Citation analysis of database publications

Erhard Rahm, Andreas Thor
University of Leipzig, Germany
{rahm | thor }@informatik.uni-leipzig.de

Abstract

We analyze citation frequencies for two main database conferences (SIGMOD, VLDB) and three database journals (TODS, VLDB Journal, Sigmod Record) over 10 years. The citation data is obtained by integrating and cleaning data from DBLP and Google Scholar. Our analysis considers different comparative metrics per publication venue, in particular the total and average number of citations as well as the impact factor which has so far only been considered for journals. We also determine the most cited papers, authors, author institutions and their countries.

1. Introduction

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* [AM00]. 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, database research results are primarily published in conferences which are not covered by the JCR citation databases. The two major database conferences, SIGMOD and VLDB, receive and publish many more papers than the two major journals, ACM TODS and VLDB Journal (VLDBJ). Furthermore, these conferences are more than twice as selective as the journals with acceptance rates for research papers of 15-20% vs. 35-45% [Be05]. The number of conference submissions has increased significantly in the last five years [Be05] underlining the high scientific importance of conferences.

The tremendous scope of new scientific archives like Google Scholar makes it possible to freely access citation data for millions of publications and authors and thus to evaluate the citations for entire conferences and journals. For our analysis we utilized our new data integration platform iFuice [Ra05] to combine bibliographic data on conferences and journals from DBLP with citation data from Google Scholar and the ACM Digital library. We evaluate citations for all papers which appeared between 1994 and 2003 in the two conferences SIGMOD and VLDB, and the three journals TODS, VLDBJ and Sigmod Record (SR). The latter is not a refereed journal but more a newsletter which also publishes short research articles of broader interest. It has good visibility in the database community favored by its free online accessibility.

In the next section we briefly discuss previous attempts to evaluate the citation impact of database conferences and journals. Section 3 provides information on the data sources and the data cleaning applied. Sections 4-9 present our comparative citation analysis for the five publication venues. In particular, section 6 analyzes the citation skew, section 7 evaluates the journal and conference citation impacts, section 8 discusses the most frequently referenced papers and authors, and section 9 determines the most referenced institutions and their countries.

2. Previous evaluations

The DBLP website contains a list of the 120 most referenced database publications with a total of about 17,000 citations[1]. The list was determined from about 100,000 citations in the SIGMOD anthology containing research papers from 1975-1999. The list contains 17 books, 11 papers from ACM Computing Surveys, 29 TODS papers, 22 SIGMOD, 6 VLDB, only 1 VLDBJ and no SR paper. Most citations go to the classic papers from the seventies and early eighties, the most referenced paper being the 1976 TODS paper by Chen on the entity-relationship model (608 citations). The youngest entry is from 1996 and only 9 publications have appeared after 1990 so that this list does not reflect the citation impact for the more recent research. Furthermore, the list only reflects citations from the database publications of the anthology but not from other publication venues or related fields.

Citeseer is a large archive of computer science publications collected from the web. Based on the citations found in its document base, publication venues (journals, conferences, workshops, newsletters etc.) which received at least 25 citations were ranked according to their average number of citations per referenced paper[2]. In a list of more than 1200 venues TODS and VLDBJ achieved ranks 51 and 52, SIGMOD rank 66, VLDB rank 106 and SR rank 414. There are several problems with this ranking. First of all, not all papers of a venue are considered but only those which have at least one citation in the Citeseer collection. Second, as already observed in [Sn03] Citeseer includes many unreviewed technical reports but lacks many publications from database journals and conferences. Thirdly, the average number of citations favors venues with a smaller number of papers like

[1] http://www.informatik.uni-trier.de/~ley/db/about/top.html
[2] http://citeseer.ist.psu.edu/impact.html

journals or workshops compared to larger conferences. As an example, the WebDB workshop (associated with the SIGMOD conference) achieved rank 35 and is thus ranked higher than SIGMOD itself. Finally, the list was last generated in May 2003 and thus does not reflect more recent publications. In general, Citeseer seems to become quickly outdated since only comparatively few new documents are added.

3. Data sources and data integration

Our study is based on data from three sources as of August 2005: the DBLP bibliography[3], Google Scholar (GS)[4] and the ACM Digital Library (ACMDL)[5]. DBLP and ACMDL provide bibliographic information on authors, publishers and complete lists of papers per conference and journal. ACMDL also provides many conference and journal documents and citation counts. However, only citations from the documents in the ACMDL collection are considered. Google Scholar covers a huge number of documents by crawling the web automatically but also includes the papers from several digital libraries including those from ACMDL, IEEE, and Springer. GS automatically extracts the bibliographic data from the reference sections of the documents (mostly in PDF and PS formats) and determines citation counts for papers in its collections as well as for citations for which the document is not available. The publications in the result of a query are typically ranked according to the number of citations.

We use our integration platform iFuice [Ra05] to combine the data from the mentioned sources and determine the number of citations for entire conferences and journals. We map each paper found in DBLP for a given publication venue to both ACMDL and GS to determine its citations. Moreover, we perform extensive data cleaning to deal with errors in the citations and limitations of the automatic extraction of references. For instance, GS frequently has several entries for the same paper, e.g. due to misspelled author names, different ordering of authors etc. On the other hand, GS may group together citations of different papers, e.g. for a journal and conference version of a paper with the same or similar title. In addition to dealing with these issues we also determine all author self-citations in order to eliminate them from the citation counts. To extend the scope of our analysis we also did some manual data preparation. In particular, we grouped papers into different types (research, industrial, demo, panel, etc.) and determined the originating institution of papers.

Fig. 1 shows the normalized number of citations for GS and ACMDL to all considered papers published between 1994 and 2003. 100% refers to the total number of GS citations including self-citations. The other curves indicate the shares for GS citations without self-citations, the GS citations from publications for which GS has an associated publication year, and the ACMDL citations without self-citations. The graph shows that on average about 10% of the GS citations are self-citations (i.e. about 90% of the GS citations remain) and that this value increases somewhat to about 17% for recent publications. Only for about 50-60% of its citations GS has the year of the referencing paper. This information is needed to determine the impact factor or the age of citations. GS derives the year of a publication X apparently from citations to X so that the year is often unknown for unreferenced papers. This also explains why the share of GS citations with a year information goes down for more recent years.

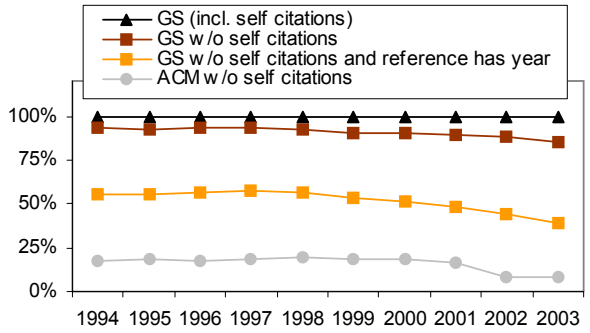

Figure 1: Comparison of different citation counts (100% = Google scholar including self citations)

The number of ACM citations is only about 20% of GS citations until 2001. For 2002 and 2003 the share goes down to about 10% indicating that the ACMDL is less current than GS (in fact, all VLDB papers from 2002 and younger were missing in ACMDL as of August 2005). For these reasons we will only consider the cleaned *GS citations without self-citations* for the remaining analysis. The calculation of impact factors is based on GS citations with a known year for the referencing publication.

4. Base statistics

We determined the number of citations for all papers listed at DBLP for the five considered publication venues and the ten years 1994 - 2003. As of August 2005, we had 81,680 cleaned GS citations for 2,338 papers.

In a first step we analyzed the distribution of citations over different types of papers. Both SIGMOD and VLDB publish not only research papers but also industrial and application papers, demo descriptions, panel and tutorial abstracts and invited papers (mostly with an abstract only). Fig. 2 compares the relative number of papers with the relative citation frequency for these publication types. The "non-research" papers account for a substantial share of the publications, namely 34% and 41% for VLDB and SIGMOD, respectively. However, they receive less than 10% of the citations for both conference series and thus have only a limited impact. This is probably because many of these pa-

[3] http://www.informatik.uni-trier.de/~ley/db/index.html
[4] http://scholar.google.com
[5] http://portal.acm.org

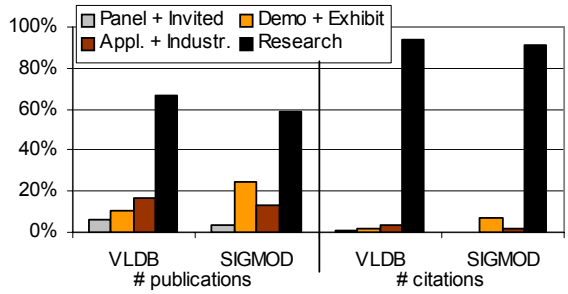

Figure 2: Relative impact of conference paper type

Conference / Journal	# Publications	# Citations	avg. # citations per publication
SIGMOD	446	31,069	70
VLDB	570	28,659	50
SIGMOD Record	327	7,724	24
VLDB Journal	189	4,919	26
TODS	130	4,162	32
Overall	1,662	76,533	46

Table 1: Number of publications and citations

pers are very short (1-4 pages) and the acceptance process is less selective than for research papers. In our remaining analysis we will therefore focus on research papers. For this purpose, we also exclude about 20% of the SR articles like editorials, book reviews, interviews, obituaries etc. from further consideration.

Table 1 summarizes the total number of publications and citations and the average number of citations received per paper for the five publication venues. Most papers (ca. 61%) are published in the two conference series. What is more the total number of citations for SIGMOD and VLDB is almost by a factor 7 higher than for TODS and VLDBJ. Hence the journals have only a comparatively small citation impact. Even with respect to the average number of citations the conferences are clearly ahead by more than a factor 2. This is likely because the most up-to-date research results are primarily published in conferences. Successful conference submissions are published within six to seven months while the so-called end-to-end time for journals (time delay between submission of a manuscript and publication time of the issue with the article) is much higher and highly variable. As outlined in [Be05] the average end-to-end time used to be 2 to 3.3 years for TODS and 1.5 to 2.8 years for VLDBJ in the nineties; currently both journals have an improved average of about 1.5 years.

SIGMOD published fewer papers than VLDB but received more citations resulting in a 40% higher average number of citations per paper. TODS and VLDB Journal are closer together. TODS achieved a better average number of citations while VLDB Journal received more citations in total. SR publishes substantially more (short) papers than either TODS or VLDBJ and achieves a surprisingly high number of citations. As we outline below this is mainly because of some heavily referenced and timely survey papers.

5. Development over time

We now drill down from the summary data into the time dimension to see the distribution over the 10 years. Figures 3-6 show for each publication venue and year the number of papers, the total number of citations, the number of citations to the 5 most referenced papers and the average number of citations per paper.

Fig. 3 indicates that TODS and VLDBJ have a relatively constant and low number of publications per year. (The atypical VLDBJ numbers for 1999 and 2000 are due to a delayed issue.) VLDB publishes most papers per year and recently increased the number of research papers from around 50 to 75 in 2003 to keep pace with an increased number of submissions.

Fig. 4 illustrates the much higher number of citations for the two conferences compared to the journals. Most citations refer to papers from the nineties while the number of citations continuously decreases since 1999. This indicates that many references reach back five and more years giving younger papers comparatively little opportunity to get referenced. The most referenced conferences achieved almost 5000 citations (VLDB94, SIGMOD96, SIGMOD98) while the more recent ones have earned less than 2000 citations so far. Despite the higher number of papers VLDB is referenced more than SIGMOD only in 4 of the 10 years (1994, 1995, 2001, 2002).

Comparing Fig. 4 with Fig. 5 illustrates that the 5 most referenced papers already account for a large portion of all citations. In fact, they receive on average about 40% of all citations for the conferences and 70% for the journals. Interestingly, the top-5 referenced conference papers are on average three times as frequently cited than the top-referenced journal papers. In a few cases (SR1997, SR1998, VLDBJ2001) survey articles helped the journals to close the gap to the conferences.

Fig. 6 shows that SIGMOD dominates w.r.t. the average number of citations especially for the four years 1995-1998 with an average of about 100 references per paper. Another observation is that in the last 5 years the averages for journals and conferences have approached each other.

6. Citation skew per venue

Metrics for entire publication venues like impact factors and total / average number of citations do not allow one to estimate the impact of individual papers or authors. This is because the distribution of citations is typically highly skewed across different papers as already seen from the impact for the 5 most referenced papers (Fig. 5). We now analyze the degree of citation skew in somewhat more detail for the five publication venues.

In a first approach we sort the papers per venue w.r.t. to their citation count and group them into 4 quarters with the

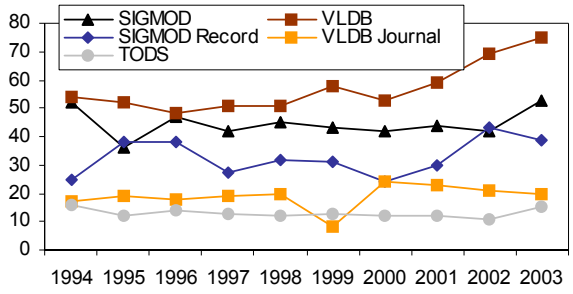

Figure 3: Number of publications

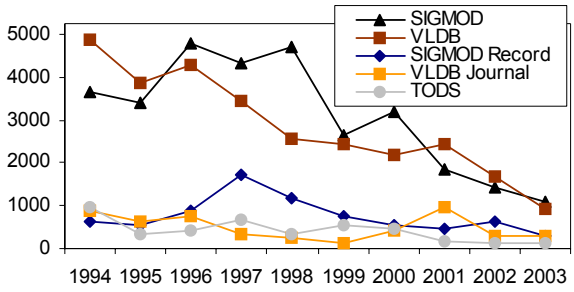

Figure 4: Total numer of citations

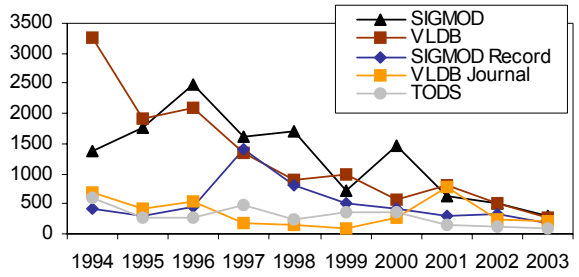

Figure 5: Number of citations for top 5 publications

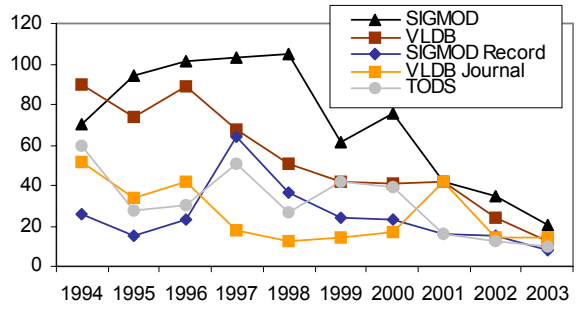

Figure 6: Average number of citations

same number of papers. For each quarter we then determine the relative cumulative citation count. As indicated in Fig. 7 the 25% top-referenced papers account for 60 to 80% of all citations while the bottom 25% of the papers merely achieve 2 - 5% of all citations. In this regard, SR exhibits the highest skew. By contrast, TODS is the most balanced publication venue. VLDB is more skewed than SIGMOD, i.e., the most referenced publications dominate the overall number of citations more and the least referenced papers have even less citation impact for VLDB than for SIGMOD. This might be influenced by the larger number of papers at VLDB compared to SIGMOD. Interestingly, VLDB and VLDB Journal have a very similar citation distribution.

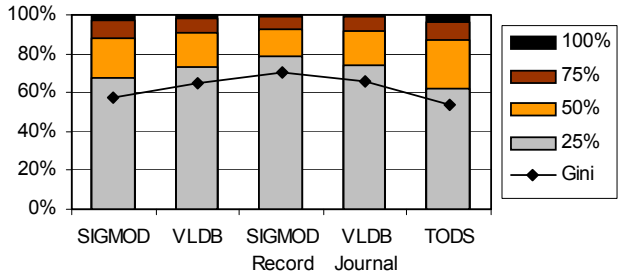

Figure 7: Citation distribution (splitted by quarters) and Gini inde.

As a second, handier metric for the citation skew we consider the so-called Gini coefficient using the Brown formula[6]. The Gini index is a measure of (in our case: citation) inequality. It is a number between 0 and 1 where 0 corresponds to perfect equality (i.e., all publications have the same number of citations) and 1 corresponds to complete inequality (i.e., one paper has all citations). As indicated in Fig. 7 the Gini coefficients confirm our observations above with the highest value for Sigmod Record (0.7) and the lowest for TODS (0.53).

7. Impact factor

The journal impact factor (JIF) determines the average number of citations per paper for a period of two years. For a given year X, the JIF is the average number of times articles from the journal published in the past two years X-2 and X-1 have been cited in year X. For example, the JIF for VLDB Journal in the year 2003 is calculated by dividing the number of 2003 citations to VLDBJ papers from 2001 and 2002 by the number of VLDBJ papers in 2001 and 2002. For the citations recorded in the Journal Citation Report (JCR) database the result is

$$JIF_{VLDBJ2003} = (148 + 52) / (23 + 21) = 4.545$$

which made VLDBJ one of the top-rated computer science journals in 2003. Fig. 8 shows the available JIF values from JCR for the three database journals. The curves indicate that all journals have increased their impact factors in the last two years (which might be influenced by an increased number of evaluated papers). VLDBJ has seen the largest increase thereby outperforming TODS which in turn outperforms SR[7]. The JCR contains additional metrics like the number of citations within a year to all previous articles of a journal (not

[6] http://en.wikipedia.org/wiki/Gini_coefficient

[7] The JCR impact factors for 2002 and 2003 are likely flawed (too low) for SR because they are based on a too high a number of papers (105 and 109 papers for 2000 and 2001 compared to 47 for 2002 and 2003). This underlines the importance of data quality (data cleaning) for citation analysis

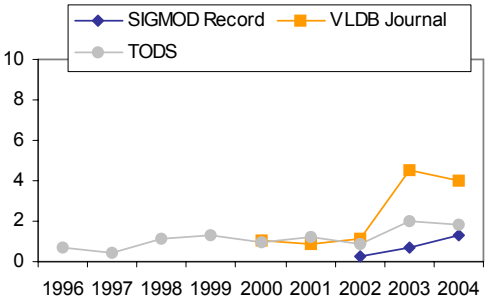

Figure 8: JCR impact factors for journals (2 years)

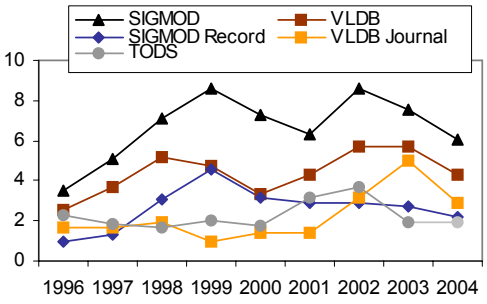

Figure 9: GS-based impact factors (2 years)

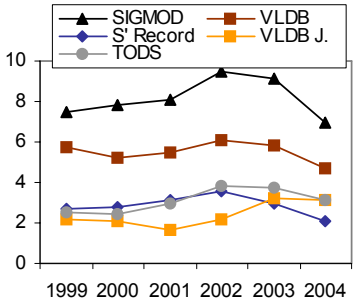

Figure 10: GS-based impact factors (5 years)

only to articles of the two previous years) and the so-called half-life, i.e. the number of recent years accounting for 50% of all references. For instance, there are 870 TODS citations vs. 755 VLDBJ citations in 2004. However, the half-life of TODS (launched in 1976) is more than 10 years, i.e. most citations refer to older articles, which is in line with our observations from the DBLP ranking (section 2). On the other hand, VLDBJ was launched in 1992 and has a half-life of only 4 years, i.e. 50% of the 2004 citations refer to articles from 2001 and younger.

We used the GS citations with a known year (section 3) to determine the impact factors not only for the journals but also for the conferences. In addition, we not only calculated the impact factors over two but also over five years (for instance the five-year impact factor for 2004 indicates the average number of citations in 2004 publications to publications from 1999-2003). The consideration of more than two years was proposed in [Am00] to improve the impact factors' coverage and reduce fluctuations due to a few highly cited articles.

Fig. 9 and Fig. 10 show the resulting impact factors for the five publication venues. We see that the impact factors for the journals are mostly higher than in Fig. 8 because GS provides more citations than the JCR database (despite the elimination of self-citations and citations for which the year of the citing publication is unknown). Furthermore, the impact factor is more stable than the total and average number of citations of Figs. 4 and 6 since the uniform window of 2 and 5 years reduces the bias against younger papers which have less time than older papers to get referenced. The most striking result is that the two conferences achieve excellent impact factors and outperform the journals in all years. SIGMOD achieves the best impact factor in all years. As for the two-year JCR impact factor, VLDBJ reaches the best impact factor of the three journals in the last two years (Fig. 9). SR achieves a surprisingly good impact factor favored by our elimination of non-research papers during data preparation. The five-year impact factors (Fig. 10) are less influenced by a few heavily cited papers. For this extended evaluation period, all journals remain clearly behind the two conferences in all years.

8. Most referenced papers and authors

Tables 2 and 3 show the 5 most cited conference and journal publications, respectively. Longer lists and the top 5 papers per publication venue and year can be found online at *www.ifuice.de*. The by far most cited publication in the considered time frame is Agrawal's and Srikant's 1994 association rule paper (which received the 10-Year-Best-Paper-Award at VLDB 2004). Data mining papers actually contribute substantially to the high citation counts of the conferences in the nineties: 10 from the 20 most referenced papers belong to this category. Their high impact is attributable to

	Title	Authors	Published in	#Cit.
1.	Fast Algorithms for Mining Association Rules	R. Agrawal, R. Srikant	VLDB '94	2261
2.	Querying Heterogeneous Information Sources Using Source Descriptions	A.Y. Levy, A. Rajaraman, J.J. Ordille	VLDB '96	692
3.	BIRCH: An Efficient Data Clustering Method for Very Large Databases	T. Zhang, R. Ramakrishnan, M. Livny	SIGMOD '96	617
4.	Mining Frequent Patterns without Candidate Generation	J. Han, J. Pei, Y. Yin	SIGMOD '00	573
5.	Implementing Data Cubes Efficiently	V. Harinarayan, A. Rajaraman, J.D. Ullman	SIGMOD '96	559

Table 2: Most referenced conference publications (1994-2003)

	Title	Authors	Published in	#Cit.
1.	An Overview of Data Warehousing and OLAP Technology	S. Chaudhuri, U. Dayal	SR '97	634
2.	Lore: A Database Management System for Semistructured Data	J. McHugh, S.Abiteboul, R.Goldman, D.Quass, J.Widom	SR '97	392
3.	Database Techniques for the World-Wide Web: A Survey	D. Florescu, A.Y. Levy, A. Mendelzon	SR '98	391
4.	A Survey of Approaches to Automatic Schema Matching	E. Rahm, P.A. Bernstein	VLDB J '01	324
5.	An Introduction to Spatial Database Systems	R.H. Güting	VLDB J '94	280

Table 3: Most referenced journal publications (1994-2003)

	Author	# Cit.	# Pub.
1.	Rakesh Agrawal	5393	21
2.	Ramakrishnan Srikant	3654	7
3.	Alon Y. Halevy	3052	25
4.	Hector Garcia-Molina	2792	47
5.	Jeffrey F. Naughton	2657	34
6.	Michael J. Franklin	2475	26
7.	David J. DeWitt	2328	27
8.	Jennifer Widom	2176	22
9.	Jiawei Han	1997	17
10.	Christos Faloutsos	1960	22

Table 4: Most referenced authors

the fact that they successfully reached out of the database field to other communities as well. Surprisingly, journal papers on data mining did not appear in the database journals but apparently elsewhere. As Table 3 indicates 4 from the 5 most referenced journal papers are surveys on topics with a substantial research activity to follow on. These surveys helped SR to achieve good citation numbers in 1997/1998 and VLDBJ in 2001, and good two-year impact factors 2 years later.

Table 4 lists the 10 most cited authors together with the number of their papers in the considered venues and time frame. In case of several co-authors per paper, we attributed all citations to each co-author. Four of the ten authors are also in the 2002 list of the ten smallest centrality scores indicating a large co-authorship / social network [Na03].

9. Citation counts by country and institution

For our final evaluation we study the distribution of citations over originating institutions and their countries. For simplicity, we only consider the first author's institution which we extracted manually from the papers. This is obviously very time-consuming so that we restricted ourselves to the research papers with at least 20 citations. Overall, we considered the top-referenced 50% research publications receiving more than 91% of all citations so that we believe that the results are still significant.

Tables 5 and 6 list the top 10 countries and institutions, respectively, with respect to these citation counts. Note that the absolute values cannot directly be compared with the previously presented numbers due to the reduced set of papers. Moreover we attribute the citations of a paper only to one country and one institution, while for Table 4 the citations were assigned to each co-author.

Table 5 shows that almost three quarter of all citations go to papers from US institutions which also contribute by far the most papers. Runners-up are Germany and Canada but with a huge difference to the US. Papers from France and Italy still achieve a noticeable citation impact while countries like UK do not make it on the list.

Table 6 shows that IBM and Stanford are the institutions with the highest citation impact. Only two non-US institutions are listed, the French institute INRIA and the German University of Munich. Comparing Table 5 with Table 6 indicates that papers from some institutions receive more citations than entire countries other than the USA. For example, papers from all German universities combined receive fewer citations than Stanford university alone.

10. Conclusions

We presented a detailed comparative citation analysis for five database publication venues. We note that the two main database conferences SIGMOD and VLDB have a substantially higher citation impact than the database journals TODS, VLDBJ and Sigmod Record, not only in terms of the total number of citations but also with respect to the two-year and five-year citation impact. SIGMOD achieves a higher citation impact than VLDB. Sigmod Record and VLDBJ have benefited from survey articles, while TODS has the least skewed distribution of citations. US institutions receive almost 75% of all citations with papers from IBM and Stanford receiving most citations. The study underlines the high usefulness of Google Scholar for evaluations like this. Data preparation, data cleaning and integrating data from several sources are important to achieve useful and correct results showing the value of tools like iFuice.

Acknowledgements We thank Phil Bernstein, Tamer Özsu, and Rick Snodgrass for their insightful comments.

	Country	# Cit.	in %	# Pub.
1.	USA	51222	73.2	567
2.	Germany	4291	6.1	64
3.	Canada	3270	4.7	34
4.	France	2222	3.2	29
5.	Italy	2025	2.9	22
6.	Israel	858	1.2	6
7.	Japan	724	1.0	6
8.	Denmark	644	0.9	7
9.	Switzerland	612	0.9	8
10.	Greece	590	0.8	12

Table 5: Citations by country

	Institution	# Cit.	# Pub.
1.	IBM	9540	70
2.	Stanford University	7045	62
3.	University of Wisconsin-Madison	5132	60
4.	Bell Labs & AT&T Labs	4500	55
5.	University of Maryland	3153	32
6.	Microsoft	2360	24
7.	University of California, Berkely	1908	24
8.	INRIA (France)	1854	20
9.	University of Washington	1487	15
10.	University of Munich (Germany)	1342	13

Table 6: Citations by institution

11. References

AM00 Amin, M., Mabe, M.: Impact Factors: Use and Abuse. Perspectives in Publishing, Oct. 2000

Be05 Bernstein, P.A. et.al.: Database Publication Practices. Proc. 31st VLDB Conf., 2005

Na03 Nascimento, et al.: Analysis of SIGMOD's co-authorship graph. SIGMOD Record 32(3), 2003

Ra05 Rahm, E., Thor, A., et al.: iFuice - Information Fusion utilizing Instance Correspondences and Peer Mappings, Proc. 8th WebDB, 2005

Sn03 Snodgrass, R.: Journal Relevance. SIGMOD Record 32(3), 2003