estado del arte de la disciplina de artículos publicados tanto en revistas como en conferencias, tomando la palabra conferencia como sinónimo de congreso, workshop o simposio. Por publicaciones en conferencias nos referimos a artículos completos evaluados por un comité científico internacional y publicados por editoriales reconocidas tales como IEEE-CS, ACM y LNCS de Springer, donde cada artículo recibe al menos tres evaluaciones y donde la decisión de aceptación es tomada por un comité de programa especialista en el área a la cual pertenecen los artículos, y con tasas de aceptación bajo el 30%. Dichos artículos son habitualmente indexados por medios alternativos a la indexación ISI del "Web of Science" tales como DBLP, "ACM Digital Library Portal", "IEEE Computer Society Digital Library" y CiteSeer^X, los cuales son especializados en Ciencia de la Computación y son ampliamente conocidos por la comunidad.

Los resultados muestran que las publicaciones en conferencias pueden ser muy relevantes para determinadas áreas de Ciencia de la Computación, y el impacto de estas publicaciones en el desarrollo de la disciplina ha ido creciendo en importancia en los últimos años. Aplicando en CiteSeer^X métricas tales como el total acumulado de citas que reciben estos artículos — suma realizada sobre las 100 publicaciones de cualquier tipo (incluyendo revistas y libros) más citadas desde 1990 — se obtiene que las publicaciones en conferencias son responsables de al menos 1/3 de las contribuciones más relevantes para el estado del arte en Ciencia de la Computación. Esto sube a 1/2 cuando consideramos las 10 mil publicaciones más citadas de un total de 22 millones indexadas desde 1990, donde dentro de estos 10 mil artículos existe una cantidad muy similar de artículos en revistas y conferencias.

Los resultados obtenidos desde CiteSeer^X fueron validados utilizando el sistema ISI "Web of Science" (datos a Mayo del 2008), desde el cual hemos contabilizado el total de referencias bibliográficas presentes en cada artículo de la colección de Ciencia de la Computación y Sistemas de Información. En este caso los resultados muestran que, en promedio, más del doble de dichas citas son hacia artículos que no están indexados en ISI, los cuales en su gran mayoría pertenecen a conferencias. Así mismo, una muestra de más de 110 mil citas bibliográficas presentes en las listas de referencias bibliográficas de libros y artículos en el sistema DBLP, nos indican una distribución de 60 mil citas a artículos de conferencias (54%), 42 mil citas a artículos de revistas (38%) y 8 mil citas a libros (8%).

El propio Web of Science da cuenta de la gran importancia de los artículos en conferencias en la literatura de Ciencia de la Computación. Al 29 de Noviembre del 2008, el servicio "Highly Cited Papers (last 10 years)" ejecutado sobre la colección "Computer Science", muestra que el LNCS de Springer (una serie de libros que publica artículos de conferencias – "proceedings articles") es el quinto medio de publicación que aporta los artículos más citados en la disciplina. Es un quinto lugar entre 159 revistas (journals) indexadas en ISI, lo cual objetivamente es excelente si se considera que a partir de la segunda mitad en el ranking de las 159 revistas, cada revista sólo aporta con 1 o 2 artículos. La distribución muestra que en los 10 primeros lugares se concentra el 64% de los artículos más citados en los últimos 10 años. Pero este ranking del Web of Science no considera los artículos de conferencias publicados por otras casas editoriales distintas a Springer. Es decir, los proceedings LNCS representan sólo una parte de las publicaciones en conferencias, existiendo una gran cantidad de artículos publicados en conferencias de muy buen nivel que son publicados por editoriales como ACM y IEEE-CS, y que no están considerados en este ranking del Web of Science. En CiteSeer^X se puede comprobar que muchos de estos artículos son más citados que los artículos publicados por el LNCS de Springer.

No obstante, si ahora en el Web of Science el ranking de revistas se hace por la métrica "Citations Per Paper", es decir, la razón entre el total de citas y el total de artículos publicados en cada medio, el LNCS baja al lugar 30 de un total de 33 revistas que han publicado al menos mil artículos en las últimas décadas. El promedio de citas por artículo en LNCS es de sólo 0,96 siendo el promedio de 9,2 para los artículos publicados en el primer tercio del ranking de revistas. En este ranking, el LNCS concentra el 55% del total de artículos, es decir, es un medio muy masivo de publicación respecto de las revistas. Estos resultados son una clara evidencia de que existen conferencias de muy buen nivel pero también de muy bajo nivel que son publicados por LNCS y presumiblemente también por IEEE-CS y ACM.

También se observa en CiteSeer^X que en muchos casos los factores de impacto medidos según la métrica ISI alcanzados por las conferencias pueden ser muy superiores a los de las revistas. Lo mismo indica una muestra tomada desde DBLP y otra con el 100% de los artículos indexados por el Portal ACM Digital Library. Además las tasas de aceptación de artículos para algunas conferencias pueden ser bastante exigentes. Por ejemplo, el año 2008 el CIKM aceptó un 17% de más de 750 artículos enviados a esta conferencia. Esto contrasta con la situación en otras disciplinas de la Ciencia y la Ingeniería, o incluso en áreas más teóricas de Ciencia de la Computación, donde las conferencias tienen una importancia menor y los trabajos son considerados simples presentaciones o ponencias orales y, por lo tanto, constituyen una publicación de calidad inferior respecto de los artículos en revistas, o simplemente no son considerados publicaciones y no cuentan a la hora de evaluar los antecedentes de los investigadores de la disciplina.

Estas dos realidades resultan conflictivas al momento de emitir opiniones sobre la calidad de un investigador de Ciencia de la Computación dado que los juicios pueden depender fuertemente del tipo de área (o incluso disciplina) de la cual provienen los evaluadores. Pensamos que el estudio presentado en este artículo puede ser útil para argumentar la necesidad de diseñar una forma de evaluación que caracterice de mejor manera la naturaleza de la disciplina y, por ende, la calidad de la contribución científica de sus investigadores en el contexto nacional.

El sistema nacional actualmente utiliza las publicaciones en revistas indexadas en ISI como estándar de referencia para medir calidad de manera transversal a todas las disciplinas del País. Actualmente los artículos publicados por LNCS, IEEE-CS y ACM no son considerados como artículos válidos para efectos de evaluación de investigadores porque no están clasificados como artículos de revista ISI en el Web of Science. Cabe hacer notar que a partir del 20 de Octubre del 2008, el Web

of Science ha integrado en su sistema normal de búsquedas las versiones actualizadas de colecciones que ellos llaman "ISI Proceedings", colecciones que contienen artículos de conferencias que, para al menos Ciencia de la Computación, son de una gran diversidad en términos de nivel de exigencia en tasas de aceptación, cantidad de artículos enviados al evento y calidad del comité de programa. Para tener acceso a estas colecciones es necesario pagar una suscripción adicional.

Otros países se han preocupado del tema y actualmente existen al menos dos fuentes confiables de ranking de conferencias que son utilizados para evaluación de la contribución científica de investigadores en Ciencia de la computación: CiteSeer^X (citeseerx.ist.psu.edu) y CORE (www.core.edu.au). Por ejemplo, actualmente el sistema Español considera los artículos de conferencia indexados por estos dos sistemas de indexación de literatura técnica para evaluar a sus investigadores de Ciencia de la Computación (Sexenios). Cabe hacer notar que en la normativa oficial Española existe un apartado especial para Ciencia de la Computación la cual reconoce sus características especiales y la distingue de otras disciplinas tales como Matemática o Física. Por otra parte, en Brasil se considera desde hace ya varios años los artículos en buenas conferencias para establecer un ranking de investigadores nacionales. En varios otros países tales como USA y Canadá, el sistema ISI no es impuesto como estándar transversal a todas las disciplinas. La evaluación de investigadores y proyectos es realizada estrictamente por pares del área y de acuerdo a la naturaleza de cada área, lo cual, como es lógico, es definido por los investigadores de la propia disciplina.

Al ser el artículo de revista indexada en ISI la única métrica de evaluación de calidad imperante en Chile para todas las disciplinas de la Ciencia y la Ingeniería, se puede producir una evaluación sesgada o incompleta del trabajo de los investigadores en ciertas áreas de Ciencia de la Computación, que por sus características más tecnológicas, son más dependientes de las publicaciones en conferencias para difundir sus resultados en el momento oportuno. El tiempo de evaluación de artículos enviados a revistas puede ser de uno o dos años mientras que la escala de tiempo en conferencias es de meses, lo cual ciertamente puede ser crítico al considerar aspectos tales como la difusión oportuna de resultados conducentes a patentes en áreas de trabajo que son más tecnológicas que teóricas. Por otra parte, no todas las revistas ampliamente reconocidas en la disciplina están indexadas en el ISI Web of Science.

Los resultados presentados en las siguientes secciones de este artículo constituyen una evidencia clara de que los artículos de conferencia son un medio muy utilizado para difundir resultados relevantes de investigación en Ciencia de la Computación. Los resultados también explican los bajos factores de impacto ISI que reciben las revistas de la disciplina con respecto a otras disciplinas de la Ciencia e Ingeniería. Simplemente en Ciencia de la Computación los artículos más citados por sus investigadores se distribuyen en un medio adicional de publicación, es decir, las buenas conferencias, las cuales, como mostramos en este artículo, pueden alcanzar factores de impacto ISI superiores a los de las mejores revistas de la disciplina.

Donde Publican los Investigadores más Productivos

Elaboramos una lista de los investigadores en Ciencias de Computación más reconocidos a nivel internacional. Para esto obtuvimos los autores con índice-H mayor o igual a 40, los cuales representan a los autores más citados por la comunidad (http://www.cs.ucla.edu/∼palsberg/h-number.html). El índice-H es definido como el número de artículos de un autor dado, los cuales tienen un número de citas mayor o igual a H. Este indicador es más exigente que el total acumulado de citas y es considerado un índice que refleja bien la productividad de un investigador.

	Promedio	Desv. Std
Total de publicaciones por investigador	137	66
Total artículos en conferencias por investigador	88	45
Total artículos en revistas por investigador	51	31

Table 1: Publicaciones de los investigadores más productivos en Ciencia de la Computación.

Utilizando la lista de los primeros 100 investigadores con índice-H mayor o igual a 40, y la información de sus publicaciones extraída desde la base de datos del sistema DBLP, se determinó la relación entre el número de publicaciones en conferencias versus el número de publicaciones en revistas de estos investigadores. La Tabla 1 muestra los resultados. Como anexo se adjunta los datos completos con la identificación de los top 100 investigadores a julio del 2008. Los promedios muestran que las publicaciones en conferencias representan un porcentaje dominante en el currículum de los top-100 investigadores. No obstante, los valores de la desviación estándar muestran que una parte significativa de ellos está en los dos extremos (ver anexo).

Citas desde Publicaciones Indexadas en ISI

De la lista de referencias bibliográficas de revistas indexadas por el Web of Science y asociadas a Ciencia de la Computación y Sistemas de Información, se determinó el número de citas que hacen estas publicaciones a otras publicaciones ISI y no ISI. Toda cita que hace referencia a una publicación que pudiera ser accedida por el sistema Web of Science fue considerada como ISI. Es sencillo detectar por software este tipo de artículos puesto que en la descripción de cada referencia bibliográfica se incluye un enlace al Web of Science.

Por el contrario, toda publicación que no pudiera ser accedida (sin enlace) fue clasificada como no-ISI. Dada la imposibilidad de detectar automáticamente por software de manera exacta las publicaciones de conferencias (existen miles de conferencias, cada una con su propia secuencia de caracteres que la identifica), la cuenta de publicaciones no-ISI incluye también las citas a otras tales como libros, tesis, reportes técnicos o revistas que no están indexadas en ISI. No obstante, es sabido que habitualmente en la lista de referencias de los artículos en Ciencia de la Computación, las citas a otras publicaciones representan un porcentaje menor de la lista de referencias bibliográficas. Por ejemplo, en el sistema DBLP comprobamos de que las publicaciones indexadas por este sistema que no son revistas ni conferencias representan menos del 10% del total.

El procedimiento empleado fue el siguiente. Se tomó como semilla la lista de revistas del área de Ciencia de la Computación y Sistemas de Información que el Web of Science declaraba como indexadas en ISI en Mayo del 2008. Por cada una de esas revistas utilizamos un script en PHP que envió consultas al buscador del Web of Science para recuperar todos los artículos de cada revista indexada, y por cada revista se obtuvo su lista de referencias bibliográficas. Luego aplicamos scripts Unix sobre estos resultados para formar una base de datos relacional para luego poder contar el número de publicaciones de cada tipo.

Las citas presentes en las listas de referencias de cada artículo nos permitió encontrar la cantidad de citas a publicaciones que no están indexadas en ISI (es decir, no tienen un enlace a un documento que entrega los detalles del artículo en el Web of Science), esto tanto en número acumulado de citas como cantidad de artículos distintos que son citados.

Muestra de artículos		Citas recibidas desde el año de publicación			Citas totales hechas por la muestra de cada año		
Año	**A**	**B**	**C**	**C/B**	**D**	**E**	**E/D**
2006	14.394	7.370	12.314	1,67	121.830	248.121	2,03
2005	13.058	19.928	33.463	1,68	100.520	222.926	2,22
2004	12.412	31.748	57.508	1,81	92.768	206.513	2,23
2003	11.923	43.357	76.481	1,76	84.300	196.934	2,33
2002	10.822	50.834	90.366	1,78	72.429	176.578	2,44
2001	10.099	54.557	99.175	1,82	65.258	165.516	2,54
2000	9.660	63.179	106.277	1,68	58.013	156.497	2,70
Total	82.368	270.973	475.584	1,74	595.118	1.373.085	2,36

Table 2: El tipo de publicaciones que figuran en las listas de referencias bibliográficas de los artículos de revistas ISI indexados en el Web of Science (Mayo 2008). La columna **A** es el total de artículos por año de publicación desde los cuales se obtuvo las listas de referencias bibliográficas. La columna **B** es el total acumulado de citas que reciben los artículos ISI a partir de su año de publicación. La columna **C** es la misma métrica pero contando los artículos que no están indexados en el Web of Science, es decir, artículos que no son considerados ISI. Las columnas **D** y **E** representan la cantidad de citas realizadas hacia artículos ISI y no-ISI, respectivamente, que fueron encontradas en las listas de referencias bibliográficas de los artículos de la columna **A**.

Los resultados se muestran en la Tabla 2. En promedio, más del doble de las citas realizadas por los artículos indexados en ISI son hacia documentos que no son ISI. Como la cantidad de artículos distintos de ambos tipos es similar, entonces los artículos no-ISI son más citados que los ISI.

Las 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. 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 indexado en el Portal, el cual indexa tanto artículos de conferencias como artículos de revistas indicando el tipo de publicación, el cual está claramente diferenciado en el contenido del respectivo enlace HTML. Con esto podemos calcular de manera exacta la proporción entre artículos de conferencias y de revistas citados en las

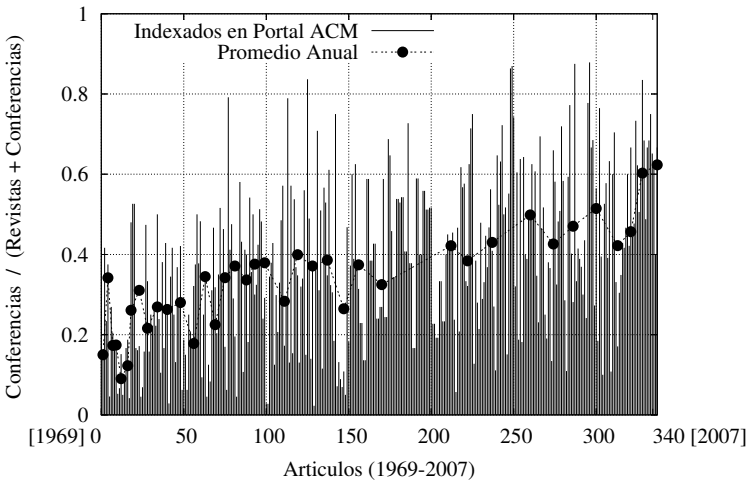

Figure 1: Proporción de artículos de conferencias citados en cada artículo del ACM Computing Surveys, los cuales están indexados por el portal ACM Digital Library.

---

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 artículos de 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 un 30%).

Factores de Impacto ISI 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 1 millón de artículos y registra más de 22 millones de citas a los artículos indexados.

En la Figura 2, mostramos los factores de impacto ISI calculados por CiteSeer^X, los cuales 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. En los archivos HTML se puede detectar sin error cuando se trata de una revista o una conferencia. Los resultados muestran que las conferencias tienen factores de impacto

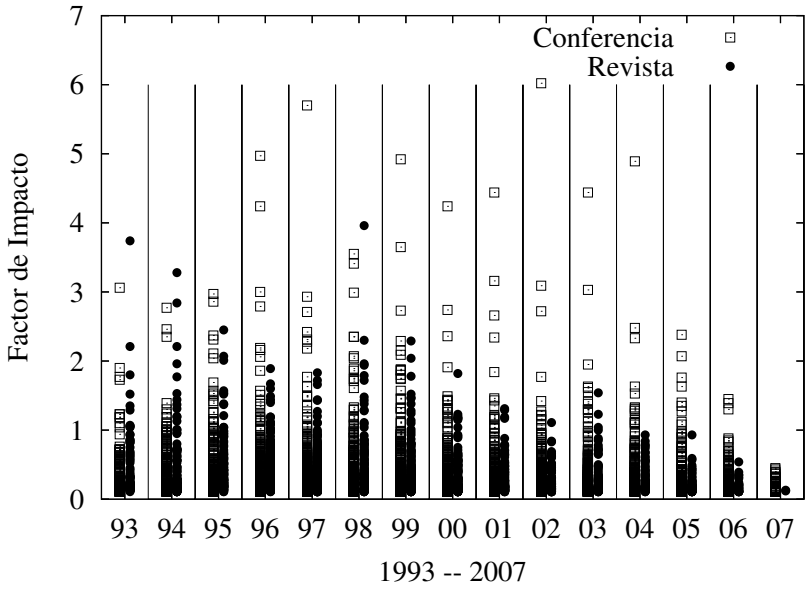

Figure 2: Factores de impacto tipo ISI para las conferencias y revistas indexadas por CiteSeer^X.

superiores a los de las revistas cuando no se hace diferencia entre áreas específicas.

En la tabla 3 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 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. Estos resultados muestran la misma tendencia general de la Figura 2.

En la Figura 3 se muestra la proporción de revistas versus conferencias dentro de los 100 artículos más citados en la disciplina. La curva etiquetada con un círculo blanco muestra la división entre el total de artículos de revistas que figuran dentro de los 100 artículos más citados en cada año y el total de artículos de conferencias que también están dentro de los 100 artículos más citados. La curva etiquetada con un triángulo negro es la división entre la suma acumulada de citas que reciben los artículos de revistas ubicados dentro los 100 más citados y la suma acumulada que reciben los artículos de conferencia también dentro de esos 100 más citados. La relevancia de las revistas sobre las conferencias no alcanzan al doble según estas métricas.

La curva etiquetada con un cuadro blanco en la Figura 3, muestra la división entre el total de revistas que figuran dentro de los 100 medios de publicación con mayor factor de impacto en cada año y el total de conferencias que también figuran dentro de los 100 medios de mayor impacto. Estos resultados muestran que en el CiteSeer^X la cantidad de conferencias con factores de impacto relevantes es más del doble que las revistas. La curva con el círculo negro muestra la proporción total de medios de publicación de artículos entre revistas y conferencias. Estos resultados muestran que el total de artículos de conferencias que se publican cada año es muy superior al total de artículos que publican las revistas. Dentro de este grupo de élite con los 100 medios de publicación mejor rankeados en CiteSeer^X, los artículos de revistas son de mejor calidad, pero los resultados también muestran de que existen conferencias capaces de aportar artículos de calidad comparable.

Considerando los totales sobre todos los años, es decir, los artículos más citados en el periodo que va desde 1990 a 2008 (citeseerx.ist.psu.edu/stats/articles), la cantidad de artículos que están

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

(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)

Table 3: 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.

dentro de los 100 más citados corresponden a 54 artículos de revistas y a 32 artículos de conferencias, es decir, la proporción está dada por $54/32 = 1{,}7$ donde las 14 referencias restantes corresponden a libros y manuales de software, y referencias indeterminadas (probablemente reportes técnicos). Respecto a la suma acumulada de citas de cada tipo de publicación abarcando el periodo 1990-2008, encontramos que las revistas suman un total de 60.683 citas mientras que las de conferencias acumulan un total de 29.170 citas. Entonces pensamos que no es arriesgado decir que los artículos de conferencias son responsables de al menos 1/3 del estado del arte en Ciencia de la Computación dado que estamos considerando un grupo de elite de sólo los 100 artículos más citados.

En la misma línea, si ahora consideramos las 10 mil publicaciones más citadas observamos $5091/4909 = 1{,}04$ para la proporción total de revistas versus conferencias, y la razón de la suma acumulada de cantidad de citas alcanza el valor 1,20. Es decir, en el rango de las 10 mil publicaciones más citadas el aporte de las revistas y conferencias es casi idéntico. Esto indica que los resultados de la Figure 3 son muy conservadores puesto nos pusimos como exigencia los 100 más citados en toda la colección mantenida por el sistema CiteSeer^X.

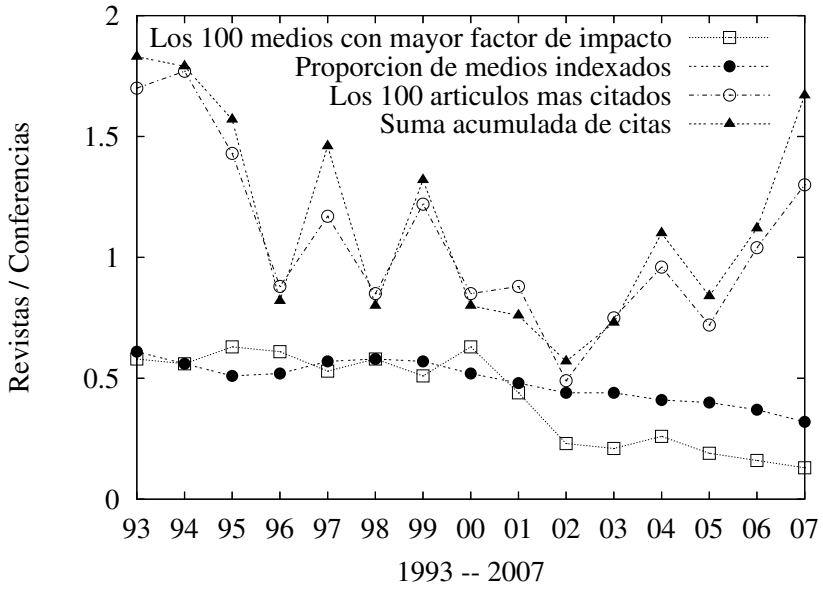

Figure 3: Proporción entre revistas y conferencias considerando el ranking de los mejores 100.

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. 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 CiteSeer^X otra sección de este sistema, la cual presenta los 10 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 CiteSeer^X.

Acumulando el total de citas de los 10 mil artículos más citados en las respectivas revistas y conferencias donde fueron publicados, podemos establecer un ranking de los medios de publicación que concitan el mayor número de citas según CiteSeer^X. En la Figura 4 se muestran los resultados, los cuales provienen desde poco más de 200 revistas distintas y casi 400 conferencias distintas. 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 10 mil artículos más citados en CiteSeer^X. 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 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 5 muestra el factor de impacto C/P que los 10 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 5 es el mismo que el orden dado en la Figura 4.

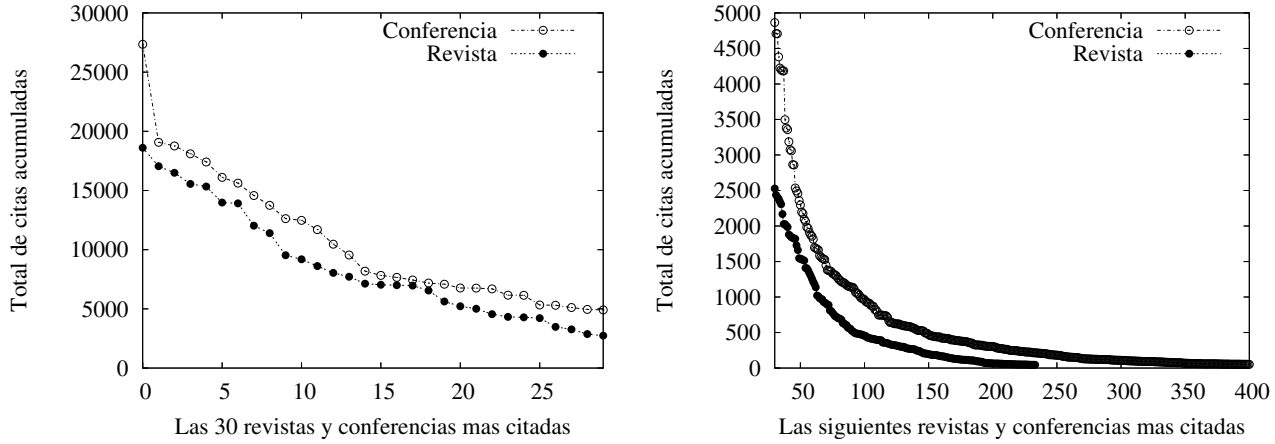

Figure 4: Total acumulado de citas por conferencia y revista utilizando CiteSeer y DBLP.

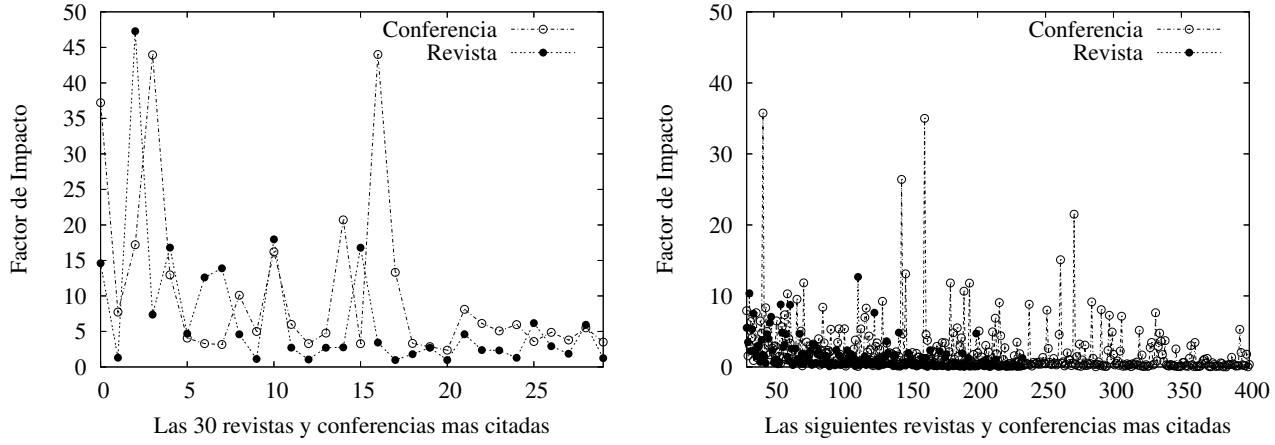

Figure 5: Número promedio de citas por artículo de conferencia/revista en CiteSeer y DBLP.

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 100 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 citas mientras que el total de citas a los artículos de revistas es de 854 mil citas. Nuestros scripts también detectaron un total de 2.428 conferencias distintas 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 6 y 7 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.

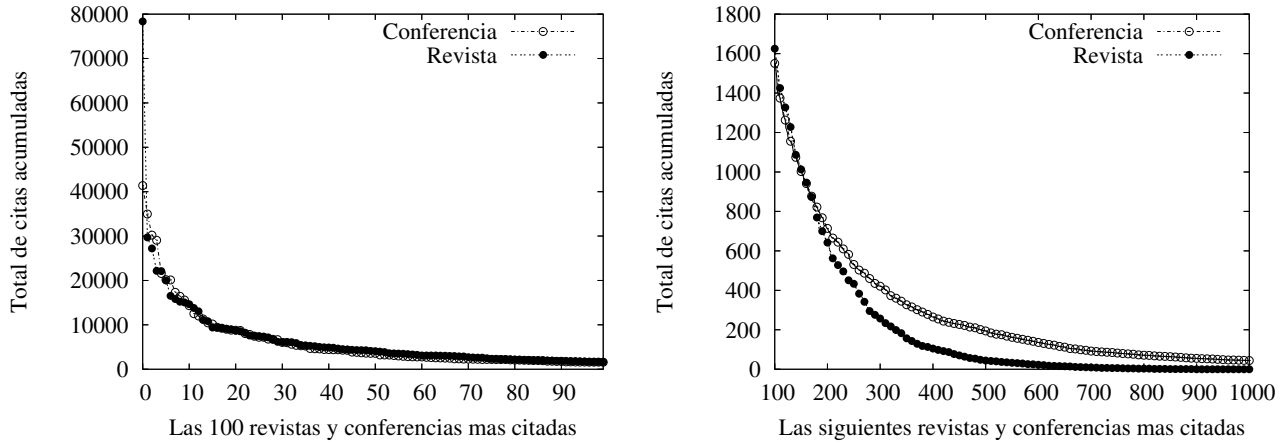

Figure 6: Total acumulado de citas por conferencia y revista utilizando los artículos indexados en el Portal ACM Digital Library.

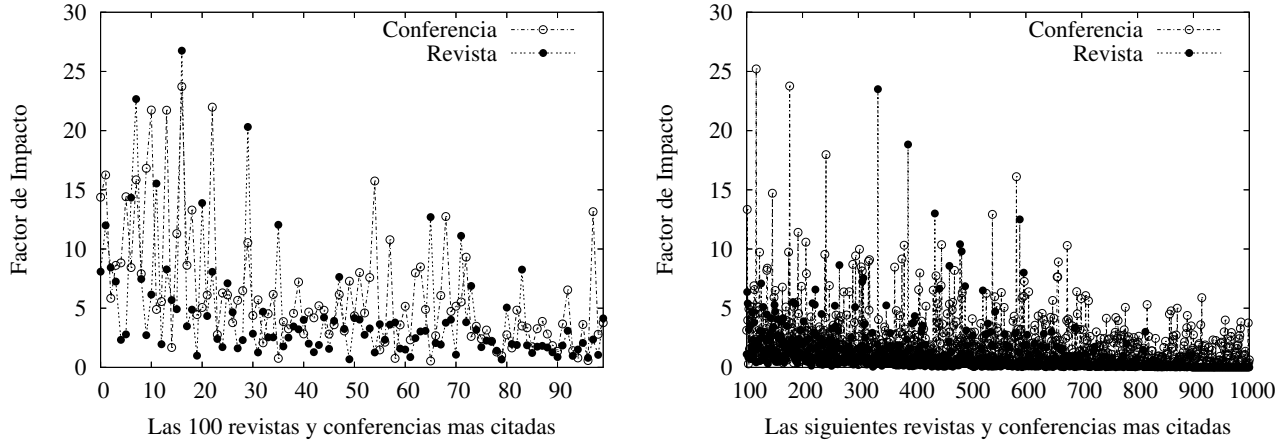

Figure 7: Citas promedio por conferencia/revista según en el Portal ACM Digital Library.

Búsquedas en el Sistema "DBLP"

En el XML que bajamos desde DBLP (http://dblp.uni-trier.de/xml) en Septiembre del 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 las ediciones anuales de 22 conferencias y 8 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 a artículos publicados en poco más de 150 revistas y 300 conferencias. En el XML de cada cita se indica claramente cuando se trata de una cita a un artículo de revista o a uno de conferencia y, por lo tanto, no hay margen de error en el conteo de citas que recibe cada tipo de artículo. También el nombre de la conferencia y revista puede ser detectado fácilmente por nuestros scripts puesto que estos datos están bien formateados con XML.

En la Figura 8 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 las ediciones anuales de las 22 conferencias y 8 revistas. En la Figura 9 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.

En la tabla 4 mostramos nuestro factor de impacto (C/P) para las 5 conferencias y revistas que figuran como las que acumulan la mayor cantidad de citas en la figura 8. El factor de impacto lo hemos calculado de la misma manera que las secciones anteriores, es decir como C/P donde C

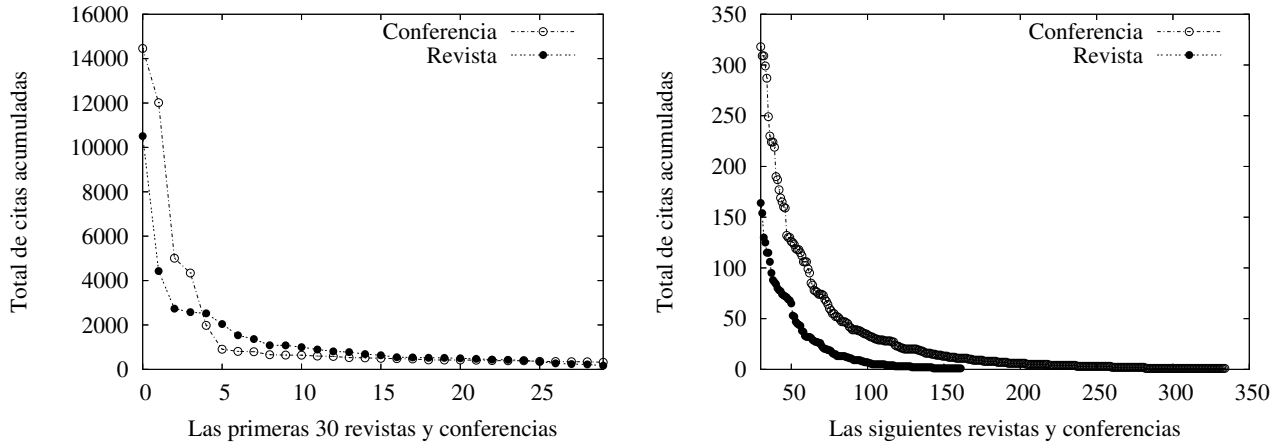

Figure 8: Total de citas acumuladas por artículos de cada conferencia y revista en DBLP.

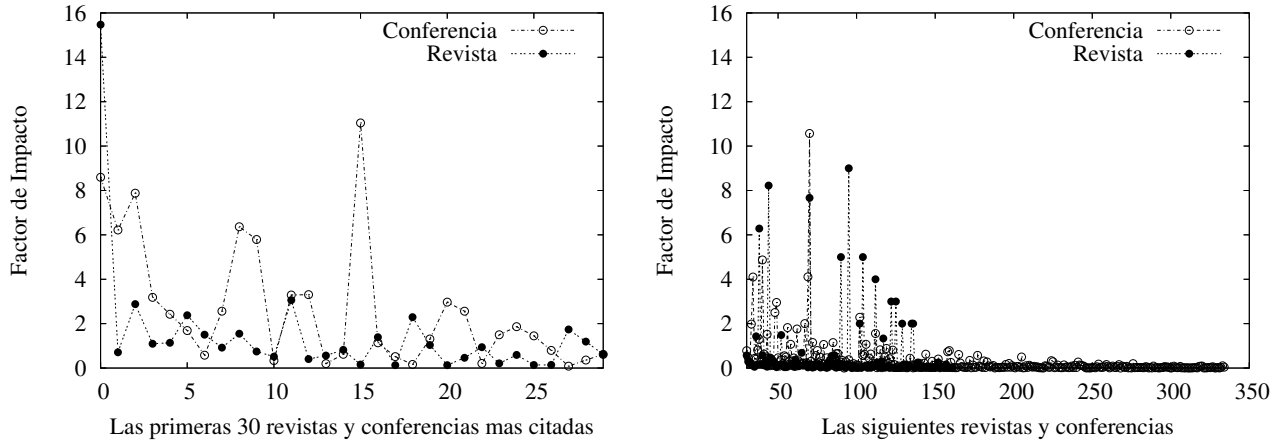

Figure 9: Factores de Impacto de cada conferencia y revista en DBLP.

son los valores de la columna "T. Citas" (i.e., la suma total de citas que reciben los artículos de la conferencia o revista) y P el total de artículos publicados por la respectiva conferencia o revista. Los valores de P están dados por la columna etiquetada con "M*A" donde "M" es el promedio anual de artículos publicados por la conferencia o revista (valor tomado desde el sistema "Faceted DBLP", opciones "Most prolific journals/conferences") y "A" es número de años que abarcan las citas a los artículos de cada conferencia o revista. Consideramos como año válido sólo aquellos en que el total de citas superó el valor 10. Esto se hizo para evitar contar un año adicional en revistas con muy pocas citas en ese año particular (en las conferencias, en cada año siempre se registraron más de 10 citas). Los resultados muestran que para esta área las top-5 conferencias superan en valor de factor de impacto a varias de las revistas más conocidas en la disciplina.

Es interesante observar si los artículos de conferencias tienden a citar más a las conferencias y viceversa respecto de las revistas, y lo que sucede con los libros. La distribución de citas a artículos de revistas y conferencias, y citas a libros la representaremos por la tupla (*conf*, *rev*, *lib*) donde *conf*

Conferencia	T. Citas	P*A	F. Impacto	Revista	T. Citas	P*A	F. Impacto
SIGMOD	14.459	64*29	7,79	TODS	10.507	21*22	20,74
VLDB	12.016	74*24	6,76	CACM	4.431	168*32	0,82
PODS	5.011	32*18	8,69	CSUR	2.735	28*26	3,75
ICDE	4.342	98*15	2,95	TSE	2.579	84*20	1,53
ER	1.983	36*17	3,24	JACM	2.526	45*30	1,87

Table 4: Factores de Impacto promedio. (Nota: el factor de impacto estimado para la revista VLDB Journal es 547/(23*7) = 3,39).

es el total de citas a artículos de conferencias, *rev* total de citas a artículos de revista, *lib* el total de citas a libros. Las tuplas considerando individualmente las listas de referencias de los artículos de conferencia, de revista y de libros son las siguientes:

```
Citas desde CONFERENCIAS ( conf= 44.037 (56%) rev= 29.092 (37%) lib= 5.514 (7%) )
Citas desde REVISTAS     ( conf= 14.562 (49%) rev= 13.238 (44%) lib= 2.181 (7%) )
Citas desde LIBROS       ( conf=  1.583 (52%) rev=  1.238 (40%) lib=   249 (8%) )
```

Los resultados muestran que no existen variaciones relevantes en la tendencia general ya comentada.

Dada la gran cantidad de información no pudimos calcular manualmente el denominador "P*A" (Tabla 4) para los resultados presentados en la Figura 9. Para ello simplemente detectamos para cada conferencia y revista el año más reciente en que fue citada, y desde la base de datos de DBLP obtuvimos el total de artículos publicados por la conferencia o revista hasta ese año. Este valor corresponde al denominador en la fórmula y el numerador es el respectivo número total de citas de la Figura 8.

Otro dato de interés es comparar el nivel de producción de artículos de conferencias y revistas. El archivo XML que bajamos desde DBLP contiene artículos hasta comienzos del 2008. De estos artículos un total de 668 mil pertenecen a conferencias y 408 mil a revistas, es decir, los artículos de conferencia representan el 62% de la colección. Los artículos provienen de 2.487 conferencias distintas y de 655 revistas distintas. El total acumulado de las ediciones anuales de conferencias y revistas es de 11.819 y 7.684 respectivamente, es decir, una relación del tipo 60%-40%.

En la Figura 10 mostramos el promedio anual de artículos publicados por las 500 conferencias y revistas consideradas como las más prolíficas en el sistema "Faceted DBLP" (http://dblp.l3s.de), opciones "most prolific journals/conferences". Esos resultados muestran que tanto las conferencias como las revistas publican tanto o más artículos que uno del otro tipo. El promedio para las conferencias es de 94 artículos por año mientras que para las revistas es de 59 artículos, lo que también entrega una relación 60%-40% para las 500 conferencias y revistas. Note que al ver los promedios de la Tabla 1 encontramos que también existe una relación 60%-40% (es decir, 88/(88+51) versus 51/(88+51)). Los valores en sí mismos pueden ser sólo una consecuencia del hecho de que las publicaciones de cada investigador se obtuvieron del propio DBLP. Lo interesante es que esta proporción se mantiene incluso entre los 100 investigadores de mayor índice H.

El sistema "Faceted DBLP" (http://dblp.l3s.de) utiliza la base de datos de DBLP para presentar distintas vistas de las referencias bibliográficas. En particular entrega información del total de artículos de revistas y conferencias, y los autores con más artículos. Realizamos búsquedas uti-

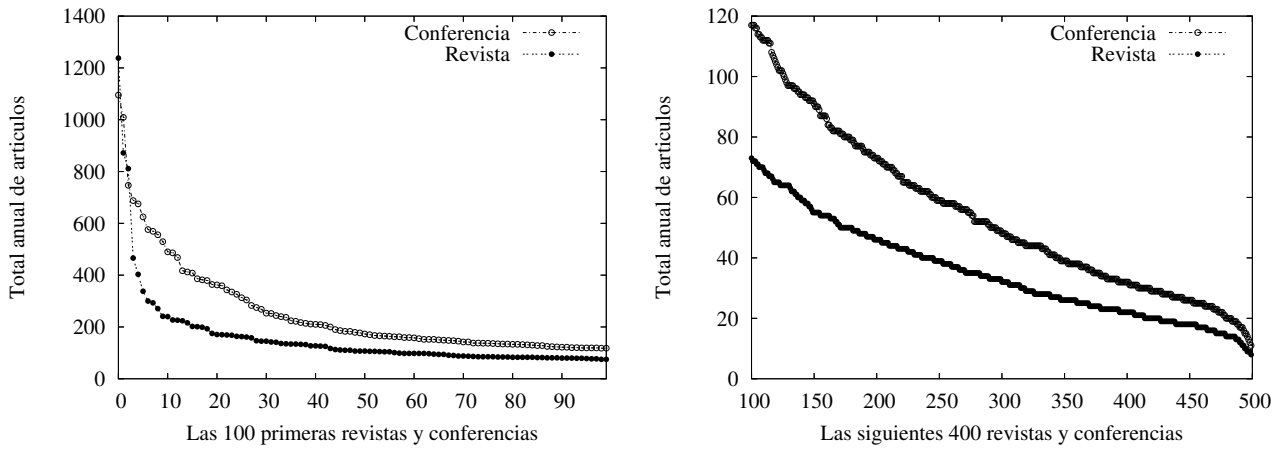

Figure 10: Total anual promedio de artículos publicados por las revistas y conferencias con mayor número de artículos en DBLP (periodo 1953-2008).

lizando distintas palabras clave que identifican áreas de Ciencia de la Computación bajo la opción "Venues only". Por cada resultado de las búsquedas anotamos el total de artículos en revistas y conferencias, y también anotamos la cantidad de artículos de revistas y conferencias para sólo los tres autores con más publicaciones en el área. En la Tabla 5 se muestran los resultados que obtuvimos en Noviembre del 2008, los cuales son interesantes puesto que muestran las diferencias para distintas áreas de Ciencia de la Computación respecto de la proporción entre publicaciones en revistas y conferencias. Los totales globales indican que en todos los casos los autores en general publican mucho más en conferencias que en revistas. Sin embargo, para los tres primeros autores que registran el mayor número de publicaciones en cada área, se observan diferencias significativas respecto de la proporción revistas versus conferencias. Esto es consecuente con lo reportado en la Tabla 1 y los detalles entregados en el Anexo de este artículo.

Validación de Citas "Web of Science" utilizando "DBLP"

El gran problema con el Web of Science es que las citas a artículos de conferencias han sido escritas de manera muy poco prolija. Al contrario de las revistas, prácticamente no existe una notación estándar para identificar a cada conferencia. Algo similar se observa en la colección "ISI Proceedings" en la cual no es difícil ver que las ediciones anuales de una misma conferencia figuran con notaciones distintas lo cual dificulta las búsquedas de artículos. Por ejemplo si en el Web of Science uno realiza una búsqueda inicializando el campo "Conference" con el valor "VLDB" y luego ejecuta la opción "Create Citation Report" puede observar en el gráfico de "Published Items in Each Year" que hay años en que se reporta que no se publicó ningún artículo, lo cual no es efectivo. Una búsqueda por el título de artículos que se sabe fueron publicados en uno de esos años revela que la conferencia figura con variantes en el nombre que la identifica. En otros casos simplemente no existen artículos de un cierto año ingresados en la base de datos. Otro problema es que en varios casos no se hace distinción entre los artículos de la conferencia propiamente tal y los workshops satélite que se realizan junto con la conferencia. Esto hace que en ciertos años exista una cantidad exageradamente grande de artículos publicados en la conferencia respecto de otros años. Pareciera

Palabra(s) de Búsqueda		Revistas	Conferencias
Algorithms	Totales	3.343	9.780
	1^{er} autor	159	67
	2^{do} autor	252	106
	3^{er} autor	114	147
Artificial Intelligence	Totales	2.185	22.226
	1^{er} autor	20	94
	2^{do} autor	41	74
	3^{er} autor	38	97
Computer Graphics	Totales	716	3.608
	1^{er} autor	135	226
	2^{do} autor	42	0
	3^{er} autor	41	52
Dabatases	Totales	1.172	9.076
	1^{er}autor	29	118
	2^{do} autor	59	125
	3^{er} autor	61	146
Cybernetics	Totales	1.368	3.783
	1^{er} autor	30	24
	2^{do} autor	31	4
	3^{er} autor	58	16
Distributed	Totales	1.302	21.823
	1^{er} autor	87	190
	2^{do} autor	55	137
	3^{er} autor	51	193
Information Systems	Totales	546	11.510
	1^{er} autor	56	256
	2^{do} autor	58	139
	3^{er} autor	26	110
Logic	Totales	3.162	9.025
	1^{er} autor	224	9
	2^{do} autor	33	131
	3^{er} autor	88	50
Parallel	Totales	7.756	25.828
	1^{er} autor	51	193
	2^{do} autor	103	164
	3^{er} autor	33	177
Software Engineering	Totales	1.673	11.770
	1^{er} autor	105	78
	2^{do} autor	67	105
	3^{er} autor	32	58

Table 5: Búsquedas en el Faceted-DBLP

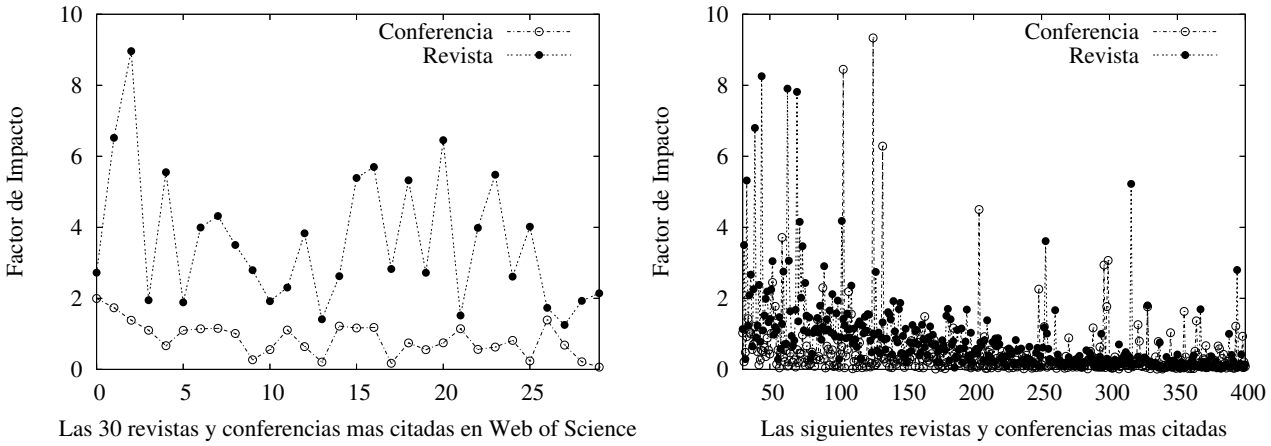

Figure 11: Factores de Impacto de conferencias y revistas utilizando Web of Science y DBLP.

ser que el enfoque es orientado al libro del proceedings particular más que un enfoque orientado a mantener la consistencia a nivel de ediciones anuales. Esto hace extremadamente complicado realizar una comparación entre la relevancia de los artículos de conferencias y revistas.

Aún en este contexto de información imperfecta decidimos realizar cálculos similares a los presentados en la Figura 5, es decir, tomamos los títulos de artículos detectados en las listas de referencias de los artículos de revista encontrados en el Web of Science, y utilizamos esos títulos como llave para obtener la revista o conferencia en la base de datos de DBLP (dblp.xml).

Los resultados se presentan en la Figura 11, la cual muestra una situación desmejorada para el impacto de los artículos de conferencia ubicados dentro de los 30 primeros lugares respecto de cantidad acumulada de citas por medio de publicación (gráfico lado izquierdo). La razón es que ahora nuestros scripts detectan mucho menos citas a conferencias que a revistas. Sin embargo para los siguientes lugares en el ranking de citas acumuladas (gráfico lado derecho) los factores de impacto son similares en conferencias y revistas. El total global acumulado de citas a todos los artículos de revista es de 488.112, mientras que para conferencias este valor es sólo de 86.753, es decir un 15%, lo cual es no es cierto por la siguiente razón. La lista de citas obtenidas en el Web of Science a los que les pudimos detectar los títulos con nuestros scripts contiene 1,46 millones de artículos, mientras que el total de citas detectadas es de 3,9 millones, es decir, más del doble, y estas citas sin título necesariamente corresponden a conferencias y algunas revistas no indexadas en ISI. Esto porque para las revistas ISI es sencillo detectar automáticamente el título debido al formato especial con que son presentadas al usuario en el Web of Science. No obstante, los resultados de la Figura 11 siguen siendo útiles para mostrar que ciertas conferencias pueden tener tanto impacto como los artículos de revistas indexadas en ISI.

Agradecimientos

Las revistas y conferencias mencionadas en la Tabla 3 fueron seleccionadas por John Atkinson de la Universidad de Concepción, y Pablo Barceló y Eric Tanter de la Universidad de Chile. La Sociedad Chilena de Ciencia de la Computación ha financiado parcialmente este estudio.

ANEXO: Lista Alfabética de Investigadores con mejor índice H

Investigador	Revistas	Conferencias	Total
Alberto L. Sangiovanni-Vincentelli	96	261	357
Alex Pentland	52	101	153
Amir Pnueli	58	167	225
Amit P. Sheth	46	63	109
Andrew S. Tanenbaum	47	45	92
Andrew Zisserman	32	113	145
Anil K. Jain	146	149	295
Barbara Liskov	26	64	90
Barry W. Boehm	53	78	131
Ben Shneiderman	108	106	214
Carl Kesselman	23	60	83
Christos Faloutsos	45	93	138
Christos H. Papadimitriou	128	130	258
Craig Chambers	18	51	69
David A. Patterson	47	59	106
David E. Culler	23	90	113
David E. Goldberg	32	117	149
David Haussler	44	55	99
David J. DeWitt	42	51	93
David L. Dill	25	105	130
David S. Johnson	66	31	97
Deborah Estrin	40	85	125
Demetri Terzopoulos	26	49	75
Didier Dubois	86	126	212
Donald E. Knuth	77	10	87
Donald F. Towsley	101	160	261
Douglas C. Schmidt	42	108	150
Edmund M. Clarke	61	126	187
Geoffrey E. Hinton	27	59	86
George Karypis	27	70	97
H. V. Jagadish	31	65	96
Hari Balakrishnan	22	67	89
Hector Garcia-Molina	73	153	226
Henry M. Levy	27	65	92
Herbert A. Simon	37	17	54
Herbert Edelsbrunner	109	88	197
Ian T. Foster	69	133	202
Jack Dongarra	77	132	209
James F. Kurose	39	99	138
Jason Cong	47	139	186
Jeffrey D. Ullman	94	85	179
Jeffrey Scott Vitter	77	107	184
Jennifer Widom	43	63	106
Jiawei Han	59	158	217

Investigador	Revistas	Conferencias	Total
John A. Stankovic	62	93	155
John McCarthy	22	25	47
John Mitchell	41	88	129
Jon M. Kleinberg	33	69	102
Jose Joaquin Garcia-Luna-Aceves	32	94	126
Jose Meseguer	63	95	158
Joseph A. Goguen	41	69	110
Judea Pearl	51	79	130
Kai Li	20	72	92
Ken Kennedy	48	97	145
Krithi Ramamritham	68	114	182
Leonard Kleinrock	41	26	67
Leslie Lamport	60	50	110
Lixia Zhang	25	50	75
Luca Cardelli	38	64	102
M. Frans Kaashoek	28	60	88
Maja J. Mataric	26	49	75
Mario Gerla	64	129	193
Martin Abadi	52	104	156
Martin Vetterli	45	63	108
Micha Sharir	194	162	356
Michael A. Arbib	48	20	68
Michael I. Jordan	33	89	122
Michael J. Carey	30	54	84
Michael J. Franklin	40	52	92
Michael Stonebraker	57	73	130
Mihalis Yannakakis	73	88	161
Mihir Bellare	20	87	107
Miron Livny	22	68	90
Monica S. Lam	13	78	91
Moshe Y. Vardi	87	187	274
Nancy A. Lynch	72	124	196
Nicholas R. Jennings	64	116	180
Oded Goldreich	128	107	235
Olivier D. Faugeras	36	77	113
Pat Hanrahan	14	59	73
Philip S. Yu	119	228	347
Philip Wadler	24	47	71
Prabhakar Raghavan	55	82	137
Raghu Ramakrishnan	32	67	99
Rajeev Alur	44	106	150
Rajeev Motwani	59	81	140
Rakesh Agrawal	32	66	98
Ramesh Govindan	29	53	82
Randy H. Katz	39	100	139

Investigador	**Revistas**	**Conferencias**	**Total**
Richard J. Lipton	64	92	156
Richard M. Karp	68	94	162
Robert Endre Tarjan	137	80	217
Robin Milner	24	54	78
Ronald Fagin	59	48	107
Ronald L. Rivest	48	79	127
Sally Floyd	19	17	36
Saul Greenberg	21	54	75
Scott Shenker	33	104	137
Sebastian Thrun	38	110	148
Serge Abiteboul	49	81	130
Simon L. Peyton Jones	53	67	120
Stanley Osher	9	15	24
Stefano Ceri	53	72	125
Steven Salzberg	27	18	45
Sushil Jajodia	76	142	218
Takeo Kanade	60	131	191
Teuvo Kohonen	14	8	22
Thomas A. Henzinger	33	154	187
Thomas S. Huang	88	231	319
Thomas W. Reps	31	82	113
Timothy W. Finin	29	71	100
Tomaso Poggio	28	45	73
Vern Paxson	14	28	42
Victor R. Basili	82	60	142
Victor R. Lesser	39	120	159
Vipin Kumar	39	102	141
W. Bruce Croft	49	93	142
Willy Zwaenepoel	20	69	89
Won Kim	83	23	106