Internat. Math. Nachrichten
Nr. 213 (2010), 1–5

Zitate zählen

Günter Rote
Freie Universität Berlin

In den *Internationalen mathematischen Nachrichten* vom August 2008 (Heft Nr. 208) ist eine verdienstvolle Studie abgedruckt, der die Verwendung von Zitatstatistiken und insbesondere des berüchtigten *impact factors* für mathematische Zeitschriften kritisch beleuchtet. (S. 1–25, *Citation Statistics*, von Robert Adler, John Ewing, und Peter Taylor).

Als verspäteten Nachtrag dazu möchte ich an einer Stelle dieser Studie Kritik anbringen, die vielleicht vom methodisch-mathematischen Gesichtspunkt interessant ist. Wer das *IMN*-Heft nicht zur Hand hat, kann auch die online-Fassung vom Juni 2008 nehmen, die man im Netz findet.[1]

Vorab möchte ich betonen, dass ich mit den Grundaussagen und den Schlussfolgerungen dieses Berichts voll einverstanden bin. Wenn ich an einer Stelle herumkrittele, wo ich finde, dass die Autoren über das Ziel hinausgeschossen haben, so soll das keineswegs als grundlegende Kritik an der Studie verstanden werden.

Qualität von Zeitschriften. Im dritten Abschnitt, *Ranking papers*, (S. 10–14 im *IMN*-Heft, S. 9–12 in der Online-Fassung), geht es darum, dass man aus dem Qualitätsurteil über eine Zeitschrift, die sich aus dem *impact factor* ergibt, ein Qualitätsurteil über die in ihr erscheinenden Arbeiten ableitet. Die Autoren kritisieren diese Schlussweise, allerdings mit einer Methode, die sich prinzipiell dagegen richtet, dass man aus der Zeitschrift, in der etwas erscheint, ein Urteil über den Wert der Arbeit herleitet.

Nun ist dieses Vorgehen nach meiner Erfahrung gang und gebe, und mein Ziel ist es, dieses Vorgehen zu rechtfertigen, oder zumindest gegen die ungerechtfertigte Kritik zu verteidigen.

Jeder, der längere Zeit auf einem Gebiet arbeitet und publiziert, kennt dort die „guten“ und die „weniger angesehenen“ Zeitschriften, die „Hauszeitschriften“ von nur lokaler Bedeutung, und die „Spitzenzeitschriften“, bei denen man den KollegInnen gratuliert, die dort etwas untergebracht haben (oder es ihnen neidvoll

[1] *http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf*

ISSN 0020-7926

missgönnt, je nachdem). Wer etwas publizieren möchte, fragt sich oft, ob die Ergebnisse gut genug sind, um es mit Zeitschrift *A* zu probieren, oder ob der Artikel lieber gleich bei Zeitschrift *B* eingereicht werden soll.[2] Wenn ich eine Publikationsliste begutachten soll, bei der ich das Gebiet gut kenne, nicht aber die Arbeiten selbst, schaue ich zu allererst (vielleicht noch bevor der Reflex des Zählens einsetzt), *wo* die Arbeiten erschienen sind.

Es ist klar, dass man aus der „Qualität" der Zeitschrift, in der eine Arbeit erschienen ist, nicht direkt auf die Qualität dieser Arbeit und der zugrunde liegenden Forschungen schließen kann; in jeder Zeitschrift erscheinen bessere und schlechtere Artikel. Dennoch halte ich dieses Vorgehen als Näherungslösung für sinnvoll, solange man nicht *die Arbeiten selbst* lesen und begutachten will, und ich weiß aus Kommissionen und Gutachten, dass es der gängigen Praxis entspricht.

Wenn man es nun als gegeben annimmt, dass es so etwas wie „bessere" und „schlechtere" Zeitschriften gibt, stellt sich die legitime Frage, wie die „Qualität" einer Zeitschrift zustande kommt.

Natürlich speist sich unser Wissen über die Qualität einer Zeitschrift aus vielen Quellen, zum Beispiel aus Erfahrungen mit dem Begutachtungsprozess, aus Gesprächen mit Kolleginnen, usw., aber wir möchten hier einmal den „wissenschaftlichen" Blickpunkt „von außen" einnehmen, wenn wir diese Frage stellen (wie die Politiker und Forschungseinrichtungen, die uns das Geld geben).

Da wir aus der Qualität der *Zeitschrift* auf die Qualität der in ihr enthaltenen *Artikel* schließen wollen, müssen wir die Qualität dieser Artikel zu einem Qualitätsmaß für die Sammlung der Artikel (die Zeitschrift) *aggregieren*.

Zusammenfassend: Die Qualität einer Zeitschrift ergibt sich also aus der Qualität der in ihr erscheinenden Artikel, und sie strahlt wiederum, wie früher dargelegt, auf die wahrgenommene oder angenommene oder vorhergesagte Qualität der Artikel zurück.

Dies wirft zwei neue Fragen auf: (a) Wie soll man die „Qualität" eines einzelnen Artikels messen? (b) Wie kombiniert man die „Qualitäten" einzelner Artikel zu einer Maßzahl für die Qualität einer Zeitschrift?

Frage (a) wäre Anlass zu einer langen inhaltlichen Debatte. Wer als Herausgeber an einer Zeitschrift beteiligt ist, weiß vielleicht, wie schwierig das Ringen um Qualitätsmaßstäbe ist. Ich möchte die Frage ausklammern, und wir wollen der Argumentation halber als Beispiel einfach annehmen, dass die Qualität an der Anzahl der Zitierungen gemessen wird, so wie es in diesem Teil der Studie gemacht wird. (In anderen Teilen, insbesondere im fünften Abschnitt, *The meaning of citations*, wird die Eignung der Zitierhäufigkeit als Qualitätskriterium eingehend diskutiert.)

[2] Natürlich gibt es jenseits von Qualität auch andere Unterscheidungsmerkmale von Zeitschriften: solche für lange und tiefe Abhandlungen und solche für kurze und schnelle Notizen, solche für Übersichtsartikel, und für mehr angewandte oder mehr theoretische Arbeiten, usw.

Frage (b) andererseits ist eine Aufgabe, die mathematischer Analyse zugänglich ist: Wie kann man die Daten, die man über die Arbeiten in zwei Zeitschriften *A* und *B* hat, möglichst gut zu einem Vergleich von *A* mit *B* oder zu Maßzahlen für *A* und *B* zusammenfassen? Bei der Berechnung des *impact factors* wird einfach der Mittelwert aus den Zitierhäufigkeiten der Einzelarbeiten gebildet.

In der Studie wird nun anhand eines konkreten Beispiels gezeigt, dass der Vergleich, der sich aus dieser Zusammenfassung durch Mittelwertbildung ergibt, mit einem anderen Vergleich wahrscheinlichkeitstheoretischer Natur in Widersprich geraten kann: Die Grafik auf S. 12 (S. 11 der online-Fassung) zeigt für zwei verschiedene Zeitschriften, die *Proceedings of the AMS* und die *Transactions of the AMS*, die Verteilung, wie oft die Artikel aus einer bestimmte Periode bisher zitiert wurden. Daraus ergibt sich nach der gängigen Berechnungsmethode, dass die *Transactions* die einen etwa doppelt so hohen *impact factor* wie die *Proceedings* haben.

Die Autoren stellen nun die Frage:

> Wie hoch ist die Wahrscheinlichkeit, dass ein zufälliger Artikel aus den *Transactions* (der „besseren“ Zeitschrift) öfter zitiert wurde als ein zufälliger Artikel aus den *Proceedings*?

Die Autoren geben diese Wahrscheinlichkeit mit 38 % an, also deutlich weniger als $1/2$. Sie werten das als klares Zeichen, dass der Vergleich von *einzelnen Arbeiten* auf Grund der *impact factors* von *Zeitschriften* „wenig rationale Grundlage“ hat.

Zunächst befremdet es, dass der Gleichstand bei der Anzahl der Zitierungen nicht betrachtet wird. Ich will daher die Frage so erweitern:

> Mit welcher Wahrscheinlichkeit wird ein Artikel aus den *Transactions* öfter/gleich oft/weniger oft zitiert als ein Artikel aus den den *Proceedings*?

In Ermangelung der Originaldaten habe ich die Balken der Grafik mit dem Lineal abgemessen. Die Rechnung ergibt als Antwort 37 % / 44 % / 19 %.

Man kann darüber streiten, ob die 37 % (oder 38 %) für die *Transactions* genügend spektakulär gegenüber den 19 % für die *Proceedings* sind, dass eine doppelt so große Maßzahl für die „Wichtigkeit“ der Zeitschrift herauskommen sollte, insbesondere in Anbetracht der 43 % unentschiedenen Ausgänge. Dass die Frage aber als Frage mit zwei Ausgängen formuliert wird und der Fall der Gleichheit zu Gunsten der *Proceedings* unterschlagen wird, ist zumindest eine tendenziöse Argumentation.

Schwerwiegender ist jedoch, dass die gestellte Frage nach der Wahrscheinlichkeit prinzipiell zum Vergleich ungeeignet ist. Ein bekanntes Paradoxon der Wahr-

scheinlichkeitstheorie sind drei Würfel A,B,C, deren Seiten mit Zahlen derart beschriftet sind, dass beim gleichzeitigen Wurf von A und B die Wahrscheinlichkeit, dass A einen höheren Wert als B hat, größer als $1/2$ ist. Genauso gewinnt aber B über C, und C gewinnt wiederum über A.[3] Die Wahrscheinlichkeitsfrage beim Vergleich der Zeitschriften verläuft nach demselben Muster. Die paarweisen Vergleiche, die man erhält, sind nicht notwendigerweise transitiv, und daher ist dieses Verfahren prinzipiell ungeeignet, um Zeitschriften (oder irgendwelche anderen Gegenstände) in eine Rangfolge zu bringen.

Die Autoren der Studie ziehen aus Ihren Rechenbeispielen folgenden Schluss:

> Die Information, die man aus dem *impact factor* einer Zeitschrift über einzelne Artikel gewinnt, ist „überraschend vage und kann in dramatischer Weise irreführend sein".

Das mag vielleicht sein, aber durch den Wahrscheinlichkeitsvergleich wird diese Aussage nicht untermauert, denn damit ließe sich *jede* Vergleichsmethode für Zeitschriften diskreditieren: Keine wie immer geartete Methode kann Zeitschriften auf Grund der in ihnen enthaltenen Artikel in eine Rangordnung bringen, die mit dem von den Autoren vorgeschlagenen Wahrscheinlichkeitsvergleich in allen Fällen konsistent ist.

Meiner Meinung nach trifft die oben zitierte Schlussfolgerung der Autoren dennoch zu, aber das liegt an den schwachen Aussagekraft der Zitierhäufigkeit und überdies am kurzen Bewertungzeitraum von zwei Jahren, der in den *impact factor* einfließt, wie in der Studie ausführlich dargelegt wird. An sich halte ich die Praxis, die Zeitschrift als (Ersatz-)Kriterium für die Qualität einer Arbeit zu nehmen, für vernünftig. Es wäre interessant, empirisch zu untersuchen, wie stark die Qualität von Arbeiten in einer Zeitschrift korreliert. Niemand wird erwarten, dass die Aussage „Zeitschrift *A* ist besser als *B*" bedeutet, dass jeder Artikel in *A* besser sein soll als jeder Artikel in *B*, aber stimmt es vielleicht näherungsweise und mit großer Wahrscheinlichkeit, wenn man eine Liste von 10 oder 20 Arbeiten betrachtet?

Theorie des Messens. In vielen Situationen ist es erforderlich, komplexe Informationen und Daten in eine Rangordnung (zum Beispiel eine Berufungsliste) oder eine einzelne Bewertungzahl (zum Beispiel einen Gehaltsbonus) zu kondensieren. Die mathematische Disziplin, die sich mit solchen Fragen befasst, ist die *Entscheidungstheorie* und die *Theorie des Messens* (engl. theory of measurement, nicht zu verwechselnd mit der Maßtheorie). Es gibt umfangreiche Monographien zu diesem Thema, z. B. David H. Krantz, R. Duncan Luce, Patrick Suppes,

[3] Für die Leser, die dieses Beispiel nicht kennen und sich nicht selber das Vergnügen machen wollen, sich solche Würfel auszudenken, ist hier eine Lösung: A trägt die Ziffern 1,5,9, und zwar jede Ziffer doppelt; B trägt 3,4,8, und C 2,6,7. A gewinnt über B mit Wahrscheinlichkeit 5/9, und genauso ist es bei den anderen Paaren.

Amos Tversky, *Foundations of Measurement*, 3 Bände (1971–1990). Dieser Bereich berührt auch mathematische Grundlagenfragen. Breiter bekannte Ergebnisse sind vielleicht der Arrow'sche Unmöglichkeitssatz (der Satz vom Diktator) oder die Paradoxien von Verhältniswahlsystemen.

Typischerweise geht man dabei axiomatisch vor, indem man gewisse „vernünftige" Forderungen an die gesuchte Maßzahl stellt. (Soll man zum Beispiel als Autor seine Bewertung verbessern können, indem man einen Artikel, der in einer „schlechten" Zeitschrift erschienen ist, unter den Tisch fallen lässt oder unter einem Pseudonym schreibt, oder gilt das Prinzip „Mehr ist besser"?) Danach untersucht man alle Funktionen, die mit den Forderungen verträglich sind, oder man stellt fest, dass die Forderungen widersprüchlich sind.

Die Autoren der Studie loben ihre Methode des Wahrscheinlichkeitsvergleichs, weil sie „den Wert genauer statistischer Denkungsart im Gegensatz zu intuitiver Beobachtung zeigt". Man möchte entgegnen, dass auch „statistische Denkungsart" nicht vor naiven Fehlern schützt, wenn sie nicht mit dem Wissen um Entscheidungstheorie oder Theorie des Messens verbunden ist.

Die Studie wurde von der Internationalen Mathematischen Union (IMU), dem International Council of Industrial and Applied Mathematics (ICIAM) und dem Institut für Mathematische Statistik (IMS) in Auftrag gegeben. Gerade deshalb hätte sie der unkritischen Anwendung von Mathematik entgegenwirken und in dieser Hinsicht ein Vorbild sein können.