HANNA KÖPCKE · ERHARD RAHM

Analyse von Zitierungshäufigkeiten für die Datenbankkonferenz BTW

In diesem Beitrag präsentieren wir eine Auswertung zur Häufigkeit von Zitierungen der Publikationen, die im Rahmen der zehn BTW-Tagungen von 1985 bis 2003 erschienen sind. Die Daten für diese Analyse stammen von DBLP und Google Scholar. Im Rahmen unserer Analyse bestimmen wir auch die meistzitierten Publikationen und Autoren.

1 Einleitung

Anhand von Zitierungszahlen wird oftmals versucht, die Bedeutung von wissenschaftlichen Publikationen, Zeitschriften oder ganzer Forschungsinstitute zu evaluieren. Ein häufig verwendeter Indikator für die Bedeutung einer Zeitschrift ist beispielsweise der »Impaktfaktor« [Amin & Mabe 2000]. Impaktfaktoren für zahlreiche wissenschaftliche Zeitschriften werden jährlich durch Thomson ISI im Journal Citation Report (JCR) veröffentlicht.

Forschungsergebnisse aus dem Informatikbereich »Datenbanken« werden allerdings hauptsächlich auf Konferenzen veröffentlicht, die nicht von den JCR-Zitierungsdatenbanken erfasst werden. In einer kürzlich durchgeführten Auswertung [Rahm & Thor 2005] zeigten wir, dass Publikationen in den international führenden Datenbankkonferenzen SIGMOD und VLDB deutlich häufiger als Veröffentlichungen in den führenden Zeitschriften TODS und VLDB Journal zitiert werden. Für den deutschsprachigen Bereich fehlen solche Analysen bisher. Wir stellen deshalb in dieser Arbeit die Ergebnisse einer Zitationsanalyse für die Datenbanktagungsreihe BTW vor.

Die BTW ist die bedeutendste Tagung zu Datenbanken und deren Anwendungen im deutschsprachigen Raum. Sie wird seit 1985 alle zwei Jahre durchgeführt. Die Abkürzung »BTW« stand bis 2001 für »Datenbanksysteme in Büro, Technik und Wissenschaft«; seit der 10. BTW-Tagung 2003 in Leipzig lautet die Tagungsbezeichnung »Datenbanksysteme für Business, Technologie und Web«. Wir evaluieren in unserer Analyse alle Publikationen, die im Rahmen der BTW in den Jahren 1985 bis 2003 erschienen sind. Die Analyse basiert auf einer Integration und Bereinigung von Daten der Quellen DBLP und Google Scholar.

Der Beitrag ist wie folgt gegliedert: Abschnitt 2 informiert über die Datenquellen und die Durchführung der Analyse. Die Ergebnisse der Analyse werden in Abschnitt 3 präsentiert.

2 Datenquellen und Datenintegration

Die Daten für unsere Analyse haben wir aus zwei Quellen bezogen: der Trierer Informatik-Bibliografie DBLP [DBLP 2006] und Google Scholar [GS 2006]. DBLP ist eine sehr häufig genutzte, qualitativ hochwertige Onlinebibliografie wissenschaftlicher Publikationen aus dem Bereich der Informatik. Der Datenbestand wird an der Universität Trier gepflegt und ist seit 1994 auf derzeit knapp 800.000 Publikationen angewachsen.

Google Scholar ist ein seit November 2004 bestehender Suchdienst zur wissenschaftlichen Recherche, der seit April 2006 auch in deutscher Sprache verfügbar ist. Die in Google Scholar erschlossenen Dokumente stammen sowohl aus dem offenen Web als auch aus proprietären Quellen von Wissenschaftsverlagen, Fachgesellschaften und digitalen Bibliotheken von ACM, IEEE und Springer. Es werden somit auch Zitierungen aus Dokumenten der »grauen Literatur« erfasst, z.B. technische Berichte, Diplomarbeiten oder Seminarausarbeitungen. Google Scholar extrahiert automatisch die bibliografischen Daten aus den Literaturangaben der Dokumente und bestimmt die Zahl der Zitierungen auch für Publikationen, die nicht als Dokument verfügbar sind. Im Vergleich zu anderen bibliografischen Quellen mit Zitierungszahlen wie ACM Digital Library oder Citeseer ist Google Scholar wesentlich umfassender und daher besser zur Zitationsanalyse geeignet [Rahm & Thor 2005]. Dennoch kann natürlich keine Vollständigkeit erreicht werden, da viele zitierende Dokumente nicht elektronisch verfügbar sind. Dies betrifft vor allem ältere Arbeiten, die vor der breiten Web-Nutzung in den 90er Jahren erschienen sind.

Die zur Zitationsanalyse notwendige Integration der Daten erfolgte mit dem Framework MOMA (*M*apping-based *O*bject *MA*tching) [Thor & Rahm 2007] auf Basis unserer Datenintegrationsplattform iFuice [Rahm et al. 2005]. Für jede in DBLP erfasste BTW-Publikation X ermitteln wir alle korrespondierenden Einträge in Google Scholar sowie die zugehörigen Zitierungen, d.h. die anderen Dokumente, die X im Literaturverzeichnis referenzieren. Zahlreiche Qualitätsprobleme von Google Scholar erfordern eine umfassende, semiautomatische Datenbereinigung. Google Scholar listet beispielsweise oftmals redundante Einträge für eine Publikation auf (s. Abb. 1). Dies resultiert meist aus Unsauberkeiten in den Literaturverzeichnissen, insbesondere der Schreibweise und Reihenfolge der Autorennamen, unterschiedliche Tagungsbezeichnungen, uneinheitliche Trennung zwischen Autorenangaben, Titel und Tagungsbezeichnung u.a. Hinzu kommen noch Extraktionsfehler aus PDF-Dokumenten, z.B. bezüglich Sonderzeichen. Ein anderes Problem besteht darin, dass Google Scholar Zitierungen für unterschiedliche Publikationen teilweise zusammenfasst, z.B. für die Zeitschriften- und Konferenzversion eines Papiers mit gleichem oder ähnlichem Titel. Daneben listet Google Scholar unter den Zitierungen nicht nur Fremdzitierungen, sondern auch Eigenzitierungen auf. Zur besseren Abschätzung der Fremdwirkung einer Arbeit sollten diese Eigenzitierungen ermittelt und herausgerechnet werden. Eine zitierende Publikation Y stellt dabei für eine von Y zitierte Publi-

XMach-1: A Benchmark for XML Data Management - Gruppe von 9 »
T Böhme, E **Rahm** - Proceedings of BTW2001, Oldenburg, 2001 - dbs.uni-leipzig.de
Zitiert durch: 53 - Ähnliche Artikel - Im Cache - Websuche - In BibTeX importieren

[ZITATION] XMach-1: A Benchmark for XML Data Management, Datenbanksysteme in Büro, Technik und Wissenschaft
T Böhme, E **Rahm** - GI-Fachtagung
Zitiert durch: 6 - Ähnliche Artikel - Websuche - In BibTeX importieren

[ZITATON] XMach-1: A Benchmark for XML Data Management
T BöRa01 Böhme, E **Rahm** - Proceedings of German database conference BTW2001
Zitiert durch: 1 - Ähnliche Artikel - Websuche - In BibTeX importieren

Abb. 1: Beispiel für redundante Einträge in Google Scholar

kation X eine Eigenzitierung dar, wenn die Autorenmengen von Y und X nicht disjunkt sind. Ferner sind Duplikate bei den zitierenden Arbeiten zu eliminieren, um nicht deren Zitierungen mehrfach zu berücksichtigen.

3 Ergebnisse der Analyse

Die im Folgenden präsentierten Zitierungszahlen beziehen sich auf den Auswertungszeitpunkt September 2006. Damit bestehen für die einzelnen Tagungen unterschiedliche Auswertungszeiträume. Während für die BTW1985 zitierende Publikationen der letzten 21 Jahre berücksichtigt werden können, beschränkt sich der entsprechende Zeitraum für die BTW2003 auf drei Jahre. Für die Publikationen der BTW 2005 in Karlsruhe und der BTW 2007 in Aachen liegen erst wenige bzw. noch keine Zitierungen vor, sodass wir diese Tagungen bei der Analyse ausklammern.

Tabelle 1 listet für jeden Jahrgang der BTW die Anzahl der Publikationen, die Gesamtzahl der Zitierungen (mit und ohne Eigenzitierungen) sowie die durchschnittliche Anzahl der Zitierungen pro Publikation (ohne Eigenzitierungen) auf. Zu den insgesamt 362 Publikationen (eingeladene Vorträge, begutachtete Lang- und Kurzbeiträge sowie Anwendungs- bzw. Industriebeiträge) der zehn Tagungen wurden im Mittel knapp 3 Zitierungen pro Publikation erreicht. Eigenzitierungen erhöhen die Zitierungszahlen durchschnittlich um 27%.

Bezüglich der zeitlichen Entwicklung fällt auf, dass die letzten gegenüber den ersten fünf BTW-Tagungen wesentlich höhere Zitierungszahlen sowohl insgesamt als auch pro Publikation erzielen. Während die ersten fünf Tagungen durchschnittlich weniger als 70 Zitierungen erreichen (1,8 pro Papier), liegt dieser Wert für die jüngeren fünf Tagungen seit 1995 mit 150 (4,3 pro Papier) mehr als doppelt so hoch. Diese Entwicklung ist mit hoher Wahrscheinlichkeit dadurch beeinflusst, dass Google Scholar ältere Publikationen wesentlich unvollständiger erfasst als neuere Arbeiten, die zu einem großen Teil im Volltext im Web zu finden sind.

Diesen Umstand verdeutlicht Abbildung 2, die für jede BTW-Tagung den Anteil ihrer Publikationen angibt, deren Volltext von Google Scholar erfasst wurde. Für die ersten fünf Tagungen sind durchschnittlich nur 12% der Publikationen eines Jahrgangs im Volltext verfügbar, für die jüngeren fünf Tagungen liegt der Anteil dagegen bei durchschnittlich über 50%. Diese Entwicklung dürfte auch für andere Tagungen und Zeitschriften zutreffen, sodass davon auszugehen ist, dass für die Zeit vor 1995 die meisten Publikationen und Zitierungen von Google Scholar nicht ausgewertet werden konnten. Da der Großteil der Zitierungen erfahrungsgemäß in den ersten Jahren nach Erscheinen einer Publikation erfolgt, sind somit die älteren BTW-Publikationen bezüglich ihrer Zitierungszahlen vermutlich deutlich unterbewertet.

Jahr	Ort	# Publikationen	# Zitierungen mit Eigenzitierungen	# Zitierungen ohne Eigenzitierungen	Durchschnittliche # Zitierungen pro Publikation
1985	Karlsruhe	39	84	72	1,8
1987	Darmstadt	50	52	44	0,8
1989	Zürich	37	58	49	1,3
1991	Kaiserslautern	36	118	81	2,3
1993	Braunschweig	28	105	88	3,1
1995	Dresden	28	229	194	6,9
1997	Ulm	26	159	129	4,9
1999	Freiburg	28	185	127	4,5
2001	Oldenburg	43	180	141	3,2
2003	Leipzig	47	200	150	3,1
Gesamt		**362**	**1370**	**1075**	**2,9**

Tab. 1: Anzahl der Publikationen und Zitierungen

Abb. 2: Anteil der BTW-Publikationen mit Volltext

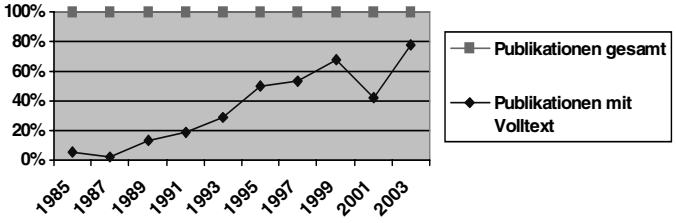
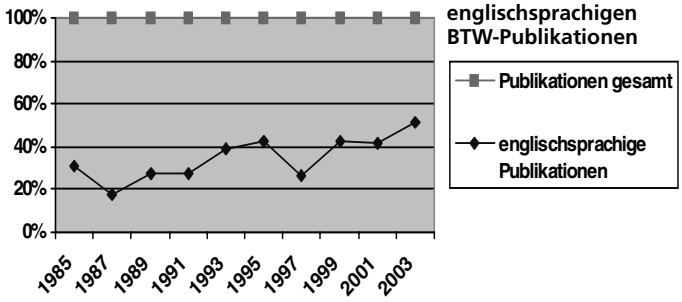


Abbildung 3 zeigt den Anteil der englischsprachigen Publikationen an der Gesamtzahl der BTW-Publikationen. Für eine Konferenz im deutschsprachigen Raum verzeichnet die BTW einen beachtlichen Anteil englischsprachiger Publikationen, der im Laufe der Jahre gestiegen ist. Insgesamt liegt der Anteil englischsprachiger Publikationen bei einem Drittel; im Jahre 2003 betrug der Anteil bereits 50%. Englische BTW-Publikationen erreichen durchschnittlich 6 Zitierungen, deutsche Publikationen durchschnittlich nur 1 Zitierung.

Abb. 3: Anteil der englischsprachigen BTW-Publikationen



Im Vergleich zu internationalen englischsprachigen Konferenzen wie SIGMOD und VLDB ist die durchschnittliche Anzahl der Zitierungen pro BTW-Publikation erwartungsgemäß sehr gering. Gemäß [Rahm & Thor 2005] kommen auf eine seit 1994 erschienene SIGMOD-Publikation durchschnittlich 70 und auf eine VLDB-Publikation durchschnittlich 50 Zitierungen (diese Werte beziehen sich auf Forschungspapiere; Industrie- und Anwendungspapiere liegen deutlich niedriger). Der Selbstzitierungsanteil liegt mit durchschnittlich 10% der Zitierungen deutlich unter dem Wert der BTW-Tagungen.

Noch stärker als bei den internationalen Tagungen gibt es für die BTW-Tagungen eine starke Ungleichverteilung der Zitierungen zwischen den einzelnen Papieren. Zu 160 Publikationen, das sind 44% aller Publikationen, werden von Google Scholar keine Zitierungen oder nur Eigenzitierungen gefunden. Bei den fünf jüngeren Tagungen seit 1995 beträgt dieser Anteil immer noch 31%. Nur 21 Publikationen, das sind gerade mal 6 % aller Publikationen, haben mehr als 10 Zitierungen. Sie decken jedoch bereits 52% aller Zitierungen ab.

Tabelle 2 zeigt die zehn meistzitierten BTW-Publikationen über den gesamten Zeitraum, die bereits 38% aller Zitierungen auf sich vereinen. Auffällig ist die Dominanz englischsprachiger Publikationen, sie belegen die ersten acht Plätze. Drei der zehn Arbeiten sind eingeladene Beiträge (Nr. 4, 6, 10).

Tabelle 3 zeigt für jeden Jahrgang der BTW-Konferenz jeweils die meistzitierte Publikation. In drei Fällen (1985, 1989, 2003) erzielte ein eingeladener Beitrag die meisten Zitierungen.

Die 362 BTW-Publikationen stammen von 579 unterschiedlichen Autoren. Tabelle 4 listet die zehn meistzitierten Autoren mit der Anzahl ihrer BTW-Publikationen auf. Bei mehreren Koautoren haben wir die Zitierungen jedem Koautor zugerechnet. Der mit Abstand meistzitierte BTW-Autor ist demnach Frank Leymann.

Tab. 2: Die zehn meistzitierten BTW-Publikationen (1985-2003)

	Titel	Autoren	Jahr	# Zitierungen
1.	Supporting Business Transactions Via Partial Backward Recovery In Workflow Management Systems	F. Leymann	1995	100
2.	Research Issues in Data Warehousing	M. Wu, A.P. Buchmann	1997	75
3.	XMach-1: A Benchmark for XML Data Management	T. Böhme, E. Rahm	2001	52
4.	Web Services: Distributed Applications Without Limits	F. Leymann	2003	36
5.	Rule-Based Dynamic Modification of Workflows in a Medical Domain	R. Müller, E. Rahm	1999	29
6.	Multimedia Database Systems - The Notion and the Issues	T.C. Rakow, E.J. Neuhold, M. Löhr	1995	29
7.	Principles of Object-Oriented Query Languages	A. Heuer, M.H. Scholl	1991	29
8.	Type Checking in XOBE	M. Kempa, V. Linnemann	2003	24
9.	Data Mining von Workflow-Protokollen zur teilautomatisierten Konstruktion von Prozeßmodellen	M.K. Maxeiner, K. Küspert, F. Leymann	2001	20
10.	Architektur von Datenbanksystemen für Non-Standard-Anwendungen	T. Härder, A. Reuter	1985	18
				Gesamt: 412

Jahr	Titel	Autoren	# Zitierungen
1985	Architektur von Datenbanksystemen für Non-Standard-Anwendungen	T. Härder, A. Reuter	18
1987	Managing Schema Versions in a Time-versioned Non-First-Normal-Form Relational Database	P. Dadam, J. Teuhola	6
1989	Query Languages for Object-Oriented Database Systems: Analysis and a Proposal	F. Bancilhon	10
1991	Principles of Object-Oriented Query Languages	A. Heuer, M.H. Scholl	29
1993	Adding Active Functionality to an Object-Oriented Database System - a Layered Approach	K.R. Dittrich	17
1995	Supporting Business Transactions Via Partial Backward Recovery In Workflow Management Systems	F. Leymann	100
1997	Research Issues in Data Warehousing	M. Wu, A.P. Buchmann	75
1999	Rule-Based Dynamic Modification of Workflows in a Medical Domain	R. Müller, E. Rahm	29
2001	XMach-1: A Benchmark for XML Data Management	T. Böhme, E. Rahm	52
2003	Web Services: Distributed Applications Without Limits	F. Leymann	36

Tab. 3: Die meistzitierte Publikation pro BTW-Tagung

	Autor	# BTW-Publikationen (1985-2003)	# Zitierungen
1.	Frank Leymann	4	156
2.	Erhard Rahm	5	88
3.	Alejandro P. Buchmann	3	78
4.	Ming-Chuan Wu	1	75
5.	Timo Böhme	1	52
6.	Gerhard Weikum	12	46
7.	Marc H. Scholl	3	39
8.	Klaus R. Dittrich	13	37
9.	Andreas Heuer	1	29
10.	Erich J. Neuhold	1	29

Tab. 4: Die meistzitierten Autoren (ohne Eigenzitierungen)

4 Zusammenfassung und Ausblick

Die Untersuchung zeigt, dass Google Scholar und Datenintegrationswerkzeuge wie iFuice auch zur Zitationsanalyse speziellerer, nationaler Tagungen wie der BTW genutzt werden können. Erwartungsgemäß liegen die Zitierungshäufigkeiten für die BTW deutlich unter denen internationaler Top-Tagungen. Englischsprachige Arbeiten wurden durchschnittlich sechsmal häufiger zitiert als deutschsprachige Arbeiten. Ein Großteil der Zitierungen entfällt auf sehr wenige Publikationen. Google Scholar hat bezüglich älterer, vor 1995 erschienener Arbeiten nur einen geringen Abdeckungsgrad. Damit sind die Zitierungszahlen für ältere BTW-Publikationen unterbewertet.

Die Studie zeigte einmal mehr die Problematik der eingeschränkten Qualität von Web-Daten, sodass ein erheblicher manueller Nachbearbeitungsaufwand erforderlich war. Eine wichtige Herausforderung besteht daher in der Entwicklung verbesserter automatischer Verfahren zur Bereinigung von Web-Daten, insbesondere zur Identifizierung korrespondierender Instanzen (Objekt-Matching).

Literatur

[Amin & Mabe 2000] *Amin, M.; Mabe, M.:* Impact Factors: Use and Abuse. Perspectives in Publishing, Oct. 2000.

[DBLP 2006] *DBLP:* Digital Bibliography & Library Project, 2006, *www.informatik.uni-trier.de/~ley/db/*.

[GS 2006] Google Scholar, 2006, *http://scholar.google.com*.

[Rahm et al. 2005] *Rahm, E.; Thor, A.; Aumüller, D.; Do, H.-H.; Golovin, N.; Kirsten, T.:* iFuice – Information Fusion utilizing Instance Correspondences and Peer Mappings. In: Proc. 8th WebDB, 2005.

[Rahm & Thor 2005] *Rahm, E.; Thor, A.:* Citation Analysis of Database Publications. In: SIGMOD Record 34 (4), 2005.

[Thor & Rahm 2007] *Thor, A.; Rahm, E.:* MOMA – A Mapping-based Object Matching System. In: Proc. 3rd Biennial Conf. on Innovative Data Systems Research (CIDR), 2007.

Hanna Köpcke
studierte Informatik an der Universität Dortmund. Seit 2006 ist sie Stipendiatin im Graduiertenkolleg Wissensrepräsentation an der Universität Leipzig und promoviert bei Prof. Dr. Rahm zum Thema »Object Matching«.

Erhard Rahm
ist Lehrstuhlinhaber für Datenbanken am Institut für Informatik der Universität Leipzig. Er promovierte und habilitierte an der Universität Kaiserslautern und verbrachte Forschungsaufenthalte in den USA bei IBM sowie Microsoft Research. Er ist Autor mehrerer Bücher und zahlreicher Konferenz- und Zeitschriftenbeiträge. Derzeit ist er Sprecher des GI-Arbeitskreises »Web und Datenbanken«.

Dipl.-Inform. Hanna Köpcke
Prof. Dr. Erhard Rahm
Universität Leipzig
Abteilung Datenbanken
Postfach 100920
04009 Leipzig
{koepcke | rahm }@informatik.uni-leipzig.de
www.informatik.uni-leipzig.de