Information Services & Use 28 (2008) 127–128
DOI 10.3233/ISU-2008-0562
IOS Press

Comparing the scientific impact of conference and journal publications in computer science

Erhard Rahm
University of Leipzig, Leipzig, Germany
E-mail: rahm@informatik.uni-leipzig.de

The impact of scientific publications is often estimated by the number of citations they receive, i.e. how frequently they are referenced by other publications. Since publications have associated authors, originating institutions and publication venues (e.g. journals, conference proceedings) citations have also been used to compare their scientific impact. For instance, one commonly considered indicator of the quality of a journal is its *impact factor* [1]. The impact factors are published yearly by Thomson ISI in the Journal Citation Report (JCR) by counting the citations from articles of thousands of journals.

However, research results in computer science are often published in high-quality conferences which are not covered by the JCR citation databases [2]. Other commercial citation data sources such as Elsevier Scopus also focus on journals and contain comparatively few conference publications. Hence these data sources cover only a fraction of quality scientific publications in computer science. Furthermore, they miss many citations even for journal articles since all references to them are not captured which originate from conference papers or other papers not included in the publication database.

Several recent system developments capture citation numbers for both journal and conference publications especially in computer science, e.g. Citeseer, the ACM Digital Library, Microsoft Libra (Libra) and Google Scholar (GS). For example, Libra holds more than 900,000 computer science publications and more than 3.5 million citations to them as of December 2007. As shown in Table 1, the majority of papers appeared in conferences and workshops, not in journals. Furthermore, the total number of citations is higher for conferences and workshops than for journals. While there are many workshops and conferences with comparatively little scientific impact the top-cited conferences are highly significant and need to be considered for a meaningful citation analysis in computer science. For example, in the Libra dataset the average number of citations per paper is similar for the 100 most cited conference series than for the 100 most cited journals. These 200 venues account for 78% of all citations.

In [4] we used cleaned citation data from GS for an in-depth citation analysis for database research, a subfield of computer science research. We analyzed all publications over a period of 10 years (1994–2003) which appeared in top database conferences and top database journals. It turned out that the two top conferences (ACM Sigmod, VLDB) not only publish many more papers than the top journals (ACM TODS, VLDB Journal) but that they receive many more citations in total and per paper. The original study used GS data from August 2005. We recently confirmed the findings with GS citation data from December 2007.

Table 1
Journal vs. conference papers and citations in computer science

	#Venues	#Papers (all)	#Cited (all)	#Papers (top 100 venues)	#Cited (top 100 venues)	#Citations per paper (top 100 venues)
Journals	471	321,000	1,655,000	190,000	1,434,000	7.5
Conference/workshop series	2,297	585,000	1,752,000	167,000	1,216,000	7.3

Source: MS Libra.

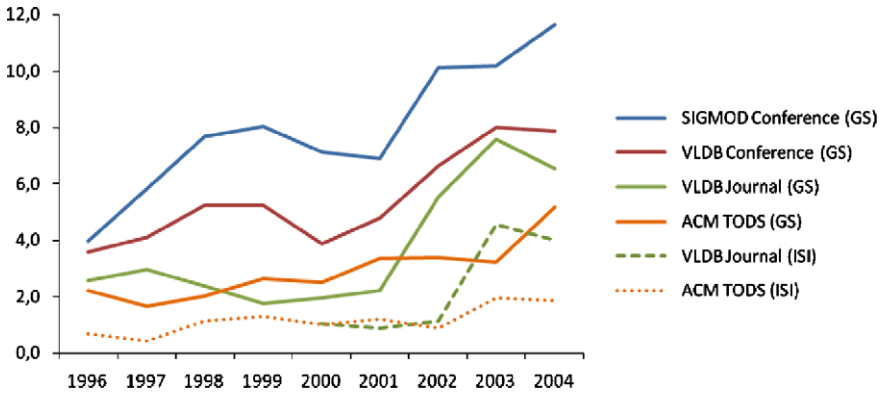

Fig. 1. Impact factors for conferences vs. journals based on data from Google Scholar (GS) and ISI Web of Science.

We also determined impact factors for the considered venues using cleaned GS citation data. The impact factor of a venue (journal, conference series) for year X, IF(X), is defined as the average number of citations in year X publications for a article published in the considered venue in the two preceding years X-1 and X-2. Figure 1 shows the impact factors from 1996 to 2004 for the two conferences and two journals based on GS data from 2007. For comparison, we also show the impact factors for the two journals as recorded in the ISI Web of Science JCR (dashed lines). The results show that for all years the impact factors for the two conferences are significantly higher than for the journals. Furthermore, the GS-based impact factors are much higher than the "official" ones from ISI Web of Science. This confirms that the latter data source misses many citations even for journal articles.

A drawback of data sources such as MS Libra and especially GS is that they incur a significant amount of postprocessing for data cleaning, e.g., to match citations, removing duplicate entries etc. These tasks were supported by several new data integration tools and aligning the data with reference bibliographies such as DBLP [3,5].

References

[1] M. Amin and M. Mabe, Impact factors: Use and abuse, *Perspectives in Publishing* **1** (2000), 1–6.
[2] H.F. Moed and M.S. Visser, Developing bibliometric indicators of research performance in computer science: An exploratory study, Technical Report, CWTS, University of Leiden, 2007.
[3] E. Rahm, A. Thor et al., iFuice – information fusion utilizing instance correspondences and peer mappings, in: *Proc. 8th WebDB*, 2005, pp. 7–12.
[4] E. Rahm and A. Thor, Citation analysis of database publications, *ACM Sigmod Record* **34**(4) (2005), 48–53.
[5] A. Thor and E. Rahm, MOMA – A mapping-based object matching system, in: *Proc. CIDR (Conf. on Innovative Database Research)*, 2007, pp. 247–258.